UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

Raissa Scariot Fernandez Camps

A proposição de um framework de *Data Analytics* para o estudo do desempenho da inovação

Porto Alegre
2022

Raissa Scariot Fernandez Camps

A PROPOSIÇÃO DE UM FRAMEWORK DE *DATA ANALYTICS* PARA O ESTUDO DO DESEMPENHO DA INOVAÇÃO

Dissertação apresentada como requisito para obtenção do grau de Mestre pelo Programa de Pós-graduação em Administração da Universidade Federal do Rio Grande do Sul.

Orientador: Prof. Dr. Pablo Cristini Guedes

Porto Alegre

2022

RESUMO

O objetivo deste estudo é propor um *framework* de *data analytics* para classificar setores econômicos em níveis de inovação – em uma escala que vai de altamente a pouco inovadores, a partir de uma base de dados com indicadores de inovação. O problema consiste em entender como se comporta o desempenho de inovação nesses setores, dado o número de empresas inovadoras que contêm e características que apresentam, e é formulado como um problema de classificação. O *framework* combina métodos para normalização da base, determinação do número de classes (níveis de inovação) encontrados nos dados, tratamento de classes desbalanceadas, seleção de variáveis (indicadores de inovação dos setores), classificação e estimação do desempenho da inovação (empresas que inovam no setor em relação ao total da amostra). Para isso, diferentes abordagens são experimentadas. Os modelos *Random Forest*, *Extreme Gradient Boosting* e *Support Vector Machine* são utilizados nas etapas de classificação das observações, seleção de variáveis e estimação da variável de saída. Na determinação do número de classes, são experimentadas abordagens gerencial e de quartis. Técnicas de *Synthetic Minority Oversampling Technique* são testadas para o balanceamento de amostras nas classes. A abordagem analítica no estudo dos dados de inovação das empresas auxilia na compreensão dos fatores que influenciam o desempenho da inovação dos setores e apoiará a tomada de decisão acerca de ações de fomento.

Palavras-chave: classificação; análise de dados; apoio à decisão; desempenho da inovação.

ABSTRACT

The aim of this study is to propose an analytics framework to classify sectors at levels of innovation - on a scale from highly to less innovative, given a database with innovation indicators for economic sectors. The problem is to understand how innovation performance behaves in these sectors, given the number of innovative companies they contain and the characteristics they present, and it is formulated as a classification problem. The framework combines methods for data normalization, determination of the number of classes (levels of innovation), deal with imbalanced classes, feature selection (innovation indicators), classification and estimation (companies that innovate in the sector in relation to the total sample). For this, different approaches are tested. The Random Forest, Extreme Gradient Boosting and Support Vector Machine models are used in the observation classification, feature selection and output estimation steps. To determine the number of classes, managerial and quartile approaches are experimented. Synthetic Minority Oversampling Techniques are tested for balancing classes. The analytical approach in the study of companies innovation data helps to understand which factors that affect the sectors innovation performance and support decision making about fostering actions.

Keywords: classification; data analytics, decision support; innovation performance.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A inovação é entendida como estratégia impulsionadora do desempenho de organizações em estudos nos mais variados contextos, e indicada como condutora de mudanças que garantem a viabilidade de instituições (ANDRONIKIDIS et al., 2020). O que se sabe é que, inovar, por vezes, é concebida como ação subjetiva. Por isso, se faz necessário avaliar quais são as atividades adotadas internamente pelas organizações que as caracterizam, de fato, como inovadoras. Para tanto, Cruz-Cázares et al. (2013) ressaltam a importância de se analisarem simultaneamente nessas organizações os esforços praticados para a inovação e os resultados obtidos, uma vez que não é tão simples estabelecer uma relação clara entre eles.

De acordo com Gomes et. al. (2014), os resultados da inovação podem ser traduzidos, basicamente, em dimensões estratégicas como eficácia, eficiência, custos e melhorias de processo. Essas dimensões tratam do impacto econômico da inovação, como a participação da empresa no mercado e de novos produtos ou processos na receita total de vendas, recursos utilizados para atingir os resultados, custos reduzidos no processo de inovação e melhorias de processo. Sendo assim, é possível medir a inovação dentro das organizações a partir de um conjunto de indicadores internos de suas atividades. E a medição do desempenho de iniciativas de inovação das empresas é necessária para garantir a eficácia de seus investimentos (DEWANGAN; GODSE, 2014). De forma análoga, é possível medir a inovação também de um setor econômico, a partir dos indicadores do conjunto de organizações pertencentes a ele. Essa análise possibilita identificar potenciais setores de crescimento futuro, além de fornecer subsídios para estreitar a gama de áreas candidatas ao desenvolvimento de uma estratégia de especialização inteligente (*Smart Specialisation Strategy*) para alavancar a vantagem competitiva de uma região.

O potencial competitivo de setores do Brasil é apurado pela Pesquisa Industrial de Inovação Tecnológica (PINTEC), visando oferecer subsídios para a definição de estratégias empresariais e políticas públicas voltadas à inovação (IBGE, 2020). Um dos resultados publicados pela PINTEC é a taxa de inovação dos segmentos econômicos selecionados, que "corresponde ao percentual do número de empresas que implementaram inovações de produto ou processo sobre o total de empresas" (IBGE, 2020). A partir dos dados da Pesquisa é possível, portanto, analisar segmentos econômicos e as características que fomentam a inovação.

Na PINTEC de 2017, os setores de atividades da indústria, comércio e serviços do Brasil apresentam diferentes taxas de inovação, o que permite classificá-los em graus de inovação. Entretanto, devido ao enorme número de características desses segmentos, que correspondem às atividades desempenhadas pelas empresas do setor, a classificação se torna inexequível com uma abordagem qualitativa. Assim, este estudo se propõe a suprir a lacuna observada quanto à utilização de métodos analíticos e abordagem *data-driven* da Pesquisa Operacional para classificar setores em níveis de inovação a partir de seus indicadores.

Frameworks que analisam o desempenho da inovação são encontrados na literatura, empregando, geralmente, métodos para predição dos efeitos da inovação. Comumente são utilizados modelos de Redes Neurais, como em Forner et al. (2019), Hajek e Henriques (2017) e Paz-Marín et al. (2012). Já Hajek e Henriques (2019) utilizam, para análise da eficácia de atividades de inovação a partir dos efeitos gerados, programação genética e, Curado et al. (2018), propõem, para uma análise semelhante, a utilização de modelos *fuzzy*. Neste estudo, a predição do desempenho da inovação surge, de certa forma, com o intuito de analisar inversamente o resultado da etapa de classificação: a partir das variáveis selecionadas nessa etapa, se poderia estimar o comportamento do desempenho da inovação dos setores econômicos.

Além da predição, estudos incluem etapas de agrupamento de observações em níveis de inovação, utilizando abordagens de *clustering*. Em Paz-Marín et al. (2012), o algoritmo *k-means* é implementado. Já Forner et al. (2019) aplicam o método *Clustering for Large Applications* como etapa de agrupamento das características correlacionadas a grupos, antes de avaliar, separadamente, o desempenho da inovação de cada um deles. A abordagem adotada neste estudo inclui agrupamento das observações em níveis de inovação como etapa prévia à classificação. Porém, entende-se relevante aqui a utilização de uma abordagem gerencial para esse agrupamento – denominado no *Framework* como etapa de definição do número de classes. A abordagem gerencial corresponde a uma análise por parte do decisor quanto ao desempenho de inovação dos setores econômicos. Ainda assim, para fins de verificação da efetividade da análise gerencial, se propôs o uso de uma abordagem estatística de quartis para comparação dos resultados e inclusão da técnica mais adequada no *Framework*.

Assim sendo, o *Framework* de métodos proposto neste trabalho busca responder à seguinte questão principal: como classificar os setores em mais ou menos inovadores a partir das suas características? Tendo em vista a resposta, o comportamento da taxa de inovação dos

setores a partir dessas características é também analisado. Portanto, estabelecer uma metodologia de classificação das observações em níveis de desempenho da inovação é o objetivo principal da análise. O *Framework* é desenvolvido a partir dos dados da PINTEC 2017 relativos a setores selecionados da indústria extrativa e de transformação de 15 estados brasileiros que detêm no mínimo 1% do Valor da Transformação Industrial (VTI) nacional (IBGE, 2020).

A pesquisa está estruturada em cinco capítulos. O primeiro capítulo, que contém esta introdução, aborda a contextualização, justificativa e objetivos do estudo; o próximo capítulo apresenta a revisão da literatura; o capítulo 3, apresenta o método e a abordagem utilizada na construção dos modelos de análise dos dados, o capítulo 4, os resultados da pesquisa, e por fim, no capítulo 5, a conclusão.

1.1 JUSTIFICATIVA

Diante da tendência das regiões e países priorizarem áreas para se impulsionar a inovação, capitalizando ativos específicos e recursos existentes na sua base de conhecimento para se tornarem competitivos (RUSU, 2013), é imprescindível que se desenvolvam métodos para classificar os setores econômicos em mais ou menos inovadores a partir das características que apresentam. A classificação de setores econômicos em níveis de inovação não fornece informações suficientes para a implementação de uma estratégia de especialização inteligente. Como afirma Krammer (2017), para se identificar mais claramente as vantagens competitivas de uma região, são examinados níveis mais desagregados como produto, nicho ou indústria, em detrimento à análise de empresas, empresários ou cluster. Porém, a classificação cria uma espécie de “filtro” dos segmentos em que a inovação está evidente, viabiliza a análise das principais características que melhoram seu desempenho e orienta a busca pelas áreas onde há oportunidade de especialização. O aumento do número de empresas que buscam na inovação um meio para aumentar o grau de diferenciação dos seus negócios frente as demais também justifica o interesse em desenvolver este estudo no tema.

Pesquisas relacionadas às características das organizações que explicam a inovação ou o seu desempenho têm se expandido nos últimos anos no Brasil. Entretanto, o assunto é predominantemente discutido em pesquisas empíricas e de constructos econômicos, em geral, sob uma perspectiva de economia industrial ou de gestão de negócios (VEGA-JURADO ET

AL., 2008). Como exemplo, Gonçalves Filho et al. (2013) classificam empresas em *clusters* a partir de diferentes níveis de desempenho, dados pela participação de mercado, lucratividade, aumento da receita com vendas, retorno sobre os ativos, desempenho geral e capacidade de inovação, e analisam as diferenças entre perfis de organizações. Longhini et al. (2018) estudam a relação entre investimentos em inovação e a receita líquida de vendas nos setores nacionais. Santos e Pestillo (2019) estudam como padrões setoriais de inovação explicam o desempenho da indústria brasileira, medido por indicadores de lucratividade e rentabilidade. Santos et al. (2014) também analisam a relação entre inovação e o desempenho econômico de empresas brasileiras, a partir de indicadores de retorno financeiro e margem operacional.

Estudos que desenvolvem métodos para apoiar à tomada de decisão de empresas, setores ou regiões quando avaliadas atividades e resultados da inovação são amplamente explorados internacionalmente, como em Park et al. (2021), Hajek et al. (2019), Forner et al. (2019), Debaere et al. (2018), Curado et al. (2018) e Hajek e Henriques (2017). Forner et al. (2019) inclusive comentam sobre a tendência crescente e a difusão de estudos que utilizam métodos analíticos para identificar os determinantes da inovação e prever o efeito de instrumentos políticos sobre o seu desempenho. Entretanto, até o momento, se desconhecem estudos que classifiquem setores inovadores a partir de suas características, especialmente sob a óptica do contexto nacional e com dados nacionais.

O informativo de divulgação dos resultados da base de dados utilizada neste estudo - PINTEC 2017 - divulga a taxa geral de inovação das empresas no triênio de avaliação, a fim de comparar o cenário brasileiro ao longo do tempo. No período 2015 a 2017, a pesquisa aponta que cerca de 1/3 das empresas foram inovadoras em produto ou processo, o que resulta em uma taxa geral de inovação de 33,6%. Já no triênio 2012 a 2014, a taxa geral de inovação é de 36%. Entretanto, a taxa geral de inovação não é avaliada setorialmente. Sendo assim, a análise da base de dados no contexto desta pesquisa dá origem ainda à proposição de um novo uso gerencial dos resultados, como um ranking de inovação dos setores econômicos.

Em atenção aos fatos, a abordagem quantitativa presente neste estudo poderá validar intuições da literatura da área de inovação, utilizando de uma abordagem *data-driven decision* para classificar setores econômicos brasileiros em graus de inovação, de acordo com as práticas implementadas pelas empresas que neles estão inseridas.

1.2 OBJETIVO GERAL

O objetivo geral deste estudo é desenvolver um *framework* de métodos para analisar setores econômicos e classificá-los em níveis de inovação a partir das suas características.

1.3 OBJETIVOS ESPECÍFICOS

Para que o objetivo geral da pesquisa seja alcançado, têm-se por objetivos específicos:

a) Identificar quantos níveis de inovação (classes) são mais adequados para segmentar o conjunto de observações;
b) Constatar quais são as principais características que influenciam na classificação e estimar valores para a variável de saída, a partir das características selecionadas;
c) Comparar e selecionar a técnica que apresenta o modelo mais adequado para a classificação de setores econômicos em níveis de inovação, possibilitando análises futuras de fomento à inovação e seleção de segmentos com potencial de especialização inteligente;
d) Propor um *framework* de métodos para analisar a base de informações de atividades e resultados de inovação de setores econômicos e compreender como essas atividades influenciam na taxa de inovação desempenhada.

2 REVISÃO DA LITERATURA

Neste capítulo são abordados referenciais sobre o problema estudado, trabalhos relacionados e métodos de solução propostos. Portanto, a seção 2.1 apresenta o problema de análise e os estudos relacionados ao desempenho de indicadores de inovação, e na seção 2.2, são abordados os métodos de classificação utilizados na pesquisa.

2.1 O PROBLEMA DE ANÁLISE DE INDICADORES DE INOVAÇÃO

Trienalmente, a partir dos resultados da PINTEC, o IBGE divulga a taxa geral de inovação das empresas do setor de Indústria, de Eletricidade e gás e de Serviços. A taxa é calculada a partir do número de empresas inovadoras, que, pela PINTEC, significa aquelas que introduziram produto no mercado ou implementaram processo internamente (IBGE, 2020). Logo, as empresas inovadoras são assim conhecidas porque obtiveram algum resultado com atividades de inovação. O problema deste trabalho pode ser caracterizado, portanto, como um problema de análise de indicadores de desempenho da inovação. O desempenho da inovação é então o resultado de um conjunto de atividades realizadas e características das organizações que as realizam.

Alguns autores têm estudado como componentes de organizações, regiões e países impactam no número de inovações lançadas no mercado ou mesmo em seus resultados financeiros e econômicos. Forner et al. (2019), por exemplo, predizem *outputs* da inovação em países, investigando como mudanças em atividades-chave impactam no índice de inovação nacional. Hajek e Henriques (2017) analisam como características intra e inter-regionais influenciam no desempenho da inovação de regiões, que é medido pelo número de inovadores e por efeitos econômicos no mercado. Kannebley et al. (2005) estudam a PINTEC de 2000 e buscam identificar as características preditoras da inovação em empresas brasileiras, sendo a inovação medida pela introdução de novos produtos e processos no mercado. Posto isso, o presente trabalho aborda o problema de análise de indicadores de inovação, buscando classificar setores em níveis mais ou menos inovadores e determinar as principais atividades e características que levam os setores a tais níveis.

Na busca das características que descrevem o perfil dos mais inovadores, sejam eles territórios ou organizações, uma série de componentes são analisados. Além dos conhecidos

grupos de indicadores de recursos dedicados à Pesquisa e Desenvolvimento e patentes lançadas, também são relevantes as atividades relacionadas à adoção de tecnologias da informação e comunicação, biotecnologia e gestão do conhecimento (OECD, 2005). De acordo com o Manual de Oslo, informações sobre atividades inovadoras podem ainda ser extraídas de diversas fontes, como pesquisas de negócios, estatísticas educacionais, indicadores de atividades em setores de alta tecnologia, recursos humanos, publicações científicas, entre outros (OECD, 2005).

A análise do desempenho da inovação tem sido estudada sob uma perspectiva analítica e quantitativa por diversos autores. Hajek e Henriques (2017), por exemplo, utilizam em sua pesquisa Redes Neurais Artificiais (RNA) com um modelo *Multilayer Perceptron* (MLP*)* para analisar combinações de características intra e inter-regionais e prever o desempenho da inovação em regiões europeias - o desempenho da inovação, nesse caso, é representado por múltiplas saídas que contemplam indicadores como o percentual de inovadores e os efeitos econômicos da inovação. Neste estudo, uma única variável de saída é abordada: o número de inovadores de produto e processo nos setores analisados. Uma análise de sensibilidade para verificar se o incremento em certas características gera efeito positivo ou negativo nas variáveis de saída também é foco de investigação por Hayek e Henriques (2017). Em oposição ao estudo de Hayek e Henriques (2017), experimentos com outros valores para as variáveis não são explorados aqui, devido ao caráter da pesquisa estar voltado à proposição de um método de aprendizado que melhor interpreta os dados disponíveis.

Também na categoria preditiva de análise do desempenho da inovação, Hajek et al. (2019) avaliam a eficácia de atividades regionais e os múltiplos efeitos em inovação que elas geram usando programação genética baseada em semântica, com um otimizador de Busca Local. Por Curado et al. (2018), um modelo *fuzzy-set* é empregado para determinar a relação entre características de empresas com seu desempenho da inovação, medido por uma única saída: inovações de produto. Os resultados da pesquisa de Curado et al. (2018) trazem ainda as combinações de características que levam a um bom desempenho de inovação, de natureza parecida à análise de sensibilidade realizada por Hayek e Henriques (2017). Similar à abordagem preditiva constatada nos trabalhos de Hayek e Henriques (2017), Hayek et al. (2019) e Curado et al. (2018), uma das etapas proposta neste estudo objetiva estimar a variável de saída – desempenho da inovação - a partir de um conjunto selecionado de variáveis de entrada.

Combinando procedimentos de *clustering*, análise de correlação de recursos e Análise de Componentes Principais (PCA), Forner et al. (2019) agrupam países em quatro famílias a partir do seu desempenho da inovação - desempenhador de inovação, seguidor da inovação, desafiador da inovação e subdesenvolvidos. Utilizando a abordagem de *Clustering for Large Applications* (CLARA) e uma Rede Neural Bayesiana, os autores identificam os principais fatores determinantes dos grupos de inovação, considerando as características correlacionadas a eles. A abordagem adotada neste trabalho se assemelha ao trabalho de Forner et al. (2019), se considerarmos que a classificação de setores em níveis de inovação também se faz a partir do agrupamento dos setores conforme seu desempenho de inovação. Porém, enquanto Forner et al. (2019) avaliam separadamente as características pertencentes a cada *cluster*, a metodologia desenvolvida nesta pesquisa examina simultaneamente todas as características para identificar aquelas que conduzem à classificação.

Abordagens de classificação também são a escolha de Paz-Marín et al. (2012) na análise do desempenho da inovação. Em sua pesquisa, os autores classificam países europeus em quatro níveis de desempenho de Pesquisa e Desenvolvimento (P&D) – inovação baixa, inovação moderada, inovação alta e motor de inovação - a partir de um conjunto de características. Os níveis de desempenho são obtidos a partir do agrupamento dos países em *clusters*, aplicando o algoritmo *k-means.* Para a classificação das observações, os autores utilizam técnicas de RNA com os modelos MLP e *Product-Unit Neural Network.*

2.2 MÉTODOS DE CLASSIFICAÇÃO

Abordagens de classificação englobam métodos e modelos treinados para o reconhecimento de padrões nos dados. Análises preditivas também podem fazer uso desses métodos. Para composição do *Framework* deste estudo, foram escolhidos dois métodos baseados em *decision-trees*, *Random Forest* (RF) e o *Extreme Gradient Boosting* (XGB) e outro amplamente conhecido, dentre métodos de *Machine Learning*, o *Support Vector Machine* (SVM). Todos eles são explorados a seguir.

2.2.1 *Random Forest*

O RF é um método que pode ser utilizado tanto para classificação de observações, em que as variáveis de resposta são categóricas, quanto para regressão, em que a resposta fornecida

é uma variável contínua (CUTLER et al., 2012). O método consiste numa floresta aleatória de árvores de decisão, formadas por amostras de dados, que operam em conjunto para determinar o resultado da classificação ou regressão. As árvores são modelos não correlacionados que votam individualmente na classe prevista, como um "comitê" (HASTIE et al., 2009). Na classificação, a predição ocorre pela agregação da maioria de votos das árvores de decisão e, na regressão, ocorre pela média das saídas (SULISTIANI; TJAHYANTO, 2017). Uma floresta aleatória é, portanto, um classificador constituído de uma coleção de árvores de decisão com vetores aleatórios *k* independentes e identicamente distribuídos, onde cada árvore lança um voto para a classe mais popular na entrada *X* (BREIMAN, 2001). A operação em conjunto é o que torna o método tão poderoso em relação à predição por modelos individuais, já que, segundo Hastie et al. (2009), a ideia em técnicas *bagging* – como o *Random Forest* - é obter a média dos votos imparciais dos modelos e assim, reduzir a variância na predição.

Para explicar o procedimento do método RF, é importante introduzir o funcionamento de árvores de decisão. As árvores de decisão usam uma sequência de partições binárias (divisões) que particionam o espaço de variáveis preditoras. O nó "raiz" da árvore contempla todo o espaço preditor, onde o conjunto de dados é introduzido. De acordo com Lauretto (2010), os nós que representam a unidade de tomada de decisão são chamados nós não terminais ou nós intermediários e se dividem, a partir de algum critério, em dois nós descendentes. Critérios que medem a qualidade da divisão em árvores de decisão são a impureza de Gini e o Ganho de Informação ou Entropia. O Índice de Gini mede a redução de impureza nos nós pela escolha da variável, e é descrito, segundo Sawangarreerak e Thanathamathee (2020), pela Equação 1:

$$Gini(S) = 1 - \sum_{i=1}^{k} P_i^2 \quad (1),$$

em que S é um conjunto de dados em que as amostras são de k diferentes classes, e S_i é um conjunto de amostras que pertencem a classe C_i. O número de conjuntos S_i é s_i. P_i^2 é a probabilidade estimada por $s_i \,/\, s$. Quando o índice de Gini é igual a 0, significa que é o máximo de informação útil que pode ser obtida da variável (SAWANGARREERAK; THANATHAMATHEE, 2020).

Já para se medir o Ganho de Informação de uma variável, se utiliza o valor de entropia, definido pela Equação 2:

$$I = \sum_j p(j,x) \log \frac{p(j,x)}{p(j)p(x)}, \quad (2),$$

onde $p(j,x)$ é a distribuição conjunta da classe j e da variável x.

Os nós intermediários contêm o nome de variáveis preditoras. Uma vez que uma divisão foi realizada por algum dos critérios, a árvore ramifica-se para um nó descendente, que é tratado da mesma maneira que o nó original. Os nós que não possuem descendentes são chamados de nós terminais ou folhas. O nó terminal é também conhecido como nó resposta que, no caso da classificação, conterá o nome de uma classe, ou será nulo, quando não houver nenhum exemplo nos dados que corresponda a esse nó (LAURETTO, 2010). O procedimento é repetido até ser encontrado o nó resposta. Isso caracteriza a recursividade de árvores de decisão (SOBRAL, 2003).

No RF, variáveis aleatórias são consideradas pelo algoritmo para dividir os nós internos das árvores de decisão que fazem parte da floresta construída (SENTHILNATHAN et al., 2020). E a divisão para particionar um nó não terminal em dois descendentes é feita escolhendo a melhor variável preditora de acordo com algum critério. Esse critério é entendido também como a função objetivo do modelo e possibilita selecionar as variáveis mais relevantes para a classificação. No RF, a importância das variáveis é calculada pela importância de Gini, que avalia a redução total da impureza trazida por uma variável (BREIMAN, 2001).

Para avaliar o desempenho dos modelos, de acordo com Tanha et al. (2020), são utilizadas métricas de classe única e métricas gerais. Algumas métricas de classe única são *precision*, *recall* e *F-measure* ou *F-score*. *Precision* é o número de observações classificadas corretamente dividido pelo número de vezes que o modelo previu a classe. *Recall* é o número de observações da classe que foi classificada corretamente dividido pelo número total de membros da classe. *F-measure* é uma combinação das métricas de *precision* e *recall* e é definida pela Equação 3, de acordo com Tanha et al. (2020). O parâmetro β corresponde à importância atribuída a cada métrica. Quando a importância das medidas *precision* e *recall* são iguais, F-*measure* é a média harmônica entre elas, conhecida também como F1-*score*.

$$F - measure = \frac{(1+\beta)PrecisionRecall}{\beta^2 Precision + Recall} \quad (3).$$

Todas as métricas são calculadas a partir da matriz de confusão, que consiste em uma matriz bidimensional onde uma dimensão é indexada pela classe real de uma observação e a outra dimensão indexada pela classe que o classificador prevê (TANHA et al., 2020). A Figura 1 a seguir apresenta um exemplo de matriz de confusão. Dada uma observação N_{ij}, com $i \neq j$, a matriz mostra o número de observações classificadas como classe C_j, mas que são da classe C_i. O melhor classificador terá valores zero nas posições que não estão na diagonal (TANHA et al., 2020).

Figura 1 - Matriz de confusão

	C_1	C_2	$\cdots$	C_n
C_1	N_{11}	N_{12}	$\cdots$	N_{1n}
C_2	N_{21}	N_{22}	$\cdots$	N_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_n	N_{n1}	N_{n2}	$\cdots$	N_{nn}

Fonte: adaptado de Tanha et al. (2020)

Dentre as métricas gerais, pode-se citar a acurácia, que calcula a fração de predições corretas em relação ao total de amostras. Se as amostras são balanceadas entre as classes alvo, a acurácia e F1-*score* são praticamente iguais.

Quando o resultado da classificação apresenta uma resposta contínua, ou seja, se o RF é utilizado como método regressor, algumas estatísticas de desempenho preditivo bem conhecidas podem ser utilizadas, como o *Mean absolute error* (MAE), o *Root Mean Squared Error* (RMSE) e o coeficiente de determinação R-*Squared* (R^2). As medidas podem variar no intervalo de 0 (zero) a 1 (um). Quanto mais próximos a zero o MAE e RMSE e mais próximo de 1 for o R^2, melhor o modelo (FOUEDJIO, 2020). As métricas são apresentadas nas Equações 4 a 6 a seguir:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \quad (4),$$

em que $\hat{y}_i$ é a predição e y_i o valor verdadeiro (WILLMOTT; MATSUURA, 2005).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad (5),$$

$$R^2 = 1 - \sum_i \frac{(y_i - \hat{y}_i)^2}{(y_i - \hat{y})^2} \quad (6),$$

e $\hat{y}$ é a média dos valores verdadeiros.

Outro tópico importante a ser observado é a avaliação do desempenho de modelos de classificação aplicados em casos com poucos dados. Segundo Tušar et al. (2017), uma abordagem benéfica para avaliação do desempenho do modelo nessas situações é a *k-fold cross-validation.* O método consiste em dividir os dados em um determinado número *k* de estratos (*folds*), utilizar um dos estratos como conjunto de teste e os demais como conjuntos de treinamento. O estrato de teste é alternado, de forma que após repetido o procedimento várias vezes, todos os *k-folds* são utilizados para teste e o cálculo da medida de desempenho do modelo será realizado pela média dos valores calculados nas iterações. Isso significa que o modelo é testado com diferentes partições dos dados. Essa técnica traz grande vantagem onde o número de amostras é muito pequeno (TUŠAR et al., 2017).

2.2.2 *Extreme Gradient Boosting*

Além de métodos *bagging*, como o Random Forest, existem métodos *boosting,* como o algoritmo XGB, também utilizados para finalidades de classificação e regressão. Nos métodos *boosting*, os erros do classificador modelado anteriormente são reduzidos de forma sequencial (TANHA et al., 2020).

O XGB foi introduzido por Chen e Guestrin em 2016. Seu processo de aprendizagem busca minimizar a impureza das predições de uma árvore. O método faz divisões nas árvores de decisão até a profundidade máxima estabelecida e após, "poda" a árvore de decisão, removendo as divisões em que não há ganho obtido (CARMONA et al., 2019). Assim, a complexidade da árvore de decisão é controlada em cada iteração e o número das folhas não é fixo. Além disso, uma técnica de aproximação é utilizada para descobrir a divisão em cada nó da árvore.

Tanha et al., (2020) explicam que, para descobrir a divisão em cada nó da árvore, são classificadas todas as instâncias a partir do valor de uma variável, e depois, uma busca linear é

realizada para encontrar a melhor divisão. O algoritmo é uma versão otimizada do *Gradient Boost*, que consiste em encontrar iterativamente o mínimo de uma função de perda que contém penalidades conhecidas como termos de regularização. A função do XGB é apresentada na Equação 7, onde, a cada iteração do algoritmo, a melhor árvore é selecionada:

$$Obj = \frac{-1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T \quad (7),$$

em que T é o número de nós "folha"; G_j e H_j são somas das estatísticas de gradiente de primeira e segunda ordem na função de perda sobre as amostras do *j*-ésimo nó folha, e λ e γ são os coeficientes de regularização da função (TANHA et al., 2020).

Também de acordo com Tanha et al. (2020), após a seleção de árvores realizada na iteração, os valores dos nós folha são calculados pelo uso de estatísticas de gradiente, a partir da Equação 8:

$$w_j^* = \frac{-G_j}{H_j+\lambda} \quad (8).$$

Ainda que conceitos básicos acerca do funcionamento do algoritmo sejam apresentados, eles são bastantes complexos. O XGB é conhecido como um modelo de caixa preta, denominado assim pela impossibilidade de um usuário compreender a estrutura composta por vários modelos únicos (SAGI; ROKACH, 2021).

Em relação ao desempenho do método, as medidas utilizadas são as mesmas para a classificação e regressão com o RF. Ainda, segundo Carmona et al. (2019), a técnica de *k-fold cross-validation* também pode ser utilizada no XGB para fornecer a estimativa do erro dos testes para cada modelo.

2.2.3 *Support Vector Machine*

O SVM é um método baseado que busca encontrar o hiperplano entre pontos de diferentes classes que os classifica corretamente. O classificador SVM promove, normalmente, uma classificação de alta qualidade, mesmo com as configurações padrão (DEMIDOVA et al., 2019).

Segundo Longjun et al. (2011), o conceito básico do método é que dados de entrada são refletidos em um espaço de alta dimensão onde um hiperplano de classificação ideal é formado. Nos casos que as classes não podem ser linearmente separadas no espaço de entrada, o método SVM transforma o espaço original de entrada em um espaço variável de dimensão superior. De acordo com Demidova et al. (2019), para encontrar o hiperplano de separação é necessário resolver um problema dual de busca de um *saddle point* da função Lagrangeana, que pode ser reduzido em um problema de programação quadrática. O problema é dado na Equação 9:

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^{S} \lambda_i + \\ \frac{1}{2}\sum_{i=1}^{S}\sum_{i=1}^{S} \lambda_i \lambda_\tau y_i y_\tau \kappa(z_i, z_\tau) \rightarrow min_\lambda, \\ \sum_{i=1}^{S} \lambda_i y_i = 0, 0 \leq \lambda_i \leq C, i = \overline{1,S} \end{cases} \quad (9),$$

sendo λ_i uma variável dual, z_i a amostra do conjunto de treino, y_i o número que caracteriza a classe da amostra z_i do conjunto de teste, $\kappa(z_i, z_\tau)$ é a função *kernel*, C é o parâmetro de regularização e S é o número de amostras no conjunto de dados de teste (DEMIDOVA ET AL., 2019).

É importante ressaltar que na classificação multiclasse pelo método SVM é utilizada uma abordagem conhecida como *one-versus-one* (OVO), de forma que os classificadores construídos treinam dados de duas classes. Conforme explica Tanha et al. (2020), a estratégia OVO consiste em criar um problema de classificação binária entre quaisquer duas classes e assim, introduzir $\mathrm{K} * \frac{(\mathrm{K}-1)}{2}$ problemas de classificação, sendo K o número de classes encontradas nos dados.

3 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo são descritos os procedimentos metodológicos conduzidos no estudo. Na primeira seção, 3.1, são apresentadas as informações acerca das bases de dados utilizadas. Na 3.2 são abordadas as ferramentas de softwares utilizadas. Por fim, a seção 3.3 indica as etapas gerenciais da pesquisa e a seção seguinte, 3.4, aborda as etapas metodológicas empregadas.

3.1 BASE DE DADOS

Os dados utilizados neste estudo fazem parte da PINTEC 2017 e foram coletados no site do Instituto Brasileiro de Geografia e Estatística (IBGE). A Pesquisa apura trienalmente as práticas adotadas pelas empresas brasileiras e as características que influenciam no seu comportamento inovador, além de alguns efeitos econômicos resultantes da inovação (IBGE, 2021). A Pesquisa tem como principais referências conceituais e metodológicas a terceira edição do Manual de Oslo e o modelo proposto pela Oficina de Estatística da Comunidade Europeia (*Statistical Office of the European Communities* - EUROSTAT), materializado na versão de 2014 da *Community Innovation Survey* (CIS).

A base obtida a partir da PINTEC 2017 apresenta, nas colunas, 426 indicadores de inovação que representam as atividades de inovação realizadas pelas empresas de 104 segmentos de estados brasileiros (linhas), selecionados de acordo com as divisões da Classificação Nacional de Atividades Econômicas - CNAE 2.0. Os 104 setores econômicos agrupam os indicadores de 116.962 empresas estudadas. Conforme IBGE (2021), a pesquisa possibilita à comunidade acadêmica estudar o desempenho e as características dos setores investigados, bem como permite ao setor público desenvolver e avaliar políticas públicas direcionadas.

3.2 FERRAMENTAS DE SOFTWARE

Neste estudo é utilizada a linguagem de programação *Python*, por meio da plataforma de código aberto Anaconda. As principais bibliotecas empregadas são *Pandas, Numpy, Sklearn e XGboost*. *Pandas* é conhecida como uma potente ferramenta de manipulação e análise de *dataframes* (PANDAS) e é utilizada principalmente na etapa de pré-processamento dos dados.

Numpy, também construída na linguagem *Python*, oferece suporte para lidar com vetores e matrizes e indexar os dados (NUMPY). *Sklearn ou Scikit-learn* é uma ferramenta eficiente para experimentação de algoritmos de *Machine Learning*, particularmente para aplicações em ciência experimental e de dados (SCIKIT-LEARN). XGboost também busca resolver problemas de ciência de dados, implementando algoritmos de *Machine Learning* sob a estrutura *Gradient Boosting* de maneira rápida e precisa (XGBOOST).

Outras bibliotecas também são utilizadas para melhor visualização dos resultados, como *Matplotlib*, que permite plotar figuras e gráficos diversos, e SNS ou *Seaborn*, que plota gráficos estatísticos a partir dos *dataframes* e matrizes, o que auxilia também na análise dos dados. Para computar os tempos de execução dos métodos implementados, também se fez uso do módulo *Python Datetime*.

3.3 ETAPAS GERENCIAIS

A execução da pesquisa compreende quatro fases principais: 1) pré-processamento dos dados; 2) definição do número de classes e balanceamentos, 3) seleção e classificação e 4) estimação. As duas primeiras etapas visam estruturar a base de dados, ainda que a etapa 2 forneça resposta para uma importante questão: quantos níveis de inovação podem ser identificados nos dados? A etapa 3, no entanto, responde ao seguinte problema de pesquisa: i) como classificar os setores em mais ou menos inovadores a partir das suas características? Por fim, a etapa 4 busca responder à seguinte pergunta: como se comporta o % de empresas que implementaram inovação a partir das características apresentadas? As respostas da terceira etapa darão subsídios para o entendimento de quais são as características que influenciam na classificação de amostras mais ou menos inovadoras e a da quarta, como melhor se estimaria o potencial de inovação de setores econômicos.

A Etapa 1, de pré-processamento dos dados, pretende normalizar os valores das variáveis pertencentes ao conjunto de dados descrito na seção 3.1. A Etapa 2 visa definir o número de níveis de inovação (classes) da base de origem, que geram novas bases de dados para serem utilizadas na Etapa 3. Para isso, são realizados experimentos com duas abordagens de definição de classes, detalhadas na próxima seção. Uma subetapa importante surge neste momento: o balanceamento das classes, cujo objetivo é tratar as classes desbalanceadas

oriundas da possível distribuição desigual das observações nos níveis de inovação definidos pela abordagem gerencial.

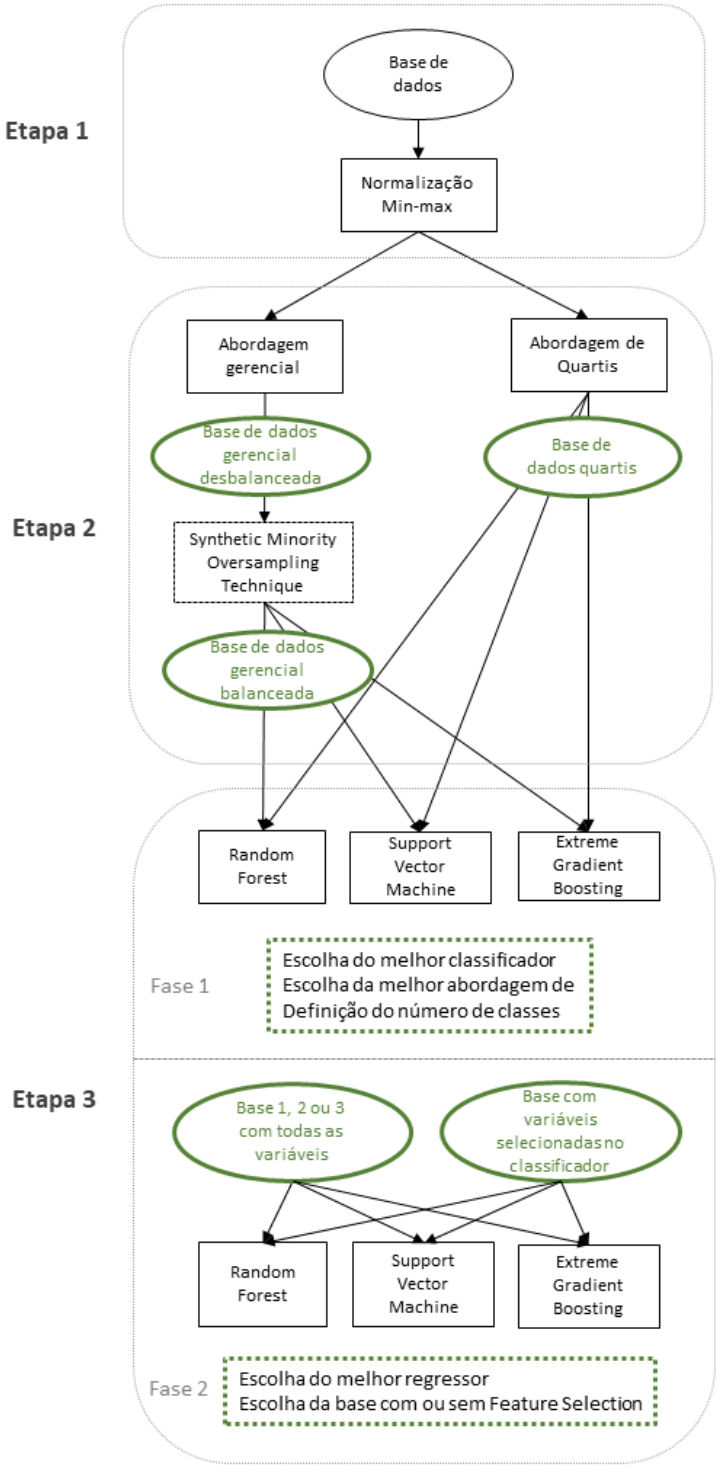
A Etapa 3 contempla duas fases. A primeira é a classificação, em que são implementados métodos para classificar categoricamente as amostras nos níveis de inovação das bases de dados 1, 2 e 3 resultantes da Etapa gerencial 2. Nesta etapa é que se verifica a melhor abordagem de definição de classes da Etapa 2. Além disso, se busca identificar as características mais relevantes para o processo de classificação a partir da seleção de variáveis embutida no modelo classificador de melhor desempenho, o que gera uma nova base de dados para utilização na Etapa 4.

Na Etapa 4 e última do fluxograma gerencial, se pretende estimar a taxa de inovação dos setores econômicos, utilizando os mesmos métodos da fase anterior – *Random Forest*, *Extreme Gradient Boosting* e *Support Vector Machine* - para classificar numericamente as observações. Isso possibilita compreender como a taxa de inovação se comporta a partir dos indicadores de inovação dos setores.

3.4 ETAPAS METOLÓGICAS

Para que os objetivos das etapas gerenciais sejam alcançados, a Figura 2 apresenta um fluxograma metodológico, detalhando os métodos utilizados em cada uma das fases, que serão exploradas nas seções seguintes.

Figura 2 - Fluxograma metodológico as etapas



3.4.1 Pré-processamento

A etapa de pré-processamento consiste em estruturar o conjunto de dados utilizado no estudo. Segundo Singh e Singh (2020), a etapa de pré-processamento dos dados é essencial para se obter bom desempenho na classificação. Por isso, identificadas variáveis de medidas diferentes nas bases de dados, utilizou-se a normalização dos valores para garantir que os recursos ou características tenham contribuição numérica igual, ainda que não signifique que tenham igual importância na decisão da classificação (SINGH; SINGH, 2020). No estudo, os dados são reescalados utilizando o método Min-Max, apresentado na Equação 10:

$$x' = \frac{x - min\,(x)}{max(x) - min\,(x)} \quad (10),$$

sendo x os valores dos dados.

Assim, evita-se que os algoritmos deem importância maior para variáveis de maior grandeza.

3.4.2 Definição do número de classes

A característica que distingue as amostras uma das outras e define, portanto, setores econômicos mais ou menos inovadores, equivale ao indicador "% percentual de empresas que implementaram inovações de produto e processo", tratado também neste estudo como "taxa de inovação" (número de empresas que inovaram em relação ao total de empresas) e que agora será abordado como característica Y. Os diferentes valores de Y no conjunto de dados não explicitam em quantas categorias as amostras podem ser separadas. Portanto, duas abordagens diferentes são propostas para definir o número de classes a ser utilizado na classificação das observações:

a) Abordagem gerencial, em que 3 classes são sugeridas: setores pouco, médio ou altamente inovadores. As observações são divididas em 3 classes, utilizando como critério os valores que apresentam para a característica Y: intervalos de 0 a 33%, 34% a 66% e 67% a 100%;

b) Abordagem de quartis, em que o conjunto de dados é dividido em quatro partes iguais, cada uma representando ¼ da amostra, representando a ordem das observações.

Por fim, a escolha da abordagem mais adequada para definir o número de classes será realizada na etapa de seleção de recursos e classificação das observações, considerando os resultados das métricas de desempenho dos métodos implementados.

3.4.2.1 Tratamento de classes desbalanceadas

Ao se atribuir classes às amostras do conjunto de dados utilizando a abordagem gerencial se geram classes de dados desbalanceadas, assim definidas quando o número de amostras de uma classe é maior do que nas demais. A etapa de tratamento de classes desbalanceadas é então proposta no *framework* metodológico pois, segundo Tanha et al. (2020), torna-se difícil para algoritmos de *Machine Learning* aprender os casos raros dos dados, porém importantes. Quando esses algoritmos são utilizados para prever saídas, os dados desequilibrados imputados no classificador produzem resultados inaceitáveis (SAWANGARREERAK; THANATHAMATHEE, 2020). Como exemplo de resultado insatisfatório, pode-se citar a acurácia de classificação, que acaba sendo inferior para as classes minoritárias.

De acordo com Sawangarreerak e Thanathamathee (2020), o tratamento de classes desbalanceadas é realizado aplicando métodos de geração de dados sintéticos, que fornecem uma distribuição de classes balanceadas a partir da adição ou remoção de amostras. O método escolhido para o tratamento é o *Synthetic Minority Oversampling Technique* (SMOTE), que consiste em gerar dados sintéticos para a classe minoritária junto com a linha entre os dados da minoria e a minoria vizinha mais próxima (SAWANGARREERAK; THANATHAMATHEE, 2020).

Segundo Zhu et al. (2019), a lógica utilizada no SMOTE é a do algoritmo k *nearest neighbours* (k-NN), que calcula os *k* vizinhos mais próximos de cada amostra minoritária x_i com a distância euclidiana como padrão. Então, para cada vizinho x_n selecionado aleatoriamente, novas amostras são geradas de acordo com a Equação 12:

$$x_{novo} = x_i + rand(0,1) * |x - x_n| \quad (12).$$

A escolha do método mais adequado para o tratamento das classes desbalanceadas também será realizada após a avaliação do desempenho dos métodos de classificação implementados na Etapa gerencial 3.

3.4.3 Seleção de recursos e classificação das observações

A classificação dos setores econômicos estaduais em níveis de inovação dá origem a um problema de classificação multiclasse com X características, dadas pelos indicadores de inovação, e um vetor Y de classes, que rotulam as observações. Três métodos são utilizados para a classificação: RF, SVM e XGB. É utilizado para cálculo do desempenho dos modelos o método *cross-validation k-fold* com *k* = 3 estratos de dados constituídos aleatoriamente a partir da base utilizada. Em cada estrato, o subconjunto de dados de treinamento é igual a k-1 partes para o treino e 1 parte para teste.

Os melhores parâmetros para os algoritmos são definidos experimentalmente. No RF e XGB são testados parâmetros da floresta, como o número de árvores, e parâmetros das árvores de decisão, como o número de variáveis utilizadas em cada divisão. A métrica da qualidade da divisão das árvores utilizada foi Gini. Para testar o modelo SVM, foram utilizados os parâmetros *default* do pacote. A avaliação da capacidade preditiva dos classificadores será realizada a partir das métricas de classe única *F1-score, precision e recall*. Como métrica geral, se avalia a acurácia dos modelos. As variáveis selecionadas durante o processo de aprendizado – técnica conhecida como *embedded* - do classificador com melhor desempenho serão utilizadas para restringir o subconjunto de características empregado na próxima etapa.

3.4.4 Estimação da variável-alvo

A taxa de inovação das observações será estimada utilizando também os três métodos, RF, SVM e XGB, sendo que agora Y são os valores contínuos da taxa de inovação, não mais discretizados em classes. Nesta etapa são utilizadas duas bases de dados: a original, com todas as variáveis (características) das amostras; e uma base com redução de variáveis, contendo apenas o subconjunto de características X selecionadas pelo modelo classificador com melhor desempenho. Os parâmetros experimentados nos métodos são os mesmos da fase de

classificação e seleção de variáveis. Os métodos são empregados para análise do comportamento do valor de Y das amostras a partir dos valores que apresentam as características X. A classificação numérica das observações consiste, portanto, em uma análise de regressão. Para separação dos conjuntos de treino e teste, também é utilizado o método *cross-validation k-fold* com $k = 3$ estratos de dados constituídos aleatoriamente a partir das bases utilizadas. As métricas MAE, RMSE e R^2 são usadas para avaliação e escolha do método de estimação que irá compor o *framework*.

4 RESULTADOS

Neste capítulo são apresentados os resultados das etapas do estudo, que consistem no pré-processamento dos dados e definição do número de classes da base utilizada, seleção de recursos e classificação pelos métodos *Random Forest, Extreme Gradient Boosting* e *Support Vector Machine* e estimação da variável-alvo pelos mesmos métodos.

4.1 DESENVOLVIMENTO DO FRAMEWORK

4.1.1 Pré-processamento e definição do número de classes

A base de dados utilizada, PINTEC 2017, apresenta indicadores de inovação de empresas pertencentes a alguns setores econômicos de estados brasileiros selecionados. As observações das bases são referenciadas, portanto, nas análises deste capítulo, como "setores-estado". A taxa de inovação de um setor econômico, utilizada neste trabalho como variável Y e a partir de qual são definidas as classes (níveis de inovação) das observações, consiste em:

$$\text{Taxa de inovação } Y = \frac{EI}{TE}$$

EI: número de empresas do setor que desenvolveram inovações de produto ou de processo;
TE: número total de empresas avaliadas do setor.

Tendo em vista a abordagem gerencial de definição do número de classes, foram definidas três classes inovação para os setores-estado:

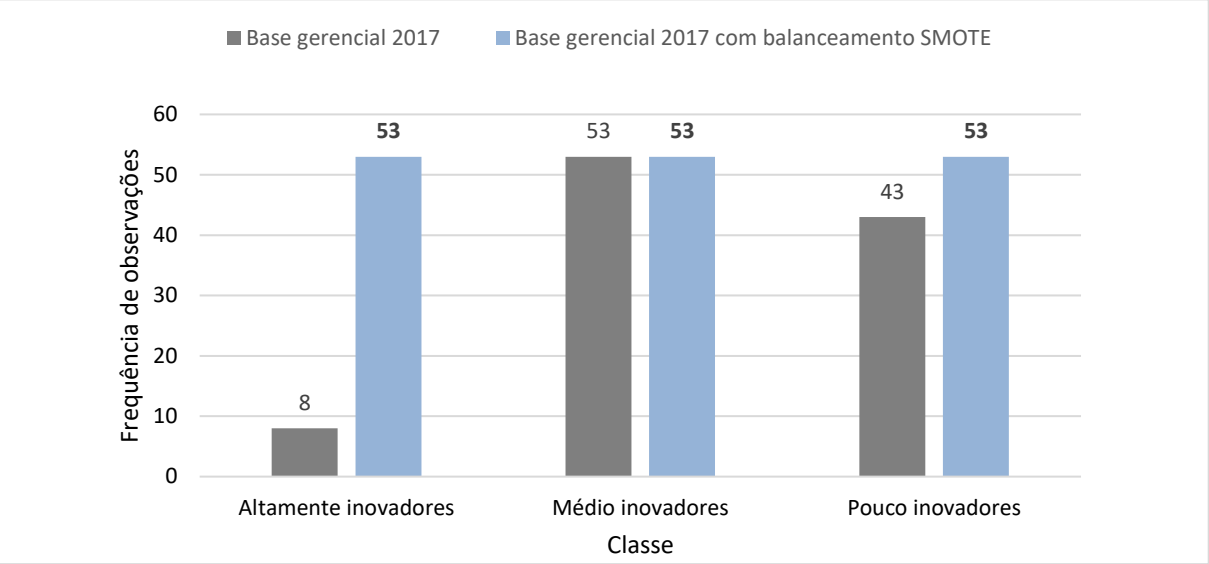
- Classe 1: setores com valor de Y de 0,66 a 1, o que significa que mais de 66% de empresas apresentam inovações de produto ou processo;

- Classe 2: setores com Y de 0,33 a 0,65, representando um percentual entre 33% e 65% de empresas com inovações;

- Classe 3: setores com Y inferior a 0,32, ou seja, menos de 32% das empresas geraram inovações.

A base PINTEC 2017 contém 104 observações e 426 variáveis. Os valores das variáveis foram normalizados utilizando o método Min-Max. A Figura 3 apresenta o número de observações alocadas em cada classe após a definição das classes pela abordagem gerencial, e o número de observações alocadas após o balanceamento com o método SMOTE.

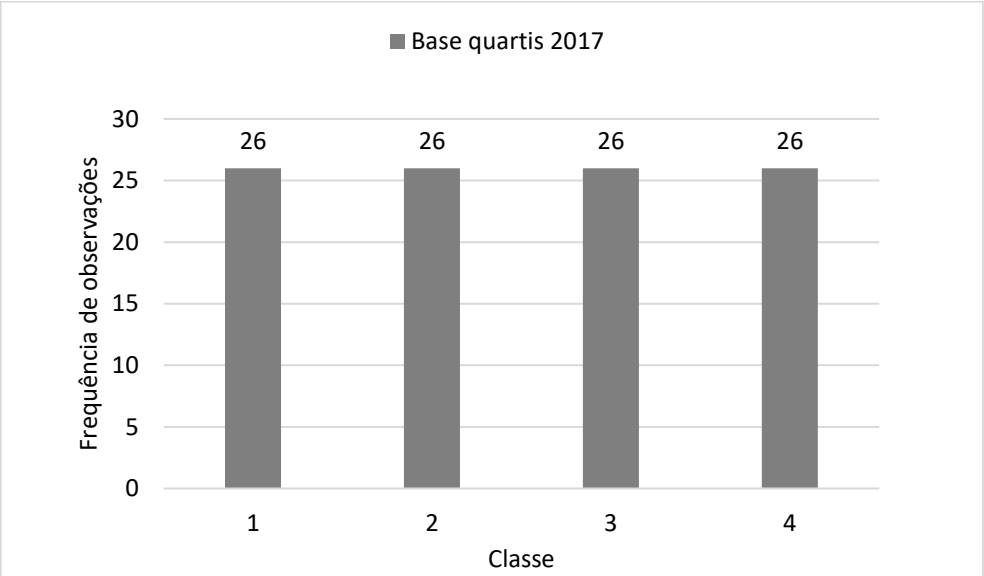
Figura 3 – Observações na Base gerencial desbalanceada X Observações na Base gerencial balanceada pelo método SMOTE (Base 2017)



Nota-se que, separando as 104 observações em 3 classes – setores-estado altamente, médio e pouco inovadores, a frequência de amostras na Classe 3 na Base gerencial desbalanceada é muito superior às Classes 1 e 2. Após o balanceamento das classes minoritárias pelo método SMOTE, foram geradas 55 amostras sintéticas, que resultou, portanto, em uma nova base de dados que será utilizada nas etapas seguintes – Base gerencial balanceada, com 159 observações. Conforme mostra a Figura 3, a distribuição passa a ser igual para as 3 classes.

Utilizando a segunda abordagem proposta nesta etapa - definição do número de classes da base por abordagem de quartis, as 104 observações são distribuídas em quatro classes. A base gerada será identificada neste capítulo como Base quartis. A Figura 4 ilustra a separação das observações de forma equilibrada em 4 níveis de inovação.

Figura 4 - Definição do número de classes com abordagem por quartis



4.1.2 Classificação:

Para implementação dos métodos RF, XGB e SVM na Base gerencial balanceada e Base quartis, as técnicas *Grid Search* e *Randomized Search* foram utilizadas para prévia definição dos parâmetros de número de variáveis na divisão e métrica de divisão das árvores, no caso de RF e XGB, e de Kernel, Regularização C e Gamma, no método SVM. Nos modelos RF e XGB se utilizaram as técnicas para definição de um ponto de partida da parametrização do número de árvores, e em seguida, outros testes foram realizados simulando valores maiores e menores para esse parâmetro.

A Tabela 1 a seguir apresenta os parâmetros utilizados em cada uma das abordagens.

Tabela 1 - Parâmetros testados

Método	Parâmetros	Candidatos
RF	Nº árvores (n)	100; 200; 300; 350; 400
	Nº variáveis na divisão	√(nº variáveis)
	Métrica de divisão	Gini
XGB	Nº árvores	100; 200; 300; 400; 450
	Nº variáveis na divisão	10
	Métrica de divisão	Gini
SVM	Kernel	Linear
	Regularização C	1
	Gamma	10-1

Fonte: a autora (2021).

Os métodos de classificação foram aplicados em três bases de dados: Base gerencial desbalanceada, Base gerencial balanceada e Base quartis. A Base gerencial desbalanceada foi testada a fim de verificar a efetividade do balanceamento das amostras nas classes previamente à implementação dos métodos de classificação.

Na separação dos conjuntos de treino e teste dos modelos foi utilizado o método de validação cruzada com *k-fold* = 3, sendo o que os dados de treino representam 2 estratos da base e os dados de teste, 1 estrato. Se utilizou como critério que amostras das 3 classes, no caso das Bases gerenciais, e das 4 classes, na Base quartis, deveriam estar presentes em todos os conjuntos de dados. Sendo assim, a Tabela 2 apresenta a média das métricas calculadas com 3 estratos de dados.

Tabela 2 - Resultados dos métodos de classificação

PINTEC 2017													
		Base gerencial desbalanceada				Base gerencial balanceada				Base quartis			
Mét.	Par.	A	P	R	F1	A	P	R	F	A	P	R	F1
SVM		0,81	0,81	0,81	0,79	0,82	0,86	0,82	0,81	0,69	0,68	0,69	0,68
XGB	n= 100	0,69	0,69	0,69	0,65	0,85	0,86	0,85	0,85	0,53	0,59	0,53	0,53
	n= 200	0,69	0,69	0,69	0,65	0,80	0,82	0,80	0,81	0,53	0,59	0,53	0,53
	n= 300	0,69	0,69	0,69	0,65	0,83	0,83	0,83	0,82	0,53	0,54	0,53	0,52
	n= 400	0,69	0,69	0,69	0,65	0,90	0,91	0,90	0,90	0,53	0,54	0,53	0,52
	n= 450	0,69	0,69	0,69	0,65	0,88	0,90	0,88	0,87	0,53	0,54	0,53	0,52
RF	n= 100	0,89	0,88	0,89	0,87	0,85	0,87	0,85	0,85	0,73	0,73	0,73	0,70
	n= 200	0,81	0,74	0,81	0,77	0,85	0,87	0,85	0,85	0,77	0,74	0,77	0,74
	n= 300	0,81	0,74	0,81	0,77	0,87	0,91	0,87	0,87	0,81	0,79	0,81	0,78
	n= 350	0,81	0,74	0,81	0,77	0,91	0,92	0,91	0,90	0,74	0,76	0,74	0,72
	n= 400	0,81	0,74	0,81	0,77	0,85	0,89	0,85	0,85	0,70	0,75	0,70	0,69

Legenda: A = Accuracy; P = Precision; R = Recall; F1=F1-score.
Fonte: a autora (2022).

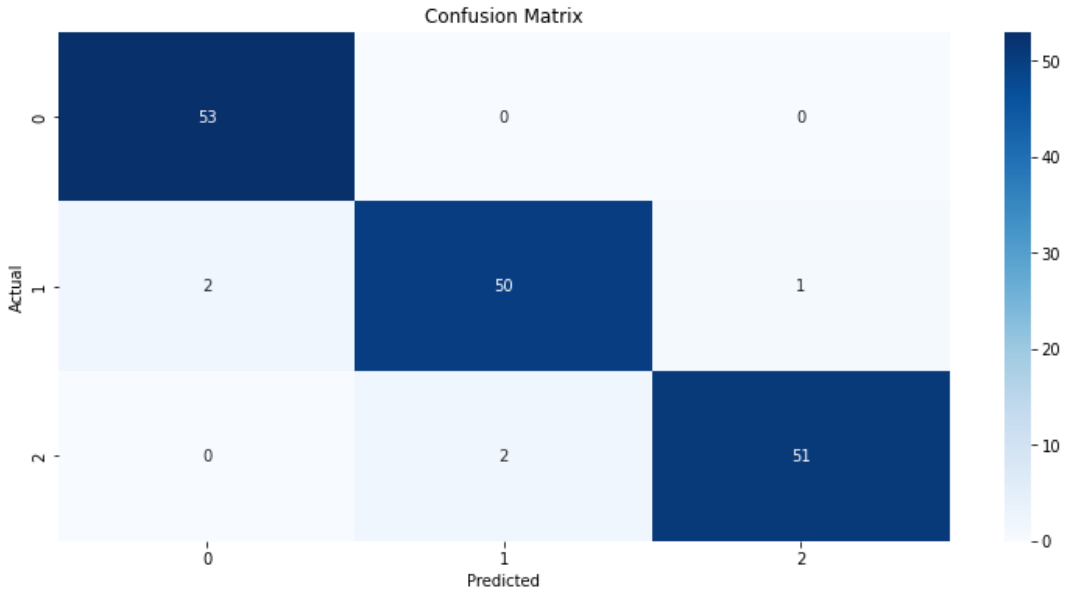
Na análise comparativa dos testes com os métodos SVM, XGB e RF aplicado na Base gerencial desbalanceada e Base gerencial balanceada, nota-se que os resultados das métricas que avaliam os classificadores, aplicados nas bases desbalanceadas, são inferiores em relação

ao uso das bases balanceadas. A inclusão no *Framework* da etapa de balanceamento das classes se justifica, portanto, uma vez que os métodos apresentam dificuldade em identificar casos especiais – ou minoritários – originais da base utilizada. No caso da separação das amostras em classes com a abordagem de quartis, o balanceamento de classes é dispensado, uma vez que os dados já ficam equidistribuídos.

Os resultados das métricas de avaliação dos métodos SVM, XGBoost e RF na Base gerencial balanceada indicam ser essa a melhor abordagem para a base de dados utilizada no desenvolvimento do framework: separação das amostras nas 3 (três) classes definidas gerencialmente, com o subsequente balanceamento pelo método SMOTE - em detrimento à distribuição das amostras em classes iguais – abordagem por quartis.

Dentre os três métodos, nota-se que o método de classificação multiclasse com melhor desempenho foi o RF, aplicado na Base gerencial balanceada. Os resultados indicam 91% de acurácia e Recall, 90% na taxa F1-score e Precision de 92%. O parâmetro utilizado, nesse caso, é um número de árvores igual a 350. Considerando o resultado superior obtido na classificação pelo método RF aplicado na Base gerencial balanceada, convém apresentar, na Figura 5, um exemplo de Matriz de Confusão obtida na classificação de um dos estratos de dados utilizados.

Figura 5 - Matriz de confusão da classificação pelo método RF – Base gerencial balanceada 2017



Fonte: a autora (2022).

Para a classe 1 – setores-estado com 66% de empresas inovadoras de produto ou processo, 100% das amostras foram classificadas corretamente, ou seja, 100% da predição com verdadeiros positivos. Na classe 2, 3 (três) amostras foram classificadas como tal incorretamente, sendo 2 (duas) delas classificadas como classe 1 e 1 (uma) como classe 3; já na classe 3, 2 (duas) amostras foram classificadas incorretamente como classe 2. Esses resultados demonstram que, no geral, o modelo RF é bom classificador, ainda que exista uma tendência de elevação do desempenho de inovação dos setores econômicos com uma taxa de inovação inferior.

4.1.3 Variáveis selecionadas: (gerencial + SMOTE + RF)

Considerando a etapa metodológica com melhor desempenho nas métricas de avaliação da classificação (Base gerencial 2017 + balanceamento SMOTE + RF com n=350), a seguir, na Tabela 3, são apresentadas as 20 variáveis mais relevantes na classificação multiclasse dos dados, com a dimensão de análise a qual pertencem, considerando para cálculo de importância das variáveis o Índice de Gini.

Tabela 3 - Variáveis mais relevantes na classificação

Variável	Importância (%)	Nome do indicador
X423	3,45%	% empresas que publicaram relatórios de sustentabilidade
X402	3,35%	% empresas que reduziram o impacto ambiental com redução da contaminação do solo, da água, de ruído ou do ar com médio grau de importância
X63	3,26%	% empresas que implementaram inovações e que adquiriram treinamento e indicam alta importância para isso
X92	3,16%	% de pesquisadores pós-graduados ocupados em atividades internas de P&D nas empresas que implementaram inovações
X118	3,15%	% empresas que implementaram inovações com impacto alto causado em ampliação da participação da empresa no mercado
X377	3,13%	% empresas que implementaram inovações de produto e processo e inovações organizacionais e/ou de marketing em técnicas de gestão
X218	2,37%	% empresas que implementaram inovações e empregaram fontes de informação - Instituições de testes, ensaios e certificações - do Brasil
X229	2,09%	% empresas que implementaram inovações com relações de cooperação com clientes ou consumidores com baixa importância da parceria
X60	2,07%	% empresas que implementaram inovações de produto e/ou processo que adquiriram máquinas e equipamentos e indicam alta importância para isso

X418	2,01%	% empresas que implementaram inovações, reduziram o impacto ambiental, por fator de contribuição para introdução de inovações ambientais - ações voluntárias
X178	1,94%	% empresas que implementaram inovações com emprego de informação das fontes externas - consultores - de importância alta
X334	1,93%	% empresas que implementaram inovações, por alto grau de importância dos problemas e obstáculos apontados
X235	1,86%	% empresas que implementaram inovações com relações de cooperação com concorrentes com baixa importância da parceria
X4	1,66%	% empresas que implementaram inovações de produto
X288	1,43%	% empresas que implementaram inovações que receberam apoio do governo - financiamento a projetos de P&D sem parceria com universidades
X371	1,42%	% empresas que não implementaram inovações de produto ou processo, mas inovações organizacionais e/ou de marketing em técnicas de gestão
X180	1,40%	% empresas que implementaram inovações com emprego de informação das fontes externas - consultores - de importância baixa
X372	1,38%	% empresas que não implementaram inovações de produto ou processo e sem projetos, mas inovações organizacionais e/ou de marketing em técnicas de gestão ambiental
X420	1,18%	% empresas que implementaram inovações, reduziram o impacto ambiental, por fator de contribuição para introdução de inovações ambientais - elevados custos
X336	1,11%	% empresas que implementaram inovações por médio grau de importância de riscos econômicos excessivos

Fonte: a autora (2022).

Na Figura 6 é apresentada a soma das importâncias por bloco de análise, conforme apresentado na estrutura da PINTEC, das 65 variáveis que representam juntas 80% de relevância na classificação das amostras da base 2017, a fim de subsidiar a análise dos tipos de indicador que mais contribuem para a taxa de inovação.

Figura 6 - Importância das variáveis 2017 agrupada por blocos de análise

Fonte: a autora (2022).

Outras 298 variáveis representaram, conjuntamente, 19,74% de importância durante a classificação dos setores-estado em níveis de inovação pelo método RF e 63 variáveis não tiveram importância no processo de classificação dos dados. As variáveis que compõem cada dimensão da Figura 6 são apresentadas no Anexo II.

4.1.4 Estimação

Na etapa de estimação dos valores da Taxa de inovação, os métodos RF, XGB e SVM são utilizados considerando os valores contínuos de Y, e não mais Y como classe. Na regressão, duas Bases de dados são utilizadas: a Base sem *Feature Selection*, que apresenta 426 variáveis, e a Base restrita ao subconjunto de variáveis selecionado durante a etapa de classificação, reduzida para 65 variáveis. Ambas contêm 104 observações. As variáveis utilizadas na Base com *Feature Selection* representam 80% da importância na classificação das amostras pelo RF – modelo classificador de melhor desempenho.

Nos métodos RF e XGB, aqui utilizados como regressores, foram utilizados os parâmetros com melhor desempenho na etapa de classificação: RF com número de árvores (n) = 350 e XGB, n = 400. Para o SVM, parâmetros *Default*. Na divisão dos conjuntos de dados, também foi utilizado o método de validação cruzada com *k-fold* = 3, de forma que k-1 estratos foram utilizados como conjunto de treinamento e um dos estratos como conjunto de teste. Para

análise dos resultados da estimação, são consideradas as médias das métricas MAE, RMSE e R^2 calculadas nas iterações com os estratos utilizados no teste dos modelos.

Na Tabela 4 a seguir são apresentados os resultados dos métodos para as duas Bases, com e sem *Feature Selection*.

Tabela 3 - Estimação de Y nas Bases com e sem *Feature Selection*

Base 2017						
	Com *Feature Selection*			Sem *Feature Selection*		
Método	MAE	RMSE	R^2	MAE	RMSE	R^2
RF	0,0313	0,0539	0,9164	0,0312	0,0592	0,8995
XGBoost	0,0385	0,0717	0,8524	0,0341	0,0575	0,9051
SVM	0,0497	0,0693	0,8622	0,0487	0,0696	0,8610

Fonte: a autora (2022).

Dentre todos os métodos de regressão aplicados nas Bases com e sem *Feature Selection*, o melhor desempenho, considerando para análise o valor de R^2 de 91,64%, foi do RF na Base com *Feature selection* (104 amostras e 65 variáveis). A regressão aplicada nessa Base, que teve a dimensionalidade reduzida, apresentou menores taxas de erro em relação à Base com todas as variáveis. Entretanto, pelo método XGB, a taxa de erro aumentou em relação à regressão aplicada na Base sem *Feature Selection*.

Considerando os valores estimados de Y a partir do RF aplicado na Base com variáveis selecionadas, é apresentada, na Tabela 5, a média, mediana, mínimos e máximos da diferença entre os valores estimados e os valores reais de Y em cada uma das classes. A estatística descritiva dos valores absolutos preditos também é demonstrada.

Tabela 4 - Variação entre valores reais e valores estimados de Y

Base 2017 com Feature Selection							
Método	Medida	Classe 1		Classe 2		Classe 3	
		Y Pred - Y	Y pred	Y Pred - Y	Y pred	Y Pred - Y	Y pred
RF	Média	-0,0794	0,7197	-0,0073	0,4196	0,0118	0,2182
	Mediana	-0,0269	0,7149	-0,0111	0,3999	0,0056	0,2200
	Mínimo	-0,3049	0,6595	-0,1115	0,3199	-0,0489	0,0794
	Máximo	0,0395	0,7943	0,1545	0,6717	0,1774	0,4359

Fonte: a autora (2022).

É possível observar que a estimação das taxas de inovação Y mais altas (amostras da Classe 1) é menos assertiva em relação às taxas menores (Classes 2 e 3). O valor de Y chega a ser estimado 30% menor do que o valor real, conforme demonstra o valor de mínima diferença. Isso ocasionaria, considerando apenas os valores de Y, a conclusão de que a observação é de classe 2, ou seja, menos inovadora do que os indicadores apontam. O ranking dos setores econômicos por taxa de inovação estimada e a respectiva classe em que foram alocados (sendo classe 1- alta inovação, 2 – média inovação e 3 – pouca inovação) é apresentado no Anexo III.

Nas Classe 2 e 3, a média e mediana estão mais próximas de zero, o que, no geral, indica uma estimação de Y assertiva. Entretanto, análises posteriores são necessárias, uma vez que a predição do valor Y poderia alocar observações em um nível maior ou menor de inovação.

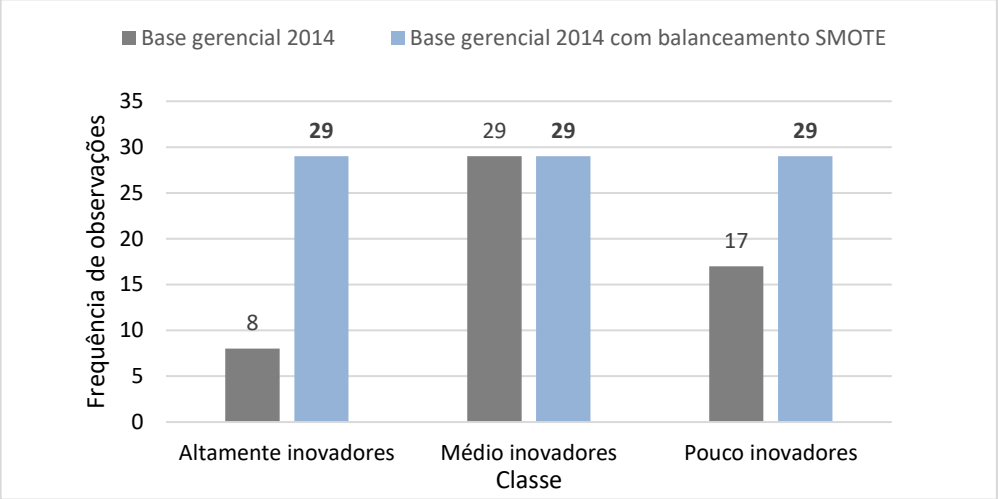
4.5 APLICAÇÃO DO *FRAMEWORK* EM OUTRA BASE DE DADOS – PINTEC 2014

A partir dos resultados obtidos durante o desenvolvimento de cada etapa metodológica proposta neste estudo, o *Framework* é aplicado em uma nova base de dados de indicadores de inovação, a PINTEC 2014. A principal diferença entre a base utilizada na construção do *Framework* (PINTEC 2017) e a Base 2014 é que, agora, as observações da base de dados são setores econômicos não desagregados nos estados brasileiros, o que resulta em um número de observações menor. A Base 2014 contêm 54 observações e 374 variáveis.

As etapas do *Framework*, portanto, são as seguintes: definição de classes por abordagem gerencial + balanceamento SMOTE + classificação e seleção pelo método RF + estimação de Y pelo método RF.

A Figura 7 apresenta o comparativo entre o número de observações da Base gerencial desbalanceada e da Base gerencial balanceada, após geradas 33 amostras sintéticas pelo método SMOTE.

Figura 7 - Observações na base de dados com definição de classes por abordagem gerencial X Observações após balanceamento SMOTE (Base 2014)



Fonte: a autora (2022).

O balanceamento das classes definidas por abordagem gerencial resultou, portanto, em uma nova base de dados, que será utilizada nas etapas seguintes: a Base 2014 passa de 54 para 87 observações (Base gerencial balanceada). Visto que o número de observações da Base 2014 é inferior à Base 2017, são realizados testes com todos os números de árvores simulados anteriormente, conforme apresenta a Tabela 6.

Tabela 5 - Resultados da classificação pelo método RF

		Base 2014			
Método	Parâmetro	A	P	R	F
RF	n= 50	0,82	0,81	0,82	0,80
	n= 100	0,82	0,79	0,82	0,78
	n= 200	0,68	0,60	0,68	0,62
	n= 300	0,73	0,60	0,73	0,64
	n= 350	0,73	0,60	0,73	0,64
	n= 400	0,68	0,62	0,68	0,61

Legenda: A = Accuracy; P = Precision; R = Recall; F1=F1-score.
Fonte: a autora (2022).

Na base de 2014, os resultados de melhor desempenho na classificação indicam 82% de acurácia e Recall, 81% na taxa *Precision* e 80% na *F1-score*. Os parâmetros utilizados no teste de melhor desempenho são 50 árvores, enquanto na base PINTEC 2017, eram 350. Tanto a queda no desempenho do método de classificação quanto o ajuste do parâmetro pode ser justificado pelo número diferente de amostras nas bases (2014 com 87 e 2017 com 159

amostras). O estudo não objetiva necessariamente fixar a metodologia na amostra utilizada, mas estabelecer um processo para análise de indicadores de inovação.

Durante a classificação das amostras da Base gerencial balanceada 2014, 78 variáveis representaram, juntas, cerca de 80% de relevância no método RF. 118 variáveis representam conjuntamente 20,18% de importância na classificação das amostras e 179 não tiveram importância. A Tabela 7 apresenta as 20 variáveis selecionadas mais relevantes.

Tabela 6 - Variáveis mais relevantes na classificação pelo método RF

Variável	Importância (%)	Nome do indicador
Q145	4,86%	Faixa de participação das inovações nas vendas internas das empresas que implementaram inovação de produto - 10% a 40%
Q98	4,51%	Impacto causado nas empresas que implementam inovações - Melhoria da qualidade dos produtos baixa ou não relevante
Q105	4,36%	Impacto causado nas empresas que implementam inovações -Ampliação da participação da empresa no mercado alta
Q143	4,25%	Impacto causado nas empresas que implementam inovações - Enquadramento em regulações e normas padrão baixa ou não relevante
Q127	3,33%	Impacto causado nas empresas que implementam inovações - Redução do consumo de energia média
Q260	2,05%	% empresas que implementaram inovações com relação de cooperação com outra empresa do grupo para outras atividades de cooperação
Q242	1,86%	% empresas que implementaram inovações com relação de cooperação com concorrentes do Brasil
Q51	1,78%	"% empresas que implementaram inovações com Aquisição externa de Pesquisa com importância média
Q288	1,65%	% empresas que não implementaram inovações e sem projetos por razão de escassez de fontes apropriadas de financiamento de importância alta
Q241	1,54%	% empresas que implementaram inovações com relação de cooperação com fornecedores do exterior
Q216	1,54%	% empresas que implementaram inovações com relação de cooperação com clientes ou consumidores de importância baixa ou não relevante
Q84	1,50%	Número de pessoas ocupadas com dedicação parcial em P&D nas empresas que implementaram inovações
Q198	1,38%	% empresas que implementaram inovações com fonte de informação empregadas - Consultoria do exterior
Q245	1,37%	% empresas que implementaram inovações com relação de cooperação com outra empresa do grupo do exterior

Q63	1,35%	% empresas que implementaram inovações com Treinamento de importância média
Q349	1,33%	% empresas que implementaram inovações de produto com obstáculos apontados -escassez de serviços técnicos externos adequados - de importância média
Q164	1,29%	% empresas que implementaram inovações com fontes de informação empregadas - Concorrentes - Importância baixa ou não relevante
Q328	1,28%	% empresas que implementaram inovações de produto com obstáculos apontados - rigidez organizacional - de importância média
Q115	1,27%	% empresas que implementam inovações com aumento da flexibilidade da produção média
Q145	4,86%	% empresas que implementaram inovações com fontes de informação empregadas - Consultoria - Importância baixa ou não relevante

Fonte: a autora (2022).

Para fins de análise gerencial comparativa dos indicadores de inovação selecionados na classificação das amostras das Bases 2017 e 2014, a Figura 8 apresenta as 78 variáveis selecionadas e agrupadas em blocos de análise.

Figura 8 - Importância das variáveis 2014 agrupada por blocos de análise

Fonte: a autora (2022).

É possível observar que os indicadores de Impacto das inovações, com importância de 6,9% na classificação dos setores econômicos como altamente, médio ou pouco inovadores na

PINTEC 2017, aparecem com relevância de 25,9% na classificação da base PINTEC 2014. Segundo IBGE (2016), a dimensão de Impacto das Inovações busca identificar nas empresas fatores como a melhoria da qualidade ou ampliação da gama de produtos ofertados, ampliação da participação da empresa no mercado ou abertura de novos, aumento de capacidade produtiva, redução de custos ou mesmo aspectos relacionados ao meio ambiente, saúde, segurança, além do enquadramento em regulamentações e normas. A redução da relevância destes indicadores nos últimos anos pode ser explicada pela inclusão de novos blocos de análise na Pesquisa mais recente, como o de Sustentabilidade e inovação ambiental, e que, a título de exemplo, fica em segundo lugar como bloco de indicadores mais importante na análise do nível de inovação dos segmentos econômicos em 2017.

O grupo de indicadores de Problemas e obstáculos à inovação, conforme IBGE (2020), monitora os fatores pelos quais a empresa não desenvolveu atividades inovativas, não obteve os resultados esperados ou dificultou a implementação de projetos, sejam de eles de custos, riscos, fontes de financiamento apropriadas, rigidez organizacional, deficiências técnicas, escassez de serviços técnicos externos adequados, falta de pessoal qualificado, falta de informações sobre tecnologia e sobre os mercados, escassas possibilidades de cooperação ou dificuldade para se adequar a padrões, normas e regulamentações. Em 2014, esses indicadores aparecem com 9% de importância na análise e, em 2017, a dimensão de Problemas e obstáculos à inovação indica que essas variáveis passam a representar uma importância de 18,7%.

Ainda que, implícito neste estudo, haja interesse na identificação dos fatores mais relevantes para a classificação dos setores econômicos como pouco, médio ou altamente inovadores, é importante considerar que se torna difícil correlacionar as variáveis selecionadas na base utilizada para construção do *Framework*, PINTEC 2017, e na base em que o *Framework* foi aplicado, PINTEC 2014. Na perspectiva de que indicadores podem mudar de uma base para outra, assim como diferem as observações (a exemplo da base 2017 serem indicadores de empresas de setores econômicos estaduais e na base 2014, de setores econômicos nacionais), a interpretação conceitual dos resultados deve ser adaptada a partir do contexto em que é realizado o estudo. Por isso a importância de se ter um *Framework* de análise de dados robusto.

A estimação dos valores de Y utiliza a Base gerencial com a *Feature Selection* resultante da etapa de classificação. Assim, a Base 2014 passa de 374 para 78 variáveis. Nesta etapa também foram mantidos os parâmetros de melhor desempenho na classificação dos dados da

Base: RF com n=50, assim como o método de validação cruzada com *k-fold* = 3 para a separação dos conjuntos de dados de treino e teste. A Tabela 8 demonstra os resultados da estimação.

Tabela 7 - Estimação de Y com a Base com *Feature Selection* pelo método RF

	Base 2014 com *Feature Selection*		
Método	MAE	RMSE	R^2
RF	0,0707	0,0899	0,7053

Fonte: a autora (2022).

A perda de desempenho no ajuste do modelo de regressão de 91,64% (Base 2017) para 70,53%, considerando a métrica R^2, pode estar atribuída ao número de observações ser consideravelmente menor na Base 2014 em relação à base de construção do *Framework*, sendo essa uma limitação identificada neste estudo. Ainda, a natureza diferente das observações em cada base, setores econômicos desagregados em estados *versus* setores agrupados a nível nacional, também pode contribuir para a dificuldade do modelo na predição da taxa de inovação das amostras.

A Tabela 9 apresenta algumas estatísticas descritivas da diferença entre a predição e os valores reais de Y, e dos valores absolutos preditos, para as observações de cada uma das classes, utilizando o modelo regressor com o método RF.

Tabela 8 - Variação entre valores reais e valores estimados de Y pelo método RF

		Base 2014 com *Feature Selection*					
Método	Medida	Classe 1		Classe 2		Classe 3	
		Y Pred - Y	Y pred	Y Pred - Y	Y pred	Y Pred - Y	Y pred
RF	Média	-0,1488	0,6042	-0,0048	0,4343	0,0698	0,3395
	Mediana	-0,1623	0,6157	0,0073	0,4107	0,0676	0,3413
	Mínimo	-0,2484	0,4712	-0,1714	0,3188	-0,0020	0,2626
	Máximo	-0,054	0,7161	0,1360	0,6406	0,1314	0,4502

Fonte: a autora (2022).

Em relação à estimação dos valores de Y na Base 2017, a partir de qual o *Framework* foi desenvolvido, nota-se um comportamento semelhante do modelo RF como regressor ao estimar a taxa de inovação das observações da Base 2014 para as classes. O valor estimado de Y para as amostras da Classe 1 – setores econômicos altamente inovadores, é constantemente menor do que o valor de Y real, o que pode ser explicado pelo número reduzido de amostras

dessa classe em ambas as bases de dados. Para a Classe 2, a diferença entre a taxa de inovação dos setores econômicos médio inovadores estimada e real é relativamente baixa, conforme indicam a média e a mediana. Nota-se ainda que nas Classes 2 e 3 os valores máximos da estimação tendem a elevar o nível de inovação das observações, especialmente na Classe 3.

Os resultados desta etapa permitem inferir que o regressor auxilia na estimação da taxa de inovação (Y) dos setores econômicos, entretanto, uma análise gerencial por parte do decisor se faz necessária para prever o potencial de inovação dos segmentos, considerando os indicadores que têm relevância no classificador.

5 CONSIDERAÇÕES FINAIS

O presente estudo teve como objetivo propor um *Framework* de *Data Analytics* para análise de indicadores de inovação, a partir da utilização da base de dados da PINTEC 2017 no desenvolvimento das etapas metodológicas que compõem a sua estrutura. Dando início à análise dos dados, na etapa denominada como Pré-processamento, foi aplicada a técnica Min-Max para normalização dos valores das variáveis, dada as grandezas diferentes dos valores apresentados nos indicadores de inovação (variáveis) dos setores econômicos (observações).

Previamente à etapa de Classificação, duas abordagens distintas foram propostas para definir as classes a que as observações pertencem, uma vez que a base de dados original não contempla tal distinção: Abordagem gerencial, que estabelece 3 classes para as observações, e Abordagem por quartis, que as separa em quatro classes. Entendendo a necessidade de as classes estarem balanceadas, visto a característica multiclasse do problema e de um número relativamente pequeno de observações na base utilizada para a construção do *Framework*, uma subetapa de balanceamento das classes foi proposta utilizando-se da técnica *Synthetic Minority Oversampling Technique*, como passo subsequente à Abordagem gerencial.

Para a Classificação categórica das observações e Estimação da variável-alvo (taxa de inovação) contínua, variável essa de qual se originaram as classes, os métodos *Random Forest*, *Extreme Gradient Boosting* e *Support Vector Machine* foram testados.

Para avaliar a efetividade do *Framework* proposto, as abordagens utilizadas em cada etapa foram comparadas mediante os resultados das etapas de Classificação e Estimação, utilizando, respectivamente, métricas de *Accuracy*, *Precision*, *Recall* e *F1-Score* para os modelos classificadores e *Mean Absolute Error*, *Root Mean Squared Error* e Coeficiente de determinação (R^2) para os modelos regressores.

Dentre os resultados obtidos, observou-se que o classificador RF obteve um desempenho mais satisfatório em relação aos métodos XGB e SVM. Na avaliação de AC, P, R e F1, também se percebeu que a abordagem gerencial para definição das classes junto ao balanceamento pelo método SMOTE se mostra mais efetiva como etapa prévia à classificação, uma vez que o desempenho dos três métodos piorou consideravelmente quando os testes foram realizados na base de dados em que as classes estão desbalanceadas. Quando se utiliza a Base de dados com definição de classes a partir dos quartis, apesar das classes ficarem naturalmente balanceadas, os modelos classificadores, de maneira geral, apresentam um desempenho ainda

pior. Isso indica a importância de uma análise mais sensível das observações da base de dados utilizada ao se definir a classe a que pertencem quando essa informação não está originalmente disponível.

Em análise dos resultados da etapa de Estimação, quando os modelos regressores dos métodos RF, XGB e SVM foram utilizados, constatou-se que o RF apresentou melhor *performance* na predição da variável-alvo. Neste momento, também foi possível comparar a eficácia do uso da seleção de variáveis da etapa de classificação para reduzir o número de indicadores da base em relação à utilização de todos eles. Considerando que a implementação do RF apresentou as menores medidas de erro e o melhor ajuste do modelo regressor nos testes, observa-se que a base de dados utilizada é aquela que contêm apenas as variáveis que representam até 80% de relevância na classificação das amostras, identificadas na seleção de variáveis embutida no classificador. Entretanto, ao se analisar os demais regressores, o método XGB aplicado na base com um número reduzido de variáveis resulta em aumento das taxas de erro da regressão, em relação à base com todas as variáveis. No SVM, o desempenho do modelo se mantém praticamente constante em ambas as bases de dados.

A partir dos resultados obtidos, o *Framework* proposto é constituído, portanto, da Normalização dos valores com o método *Min-Max*, Definição de classes por Abordagem gerencial seguida de balanceamento pelo método SMOTE, Classificação das observações e Seleção de variáveis pelo método *Random Forest* e Estimação da variável-alvo também pelo método RF, utilizando apenas as variáveis selecionadas na etapa de Classificação. Tendo em vista a combinação de métodos, o *Framework* foi aplicado a uma nova base de dados, a PINTEC 2014, que também apresenta indicadores de inovação de setores econômicos. O desempenho do *Framework* se apresenta satisfatório para subsidiar a análise dos problemas de pesquisa deste estudo, sendo eles: a classificação de setores econômicos em níveis de inovação, a identificação de quais os indicadores influenciam na classificação e como estimar os níveis de inovação das observações. No entanto, cabe ressaltar que o número de observações substancialmente menor na base de 2014, bem como a natureza diferente das observações, causa um desempenho inferior dos modelos do *Framework* nas etapas de Classificação e Estimação. Essa constatação é entendida como limitação do estudo, ainda que o objetivo seja propor uma metodologia não fixada à amostra utilizada.

Como forma de extensão deste trabalho, e considerando as características do problema estudado, podem ainda ser empregados métodos de geração de amostras sintéticas a fim de

aumentar o número de observações da base, além de métodos para redução de dimensionalidade, com intuito de diminuir o número de variáveis da base antes da classificação dos dados.

REFERÊNCIAS

ANDRONIKIDIS A.; KAROLIDIS D. & ZAFEIRIOU G. Reflections on grounding firm innovation and viability. **European Management Journal**, v. 39, n.1, p. 2-8, feb. 2020.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5-32, oct. 2021.

CARMONA, P.; CLIMENT, F.; MOMPARLER, A. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. **International Review of Economics & Finance**, v. 61, p. 304-323, may 2019.

CRUZ-CÁZARES, C.; BAYONA-SÁEZ, C.; GARCÍA-MARCO, T. You can't manage right what you can't measure well: Technological innovation efficiency. **Research Policy**, 42, n. 6-7, p. 1239-1250, jul./aug. 2013.

CURADO, C.; MUÑOZ-PASCUAL, L. & GALENDE, J. Antecedents to innovation performance in SMEs: A mixed methods approach. **Journal of Business Research**, 89, p. 206-215, aug. 2018.

CUTLER A.; CUTLER D.R.; STEVENS J. R. Random Forests. In: Zhang C., Ma Y. (eds) **Ensemble Machine Learning**. Springer, Boston, MA, 2012, 332 p.

DEBAERE, S; COUSSEMENT, K.; RUYCK, T. D. Multi-label classification of member participation in online innovation communities. **European Journal of Operational Research**, v. 270, n 2, p. 761-774, oct. 2018.

DEMIDOVA, L.A.; KLYUEVA, I.A.; PYLKIN, A.N. Hybrid Approach to Improving the Results of the SVM Classification Using the Random Forest Algorithm. **Procedia Computer Science**, v. 150, p. 455-461, 2019.

DIAZ-ROZO, J.; BIELZA, C.; LARRAÑAG, P. Machine-tool condition monitoring with Gaussian mixture models-based dynamic probabilistic clustering. **Engineering Applications of Artificial Intelligence**, v. 89, 103434, mar. 2020.

FORNER, D.; OZCAN, S.; BACON, D. Machine Learning Approach for National Innovation Performance Data Analysis. **Proceedings of the 8th International Conference on Data Science, Technology and Applications** – DATA, Prague, Czech Republic, p. 325-331, 2019.

FOUEDJIO, F. Exact Conditioning of Regression Random Forest for Spatial Prediction. **Artificial Intelligence in Geosciences**, v. 1, p. 11-23, dec. 2020.

GOMES, G.; MACHADO, D. D. P. N; ALEGRE, J. Indústria têxtil de Santa Catarina e sua capacidade inovadora: estudo sob a perspectiva da eficiência, eficácia, custos e melhoria de processos. **RAI Revista de Administração e Inovação**, São Paulo, v. 11, n. 2, p. 273-294, apr./jun. 2014.

GONÇALVES FILHO, C; VEIT, M. R; MONTEIRO, P. R. R. Inovação, estratégia, orientação para o mercado e empreendedorismo: identificação de clusters de empresas e teste de modelo de predição do desempenho nos negócios. **RAI Revista de Administração e Inovação**, v. 10, n. 2, p. 81-101, jul. 2013.

HAJEK, P.; HENRIQUES, R. Modelling innovation performance of European regions using multi-output neural networks. **PLoS ONE**, v. 12, n. 10, p. 1-21, oct. 2017.

HAJEK, P.; HENRIQUES, R.; CASTELLI, M.; VANNESCHI, L. Forecasting performance of regional innovation systems using semantic-based genetic programming with local search optimizer. **Computers & Operations Research**, v. 106, p. 179-190, jun. 2019.

HASTIE, T.; TIBSHIRANI R.; FRIEDMAN J. **Random Forests**. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY, 2009, 745 p.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa de inovação: 2014** / IBGE, Coordenação de Indústria. Rio de Janeiro, RJ, 2016, 105 p. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=299007. Acesso em 30 abr. 2022.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Pesquisa de inovação 2017**: notas técnicas. Rio de Janeiro, IBGE, 2020, 55 p. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101706_notas_tecnicas.pdf. Acesso em 14 fev. 2021.

KANNEBLEY JÚNIOR, S.; PORTO, G. & PAZELLO, E. Characteristics of Brazilian innovative firms: An empirical analysis based on PINTEC - Industrial research on technological innovation. **Research Policy**, 34, n.6, p. 872-893, aug. 2005.

LAURETTO, M. S. **Árvores de decisão.** EACH-USP, 2010. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4469825/mod_resource/content/1/ArvoresDecisao_normalsize.pdf. Acesso em 12 jun. 2021.

LONGHINI, T. M.; CAVALCANTI, J. M. M.; BORGES, S. L.; FERREIRA, B. P. Investimentos em Inovação e sua Influência na Receita Líquida de Vendas: Uma Análise com Base nos Dados do PINTEC. **BBR Brazilian Business Review** [online], v.15, n.1, p. 1-16, jan./fev. 2018.

LONGJUN, D.; XIBING, L., MING, X.; QIYUE, L. Comparisons of Random Forest and Support Vector Machine for Predicting Blasting Vibration Characteristic Parameters. **Procedia Engineering**, v. 26, p. 1772-1781, 2011.

MATPLOTLIB – Matplotlib [Online]. Disponível em: https://matplotlib.org/stable/index.html. Acesso em 05 jun. 2021.

NUMPY — Numpy. [Online]. Disponível em: https://www.numpy.org/. Acesso em 05 jun. 2021.

OECD - ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT **Oslo Manual**: Guidelines for Collecting and Interpreting Innovation Data. 3 ed. Paris, France, OECD Publications, 2005, 163 p.

PANDAS - Pandas. [Online]. Disponível em: https://pandas.pydata.org/. Acesso em 05 jun. 2021.

PARK, H; ANDERSON, T. R; SEO, W. Regional innovation capability from a technology-oriented perspective: An analysis at industry level. **Computers in Industry**, v. 129, n. 103441, aug. 2021.

PATEL, E.; KUSHWAHA, S. D. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. **Procedia Computer Science**, v. 171, p. 158-167, 2020.

RUSU, M. Smart Specialization a Possible Solution to the New Global Challenges. International Economic Conference of Sibiu 2013 Post Crisis Economy: Challenges and Opportunities, IECS 2013. **Procedia Economics and Finance**, v. 6, p. 128–136, 2013.

SAGI, O.; ROKACH, L. Approximating XGBoost with an interpretable decision tree. **Information Sciences**, v. 572, p. 522-542, sep. 2021.

SANTOS, D. F. L.; BASSO, L.F. C.; KIMURA, H.; KAYO, E. K. Innovation efforts and performances of Brazilian firms. **Journal of Business Research**, v. 67, n. 4, p. 527-535, jan. 2014.

SANTOS, D. F. L.; PESTILLO, L. Padrões setoriais de inovação e desempenho na indústria brasileira. **Revista Brasileira de Economia de Empresas**, v. 19, n.1, p. 79-110, oct. 2019.

SAWANGARREERAK, S.; THANATHAMATHEE, P. **Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression. Information**, v. 11, n. 11, p. 519, nov. 2020.

SCIKIT-LEARN - Scikit-learn roadmap. [Online]. Disponível em: https://scikit-learn.org/stable/roadmap.html. Acesso em 05 jun. 2021.

SEABORN – An introduction to seaborn. [Online]. Disponível em: https://seaborn.pydata.org/introduction.html. Acesso em 05 jun. 2021.

SENTHILNATHAN, K.; SHANMUGAM, B.; GOYAL, D.; ANNAPOORANI, I.; SAMIKANNU, R. **Deep Learning Applications and Intelligent Decision Making in Engineering**. Estados Unidos: IGI Global, p. 151-152, oct. 2020.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. **Applied Soft Computing**, v. 97, Part B, 105524, dec. 2020.

SOBRAL, A. P. B. **Hourly load forecasting a new approach through decision tree**. 2003. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2003.

SULISTIANI, H.; TJAHYANTO, A. Comparative Analysis of Feature Selection Method to Predict Customer Loyalty. **IPTEK Journal of Engineering**, v. 3, n. 1, may 2017.

TANHA, J.; ABDI, Y.; SAMADI, N.; RAZZAGHI, N.; ASADPOUR, M. Boosting methods for multi-class imbalanced data classification: an experimental review. **Journal of Big Data**, v. 7, n. 70, sep. 2020.

TUŠAR, T.; GANTAR, K.; KOBLAR, V.; ŽENKO, B.; FILIPIČ, B. A study of overfitting in optimization of a manufacturing quality control procedure. **Applied Soft Computing**, v. 59, p. 77-87, oct. 2017.

VEGA-JURADO, J.; GUTIÉRREZ-GRACIA, A.; FERNÁNDEZ-DE-LUCIO, I.; MANJARRÉS-HENRÍQUEZ, L. The effect of external and internal factors on firms' product innovation. **Research Policy**, v. 37, n. 4, p. 616–632, may 2008.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error over the root mean square error in assessing average model performance. **Climate Research**, v. 30, n.1, p. 79–82, dec. 2005.

XGBOOST – XGBoost Documentation. [Online]. Disponível em: https://xgboost.readthedocs.io/en/latest/. Acesso em 05 jun. 2021.

ZHU, L.; QIU, D.; ERGU, D.; YING, C.; LIU, K. A study on predicting loan default based on the random forest algorithm. 7th International Conference on Information Technology and Quantitative Management, **Procedia Computer Science**, v. 162, p. 503-513, 2019.

ANEXO I

Estado	**Setor**
Amazonas	Fabricação de bebidas
Amazonas	Refino de petróleo
Amazonas	Fabricação de equipamentos de comunicação
Amazonas	Outras atividades da indústria
Bahia	Fabricação de produtos alimentícios
Bahia	Fabricação de celulose e outras pastas
Bahia	Refino de petróleo
Bahia	Fabricação de produtos químicos orgânicos
Bahia	Outras atividades da indústria
Ceará	Fabricação de produtos alimentícios
Ceará	Confecção de artigos do vestuário e acessórios
Ceará	Preparação de couros e fabricação de artefatos de couro, artigos de viagem e calçados
Ceará	Outras atividades da indústria
Distrito Federal	Atividades dos serviços de tecnologia da informação
Distrito Federal	Tratamento de dados, hospedagem internet e outras atividades relacionadas
Distrito Federal	Outras atividades dos Serviços Selecionados
Espírito Santo	Indústrias extrativas
Espírito Santo	Outras atividades da indústria
Goiás	Fabricação de produtos alimentícios
Goiás	Outras atividades da indústria
Mato Grosso	Fabricação de produtos alimentícios
Mato Grosso	Outras atividades da indústria
Mato Grosso do Sul	Fabricação de produtos alimentícios
Mato Grosso do Sul	Outras atividades da indústria
Mis Gerais	Fabricação de produtos alimentícios
Mis Gerais	Indústrias extrativas
Mis Gerais	Refino de petróleo
Mis Gerais	Produtos siderúrgicos
Mis Gerais	Outras atividades da indústria
Mis Gerais	Atividades dos serviços de tecnologia da informação
Mis Gerais	Serviços de arquitetura e engenharia, testes e análises técnicas
Mis Gerais	Outras atividades dos Serviços Selecionados
Pará	Indústrias extrativas
Pará	Outras atividades da indústria
Paraná	Fabricação de produtos alimentícios
Paraná	Refino de petróleo
Paraná	Fabricação de automóveis, caminhonetas e utilitários, caminhões e ônibus

Paraná	Outras atividades da indústria
Paraná	Edição e gravação e edição de música
Paraná	Telecomunicações
Paraná	Serviços de arquitetura e engenharia, testes e análises técnicas
Paraná	Outras atividades dos Serviços Selecionados
Pernambuco	Fabricação de produtos alimentícios
Pernambuco	Fabricação de bebidas
Pernambuco	Fabricação de sabões, detergentes, produtos de limpeza, cosméticos, produtos de perfumaria e de higiene pessoal
Pernambuco	Fabricação de produtos de minerais não-metálicos
Pernambuco	Fabricação de produtos de metal
Pernambuco	Outras atividades da indústria
Rio de Janeiro	Indústrias extrativas
Rio de Janeiro	Refino de petróleo
Rio de Janeiro	Outras atividades da indústria
Rio de Janeiro	Edição e gravação e edição de música
Rio de Janeiro	Telecomunicações
Rio de Janeiro	Atividades dos serviços de tecnologia da informação
Rio de Janeiro	Serviços de arquitetura e engenharia, testes e análises técnicas
Rio de Janeiro	Outras atividades dos Serviços Selecionados
Rio Grande do Sul	Fabricação de produtos do fumo
Rio Grande do Sul	Preparação de couros e fabricação de artefatos de couro, artigos de viagem e calçados
Rio Grande do Sul	Refino de petróleo
Rio Grande do Sul	Fabricação de resinas e elastômeros, fibras artificiais e sintéticas, defensivos agrícolas e desinfetantes domissanitários
Rio Grande do Sul	Fabricação de artigos de borracha e plástico
Rio Grande do Sul	Fabricação de produtos de metal
Rio Grande do Sul	Outras atividades da indústria
Rio Grande do Sul	Edição e gravação e edição de música
Rio Grande do Sul	Atividades dos serviços de tecnologia da informação
Rio Grande do Sul	Tratamento de dados, hospedagem internet e outras atividades relacionadas
Rio Grande do Sul	Outras atividades dos Serviços Selecionados
Rio Grande do Sul	Fabricação de produtos alimentícios
Santa Catarina	Fabricação de produtos alimentícios
Santa Catarina	Fabricação de produtos têxteis
Santa Catarina	Confecção de artigos do vestuário e acessórios
Santa Catarina	Fabricação de papel, embalagens e artefatos de papel

Santa Catarina	Fabricação de artigos de borracha e plástico
Santa Catarina	Fabricação de geradores, transformadores e equipamentos para distribuição de energia elétrica
Santa Catarina	Outras atividades da indústria
Santa Catarina	Atividades dos serviços de tecnologia da informação
Santa Catarina	Outras atividades dos Serviços Selecionados
São Paulo	Indústrias extrativas
São Paulo	Fabricação de produtos alimentícios
São Paulo	Fabricação de bebidas
São Paulo	Fabricação de papel, embalagens e artefatos de papel
São Paulo	Fabricação de coque e biocombustíveis (álcool e outros)
São Paulo	Refino de petróleo
São Paulo	Fabricação de resinas e elastômeros, fibras artificiais e sintéticas, defensivos agrícolas e desinfetantes domissanitários
São Paulo	Fabricação de sabões, detergentes, produtos de limpeza, cosméticos, produtos de perfumaria e de higiene pessoal
São Paulo	Fabricação de tintas, vernizes, esmaltes, lacas e produtos afins e de produtos diversos
São Paulo	Fabricação de produtos farmacêuticos
São Paulo	Fabricação de artigos de borracha e plástico
São Paulo	Fabricação de produtos de minerais não-metálicos
São Paulo	Metalurgia de metais não-ferrosos e fundição
São Paulo	Fabricação de produtos de metal
São Paulo	Fabricação de geradores, transformadores e equipamentos para distribuição de energia elétrica
São Paulo	Motores, bombas, compressores e equipamentos de transmissão
São Paulo	Outras maquinas e equipamentos
São Paulo	Fabricação de automóveis, caminhonetas e utilitários, caminhões e ônibus
São Paulo	Fabricação de peças e acessórios para veículos
São Paulo	Fabricação de outros equipamentos de transporte
São Paulo	Outras atividades da indústria
São Paulo	Edição e gravação e edição de música
São Paulo	Telecomunicações
São Paulo	Atividades dos serviços de tecnologia da informação
São Paulo	Tratamento de dados, hospedagem internet e outras atividades relacionadas
São Paulo	Serviços de arquitetura e engenharia, testes e análises técnicas
São Paulo	Outras atividades dos Serviços Selecionados

ANEXO II

Variável selecionada	Dimensão	Indicador
X423	Sustentabilidade e inovação ambiental	% empresas que publicaram relatórios de sustentabilidade
X402	Sustentabilidade e inovação ambiental	% empresas que implementaram inovações e que reduziram o impacto ambiental, com redução da contaminação do solo, da água, de ruído ou do ar, com média importância
X63	Atividades inovativas	% empresas que implementaram inovações de produto e/ou processo que adquiriram treinamento e indicam alta importância para isso
X92	Atividades inovativas	% de pesquisadores pós-graduados ocupados em atividades internas de P&D nas empresas que implementaram inovações
X118	Impactos das inovações	% empresas que implementaram inovações com alto impacto na ampliação da participação da empresa no mercado
X377	Inovações organizacionais e de marketing	% empresas que implementaram inovações de produto e processo e inovações organizacionais e/ou de marketing em técnicas de gestão
X304	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos, por alta importância atribuída à escassez de fontes apropriadas de financiamento
X218	Fontes de informação	% empresas que implementaram inovações e empregaram fontes de informação - Instituições de testes, ensaios e certificações - do Brasil
X293	Projetos incompletos e abandonados	% de empresas que não implementaram inovações e sem projetos
X306	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos, por baixa importância atribuída à escassez de fontes apropriadas de financiamento
X229	Relações de cooperação para inovação	% empresas que implementaram inovações com relações de cooperação com clientes ou consumidores, mas atribuem baixa importância da parceria
X60	Atividades inovativas	% empresas que implementaram inovações de produto e/ou processo que adquiriram máquinas e equipamentos e indicam alta importância para isso
X418	Sustentabilidade e inovação ambiental	% empresas que implementaram inovações, reduziram o impacto ambiental, com contribuição, para introdução de inovações ambientais, de ações voluntárias
X17	Projetos incompletos e abandonados	% empresas que não implementaram inovações, com projetos incompletos e abandonados
X178	Fontes de informação	% empresas que implementaram inovações com emprego de informação das fontes externas - consultores – e atribuem a isso alta importância
X334	Problemas e obstáculos à inovação	% empresas que implementaram inovações e apontam impactos de problemas e obstáculos na velocidade ou inviabilidade de projetos
X235	Relações de cooperação para inovação	% empresas que implementaram inovações com relações de cooperação com concorrentes, mas atribuíram baixa importância da parceria
X317	Problemas e obstáculos à inovação	% de empresas que não implementaram inovações e sem projetos por falta de informação sobre mercados de média importância
X4	Inovação de produto	% empresas que implementaram inovações de produto
X316	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por falta de informação sobre mercados de alta importância

X288	Apoio do governo	% empresas que implementaram inovações que receberam apoio do governo - financiamento a projetos de P&D, sem parceria com universidades
X371	Inovações organizacionais e de marketing	% empresas que não implementaram produto ou processo e sem projetos, mas inovações organizacionais e/ou de marketing em técnicas de gestão
X180	Fontes de informação	% empresas que implementaram inovações com emprego de informação das fontes externas - consultores – mas atribuíram baixa importância a isso
X372	Inovações organizacionais e de marketing	% empresas que não implementaram produto ou processo e sem projetos, mas inovações organizacionais e/ou de marketing em técnicas de gestão ambiental
X312	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por falta de pessoal qualificado de baixa importância
X420	Sustentabilidade e inovação ambiental	% empresas que implementaram inovações, reduziram o impacto ambiental, por fator de contribuição para introdução de inovações ambientais - elevados custos
X336	Problemas e obstáculos à inovação	% empresas que implementaram inovações por médio grau de importância de riscos econômicos excessivos
X311	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por médio grau de importância de falta de pessoal qualificado
X124	Impactos das inovações	% empresas que implementaram inovações com impacto alto causado em aumento da capacidade produtiva
X358	Relações de cooperação para inovação	% empresas que implementaram inovações por baixo grau de importância atribuído a escassas possibilidades de cooperação com outras empresas/instituições
X39	Relações de cooperação para inovação	% empresas que implementaram inovações em que a principal responsável pelo desenvolvimento de produto são outras empresas ou institutos
X333	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por baixo grau de importância de centralização da atividade inovativa em outra empresa do grupo
X315	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por baixo grau de importância de falta de informação sobre tecnologia
X53	Atividades inovativas	% empresas que implementaram inovações de produto e/ou processo que desenvolveram aquisição externa de P&D e indicam baixa importância para isso ou não desenvolveram aquisição externa de P&D
X397	Sustentabilidade e inovação ambiental	% empresas que implementaram inovações e que reduziram o impacto ambiental com a substituição de energia de combustíveis fósseis por energia renovável com alto grau de importância
X401	Sustentabilidade e inovação ambiental	% empresas que implementaram inovações e que reduziram o impacto ambiental com redução da contaminação do solo, da água, de ruído ou do ar, com alto grau de importância
X119	Impactos das inovações	% empresas que implementaram inovações com médio impacto causado em ampliação da participação da empresa no mercado
X425	Sustentabilidade e inovação ambiental	% empresas que produziram energia renovável
X344	Problemas e obstáculos à inovação	% de empresas que implementaram inovações, com problemas e obstáculos – rigidez organizacional – atribuída à alta importância
X77	Atividades inovativas	% empresas com dispêndios realizados em atividades inovativas - aquisição de máquinas e equipamentos

X19	Grau de inovação	% empresas que implementaram inovações de produto novo para a empresa, mas já existente no mercado nacional, com aprimoramento de um já existente
X117	Impactos das inovações	% empresas que implementaram inovações com impacto baixo causado em manutenção da participação da empresa no mercado
X72	Atividades inovativas	% empresas com dispêndios realizados em atividades inovativas
X163	Fontes de informação	% empresas que implementaram inovações com emprego de informação das fontes internas - outras áreas - de importância alta
X287	Apoio do governo	% empresas que implementaram inovações que receberam apoio do governo - subvenção econômica
X305	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por médio grau de importância atribuído à escassez de fontes apropriadas de financiamento
X351	Fontes de informação	% empresas que implementaram inovações por médio grau de importância de falta de informação sobre tecnologia
X185	Fontes de informação	% empresas que implementaram inovações com emprego de informação das fontes externas - institutos de pesquisa ou centros tecnológicos - de importância média
X206	Fontes de informação	% empresas que implementaram inovações e empregaram fontes de informação - clientes ou consumidores - do Brasil
X364	Problemas e obstáculos à inovação	% empresas que implementaram inovações por baixo grau de importância atribuído à fraca resposta dos consumidores quanto a novos produtos
X342	Problemas e obstáculos à inovação	% empresas que implementaram inovações por médio grau de importância atribuído à escassez de fontes apropriadas de financiamento
X303	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por baixo grau de importância atribuído a elevados custos da inovação
X139	Impactos das inovações	% empresas que implementaram inovações com impacto alto causado em redução do consumo de energia
X374	Inovações organizacionais e de marketing	% empresas que não implementaram produto ou processo e sem projetos, mas inovações organizacionais e/ou de marketing em relações externas
X199	Fontes de informação	% empresas que implementaram inovações com emprego de informação das fontes externas - redes de informação informatizadas - de importância alta
X261	Relações de cooperação para inovação	% empresas que implementaram inovações com relações de cooperação com empresas de consultoria no exterior
X131	Impactos das inovações	% empresas que implementaram inovações com impacto médio causado na redução dos custos de produção
X307	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por alto grau de importância atribuído à rigidez organizacional
X89	Atividades inovativas	% empresas que implementaram inovações com dispêndios em atividades internas de P&D de caráter ocasional
X177	Fontes de informação	% empresas que implementaram inovações com emprego de informação das fontes externas - concorrentes - de importância baixa
X66	Atividades inovativas	% empresas que implementaram inovações de produto e/ou processo que desenvolveram introdução das inovações

		tecnológicas no mercado e indicam alta importância para esse fator
X122	Impactos das inovações	% empresas que implementaram inovações com impacto médio causado na abertura de novos mercados
X253	Relações de cooperação para inovação	% empresas que implementaram inovações com relações de cooperação com clientes ou consumidores no exterior
X309	Problemas e obstáculos à inovação	% empresas que não implementaram inovações e sem projetos por baixo grau de importância atribuído à rigidez organizacional
X25	Grau de inovação	% empresas que implementaram inovações de produto novo para o mercado mundial, com aprimoramento de um já existente

ANEXO III

Posição	Y pred.	Classe	Setor-estado
1º	0,79	1	São Paulo Outras atividades dos Serviços Selecionados
2º	0,79	1	Paraná Fabricação de automóveis, caminhonetas e utilitários, caminhões e ônibus
3º	0,73	1	São Paulo Fabricação de automóveis, caminhonetas e utilitários, caminhões e ônibus
4º	0,72	1	Pernambuco Fabricação de sabões, detergentes, produtos de limpeza, cosméticos, produtos de perfumaria e de higiene pessoal
5º	0,71	1	Mis Gerais Refino de petróleo
6º	0,70	1	Bahia Fabricação de celulose e outras pastas
7º	0,67	1	Santa Catarina Outras atividades dos Serviços Selecionados
8º	0,66	1	Rio Grande do Sul Fabricação de resinas e elastômeros, fibras artificiais e sintéticas, defensivos agrícolas e desinfetantes domissanitários
9º	0,66	2	Rio de Janeiro Refino de petróleo
10º	0,66	2	Paraná Fabricação de produtos alimentícios
11º	0,61	2	Amazonas Fabricação de equipamentos de comunicação
12º	0,58	2	Amazonas Fabricação de bebidas
13º	0,56	2	São Paulo Fabricação de produtos farmacêuticos
14º	0,51	2	Pernambuco Outras atividades da indústria
15º	0,51	2	Rio de Janeiro Telecomunicações
16º	0,49	2	Santa Catarina Atividades dos serviços de tecnologia da informação
17º	0,49	2	Bahia Refino de petróleo
18º	0,49	2	São Paulo Tratamento de dados, hospedagem internet e outras atividades relacionadas
19º	0,46	2	Rio Grande do Sul Atividades dos serviços de tecnologia da informação
20º	0,46	2	Mis Gerais Atividades dos serviços de tecnologia da informação
21º	0,45	2	Rio Grande do Sul Fabricação de artigos de borracha e plástico
22º	0,45	2	São Paulo Fabricação de peças e acessórios para veículos
23º	0,45	2	Paraná Telecomunicações
24º	0,45	2	São Paulo Fabricação de bebidas
25º	0,44	2	Distrito Federal Tratamento de dados, hospedagem internet e outras atividades relacionadas
26º	0,43	2	Amazonas Outras atividades da indústria
27º	0,43	2	São Paulo Fabricação de resinas e elastômeros, fibras artificiais e sintéticas, defensivos agrícolas e desinfetantes domissanitários

28º	0,43	2	Santa Catarina Fabricação de geradores, transformadores e equipamentos para distribuição de energia elétrica
29º	0,42	2	São Paulo Refino de petróleo
30º	0,42	2	Goiás Outras atividades da indústria
31º	0,42	2	Mis Gerais Produtos siderúrgicos
32º	0,42	2	São Paulo Fabricação de tintas, vernizes, esmaltes, lacas e produtos afins e de produtos diversos
33º	0,41	2	Santa Catarina Fabricação de papel, embalagens e artefatos de papel
34º	0,41	2	Paraná Serviços de arquitetura e engenharia, testes e análises técnicas
35º	0,40	2	Mis Gerais Outras atividades dos Serviços Selecionados
36º	0,40	2	São Paulo Fabricação de produtos alimentícios
37º	0,40	2	Santa Catarina Fabricação de produtos alimentícios
38º	0,39	2	Mato Grosso do Sul Outras atividades da indústria
39º	0,39	2	Rio Grande do Sul Outras atividades da indústria
40º	0,39	2	Paraná Outras atividades dos Serviços Selecionados
41º	0,39	2	São Paulo Atividades dos serviços de tecnologia da informação
42º	0,39	2	São Paulo Fabricação de sabões, detergentes, produtos de limpeza, cosméticos, produtos de perfumaria e de higiene pessoal
43º	0,38	2	Paraná Refino de petróleo
44º	0,38	2	São Paulo Motores, bombas, compressores e equipamentos de transmissão
45º	0,38	2	Pará Outras atividades da indústria
46º	0,38	2	São Paulo Fabricação de coque e biocombustíveis (álcool e outros)
47º	0,37	2	Mato Grosso Fabricação de produtos alimentícios
48º	0,36	2	Mato Grosso do Sul Fabricação de produtos alimentícios
49º	0,36	2	Goiás Fabricação de produtos alimentícios
50º	0,35	2	Rio de Janeiro Atividades dos serviços de tecnologia da informação
51º	0,35	2	Santa Catarina Outras atividades da indústria
52º	0,35	2	Bahia Fabricação de produtos químicos orgânicos
53º	0,35	2	Paraná Outras atividades da indústria
54º	0,34	2	Mis Gerais Fabricação de produtos alimentícios
55º	0,34	2	São Paulo Fabricação de geradores, transformadores e equipamentos para distribuição de energia elétrica
56º	0,34	2	Pernambuco Fabricação de produtos de metal
57º	0,34	2	Santa Catarina Fabricação de artigos de borracha e plástico
58º	0,34	2	Mato Grosso Outras atividades da indústria
59º	0,34	2	Paraná Edição e gravação e edição de música
60º	0,33	2	São Paulo Fabricação de papel, embalagens e artefatos de papel

61º	0,33	2	Bahia Outras atividades da indústria
62º	0,33	3	São Paulo Outras atividades da indústria
63º	0,32	3	Mis Gerais Outras atividades da indústria
64º	0,32	3	Rio Grande do Sul Fabricação de produtos de metal
65º	0,31	3	Bahia Fabricação de produtos alimentícios
66º	0,31	3	Rio Grande do Sul Outras atividades dos Serviços Selecionados
67º	0,31	3	Pernambuco Fabricação de produtos alimentícios
68º	0,31	3	Santa Catarina Confecção de artigos do vestuário e acessórios
69º	0,31	3	Rio Grande do Sul Fabricação de produtos alimentícios
70º	0,30	3	Espírito Santo Outras atividades da indústria
71º	0,30	3	Rio de Janeiro Edição e gravação e edição de música
72º	0,29	3	Ceará Confecção de artigos do vestuário e acessórios
73º	0,28	3	Santa Catarina Fabricação de produtos têxteis
74º	0,27	3	São Paulo Fabricação de artigos de borracha e plástico
75º	0,26	3	Rio Grande do Sul Fabricação de produtos do fumo
76º	0,26	3	São Paulo Fabricação de outros equipamentos de transporte
77º	0,25	3	Rio de Janeiro Outras atividades da indústria
78º	0,24	3	São Paulo Fabricação de produtos de minerais não-metálicos
79º	0,24	3	São Paulo Outras máquinas e equipamentos
80º	0,23	3	Rio de Janeiro Serviços de arquitetura e engenharia, testes e análises técnicas
81º	0,23	3	São Paulo Fabricação de produtos de metal
82º	0,22	3	São Paulo Metalurgia de metais não-ferrosos e fundição
83º	0,22	3	Rio Grande do Sul Preparação de couros e fabricação de artefatos de couro, artigos de viagem e calçados
84º	0,22	3	Ceará Fabricação de produtos alimentícios
85º	0,21	3	Mis Gerais Serviços de arquitetura e engenharia, testes e análises técnicas
86º	0,21	3	São Paulo Serviços de arquitetura e engenharia, testes e análises técnicas
87º	0,19	3	Mis Gerais Indústrias extrativas
88º	0,18	3	Distrito Federal Atividades dos serviços de tecnologia da informação
89º	0,18	3	Espírito Santo Indústrias extrativas
90º	0,17	3	Rio Grande do Sul Tratamento de dados, hospedagem internet e outras atividades relacionadas
91º	0,17	3	Rio Grande do Sul Edição e gravação e edição de música
92º	0,16	3	Ceará Outras atividades da indústria
93º	0,16	3	Pernambuco Fabricação de bebidas
94º	0,15	3	São Paulo Indústrias extrativas
95º	0,15	3	Rio de Janeiro Indústrias extrativas

96º	0,13	3	Rio de Janeiro Outras atividades dos Serviços Selecionados
97º	0,13	3	São Paulo Edição e gravação e edição de música
98º	0,13	3	Pernambuco Fabricação de produtos de minerais não-metálicos
99º	0,12	3	Distrito Federal Outras atividades dos Serviços Selecionados
100º	0,11	3	Pará Indústrias extrativas
101º	0,11	3	São Paulo Telecomunicações
102º	0,09	3	Ceará Preparação de couros e fabricação de artefatos de couro, artigos de viagem e calçados
103º	0,08	3	Rio Grande do Sul Refino de petróleo
104º	0,08	3	Amazonas Refino de petróleo