ipea Instituto de Pesquisa Econômica Aplicada

Título do capítulo	CAPÍTULO 5 **EMPRESAS CANDIDATAS A FINANCIAMENTO DO BNDES**
Autor(es)	Bernardo Alves Furtado Ludmilla Mattos Rafael Morais
DOI	DOI: http://dx.doi.org/10.38116/9786556350370cap5

Título do livro	**Financiar o Futuro: o papel do BNDES**
Organizadores(as)	João Alberto De Negri Bruno César Araújo Ricardo Bacelette
Volume	1
Série	**Financiar o Futuro: o papel do BNDES**
Cidade	Rio de Janeiro
Editora	Instituto de Pesquisa Econômica Aplicada (Ipea)
Ano	2022
Edição	1a
ISBN	9786556350370
DOI	DOI: http://dx.doi.org/10.38116/9786556350370

CAPÍTULO 5

EMPRESAS CANDIDATAS A FINANCIAMENTO DO BNDES

Bernardo Alves Furtado[1]
Ludmilla Mattos[2]
Rafael Morais[3]

1 INTRODUÇÃO

O objetivo deste capítulo é apresentar uma estimativa de classificação de empresas brasileiras que não receberam financiamentos do Banco Nacional de Desenvolvimento Econômico e Social (BNDES), mas que seriam candidatas em potencial. A classificação é feita por meio de um exercício de aprendizado de máquina, a partir das bases de dados das empresas brasileiras – disponibilizadas na Relação Anual de Informações Sociais (Rais) – e das informações de financiamentos concedidos pelo BNDES. Os resultados incluem as características das empresas mais bem classificadas em relação ao conjunto total de empresas e àquelas não classificadas.

2 PASSOS METODOLÓGICOS

2.1 Montagem e união de bases de dados

O primeiro passo realizado foi a montagem da base de dados do BNDES, a partir do banco de dados BNDES Transparência. Foram incluídas informações referentes a financiamentos para a administração pública (30 de junho de 1994 a 31 de março de 2021), por meio de: i) operações indiretas automáticas (1º de janeiro de 2002 a 31 de março de 2021); ii) operações de pré-embarque (1º de janeiro de 2002 a 31 de março de 2021); e iii) operações de pós-embarque de bens e serviços (1º de janeiro de 2002 a 31 de março de 2021).

As informações incluem detalhes públicos de:

- descrição do projeto;
- Cadastro Nacional da Pessoa Jurídica (CNPJ) do cliente;

1. Técnico de planejamento e pesquisa na Diretoria de Estudos e Políticas Setoriais de Inovação e Infraestrutura (Diset) do Ipea. *E-mail*: <bernardo.furtado@ipea.gov.br>.
2. Pesquisadora do Programa de Pesquisa para o Desenvolvimento Nacional (PNPD) na Diset/Ipea. *E-mail*: <ludmilla.silva@ipea.gov.br>.
3. Pesquisador do PNPD na Diset/Ipea. *E-mail*: <rafael.morais@ipea.gov.br>.

- data;
- município;
- natureza do cliente;
- valor contratado em reais;
- juros;
- prazo de carência e amortização;
- modalidade;
- apoio;
- produto;
- instrumento financeiro;
- se objeto de inovação;
- área operacional;
- setor Classificação Nacional de Atividades Econômicas (CNAE), setor e subsetor BNDES;
- fonte dos recursos; e
- nome do agente financeiro e seu CNPJ.

A partir da montagem de base do BNDES, realizou-se a compatibilização com dados gerais do CNJP das empresas constantes da Rais. Em seguida, foi montado o painel segundo os dados anuais coletados.

A base da Rais é ajustada, mantida e atualizada no Ipea, incluindo variáveis calculadas *in-house*, como a *PO_TEC* – que inclui empregados cuja Classificação Brasileira de Ocupações (CBO) esteja enquadrada no grupo de pesquisadores, ou engenheiros, ou cientistas. A base da Rais contém dados sobre os empregados (escolaridade, idade e salários), bem como sobre a própria empresa, como porte, setor, massa salarial, entre outros.

2.2 Seleção de variáveis e amostra

A base de dados em painel com informações de todas as empresas – com e sem financiamento do BNDES – fundamentou a seleção de variáveis que *caracterizassem ao mesmo tempo a empresa e seus empregados*. Sobre empregados, selecionamos:

- número de empregados;
- tempo de emprego médio;

- salários médios;
- rotatividade média; e
- tempo de estudo médio.

Sobre as empresas, selecionamos:

- natureza jurídica;
- município de localização principal; e
- número de filiais.

A partir da base completa, restrita à década de 2010, foram selecionadas 100 mil observações, com características das empresas e possíveis empréstimos para todos os anos do período (2010-2019). Essa amostra contendo empresas com e sem financiamento foi utilizada para o aprendizado de máquina.

2.3 Aprendizado de máquina – florestas aleatórias

Florestas aleatórias são um tipo de aprendizado de máquina supervisionado. Neste caso, os processos algorítmicos são utilizados para aprender regras subjacentes que ponderam os dados de entrada em relação ao objetivo escolhido e classificam de forma hierárquica o resultado (Ren, Cheng e Han, 2017; Mishina *et al*. 2015). No caso em tela, o algoritmo buscou compreender, baseado nos dados das empresas e de seus empregados, quais seriam as empresas mais próximas das que receberam financiamentos anteriores do BNDES. Em outras palavras, quais características das empresas e em que magnitude as tornam similares àquelas que efetivamente foram financiadas.

A definição dos desenvolvedores é a seguinte: "Uma floresta aleatória é um metaestimador que ajusta uma variedade de classificadores de árvores de decisão em várias subamostras da base de dados e usa médias para melhorar a acurácia da capacidade preditiva ao mesmo tempo que controla para sobreajuste" (tradução nossa).[4]

O algoritmo de florestas aleatórias é uma generalização de processos simples de árvores de decisão. Árvores de decisão implicam escolhas numéricas sucessivas das características das empresas de forma a particionar os dados originais e iterativamente gerar grupos cada vez mais homogêneos. Matematicamente, o cálculo de regiões homogêneas é um tanto simples e é feito por meio do critério de separação de cada observação entre os ramos de acordo com suas características.

4. "*A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting*". Disponível em: <https://bit.ly/37rHJsa>.

Para compreender o processo de tomada de decisão, imagine uma árvore simples. Considere a raiz como uma determinada característica da empresa – por exemplo, porte. Se o porte é "maior que quinhentos empregados", siga pelo ramo da direita. Caso contrário, siga pelo ramo da esquerda. Outras características são usadas de forma subsequente, até que o conjunto de empresas no último ramo (a folha da árvore, ou nó final) seja pequeno o suficiente. Note que ramos distintos da árvore podem ter profundidades e ramificações variadas.

Os algoritmos utilizados realizam o processo descrito no parágrafo anterior para todas as características de entrada dos dados do modelo e várias alternativas de cortes e profundidades. Dado que o processo é supervisionado, o algoritmo contém a informação de qual é o resultado esperado. Desse modo, é possível realizar previsões de classificação e verificar qual árvore gera resultados que mais se aproximam do observado.

Com isso, o conjunto de dados é subdividido em amostras, várias árvores diferentes são estimadas e apenas aquelas com maior número de acertos – dados os resultados conhecidos – são mantidas. Isso é feito repetidamente (por isso, são chamadas de florestas aleatórias, em vez de árvores de decisão), até que o algoritmo encontre o melhor resultado possível, condicionado à amostra inicial de dados fornecidos.

Um outro subconjunto de dados é reservado no início do processo, não entra ao longo do aprendizado, e é utilizado ao final para caracterizar a capacidade do modelo de realizar classificações ao observar dados desconhecidos durante o processo de aprendizagem.

2.4 Implementação

O algoritmo para este capítulo foi implementado utilizando o Python 3.7.10 e a biblioteca RandomForestClassifier disponível em *sklearn* 0.24.2.[5] O desenvolvedor caracteriza o algoritmo como do tipo perturbe-e-combine (Breiman, 1998; 2002; Lavin *et al.*, 2021). Com isso, conjuntos de árvores classificadoras aleatórias são construídos e a previsão reflete a média de classificadores individuais. A escolha de amostras no conjunto de treinamento é feita por meio de um subconjunto aleatório com reposição – o chamado *bootstrap*.

Alguns parâmetros são escolhidos para cada exercício. O número de árvores da floresta (*n_estimators*) foi escolhido como 10 mil e o tamanho dos subconjuntos aleatórios a considerar quando tomar a decisão de criar um novo nó separador de ramos foi de quinze (*max_features*). A função de separação para cada nó utilizada foi o critério de Gini.[6]

5. Disponível em: <https://bit.ly/37rHJsa>.
6. Detalhes para a chamada *impurity function* disponíveis em: <https://online.stat.psu.edu/stat508/lesson/11/11.2>.

Apresentamos apenas os resultados referentes às florestas aleatórias. Entretanto, no período de construção do exercício também foram realizados testes com outros processos de aprendizado de máquina, especificamente: regressão logística, redes neurais e classificação com suportes vetoriais, além de um processo que combina as várias alternativas por meio de votação.

O processo computacional em si contou com os procedimentos a seguir descritos.

1) Leitura e adequação das bases de dados:
   a) separação da variável de interesse (número de contratos);
   b) transformação em variáveis *dummies* para informações qualitativas, utilizando-se somente de códigos de Grandes Regiões do Instituto Brasileiro de Geografia e Estatística (IBGE) e natureza jurídica (código município e natureza jurídica da empresa); e
   c) separação em amostras de treinamento e reserva para testes (25%).
2) Ajuste do modelo.
3) Predição do resultado para a base completa.
4) Vinculação da probabilidade de receber financiamento para as empresas que não receberam financiamento ao CNPJ – oito dígitos da empresa.
5) Classificação descendente de acordo com a probabilidade.
6) Separação do grupo de empresas referente àquelas entre os 5% mais bem classificados e probabilidade positiva de financiamento.
7) Descrição dos grupos:
   a) a base completa com todas as firmas;
   b) a base com aquelas que se classificaram no *top* 5% – as selecionadas; e
   c) a base com aquelas não classificadas.

Finalmente, vale notar que todo o processo considerou dados anuais das firmas no período 2010-2019. Nesse sentido, dados da firma – de porte e salários, por exemplo – variaram e caracterizam observações distintas para a mesma firma em anos diferentes.

3 RESULTADOS

Os resultados sugerem que há um grupo de firmas que não receberam financiamento do BNDES, mas se configuram como distintas das demais. Partindo dos 5% do número de observações da amostra completa, chegamos a quase 500 mil firmas que seriam prioritárias (tabela 1).

Essas firmas têm o porte bem maior que as comuns da amostra, com mediana de quase cinquenta empregados, em comparação com as firmas típicas de dois ou três empregados. São ainda bem mais consolidadas que as firmas da amostra geral, com mais de quatorze anos de existência, sendo mais que o dobro do padrão observado em geral.

Os empregados das firmas selecionadas, por sua vez, permanecem mais tempo no emprego em média, aproximando-se de três anos, sendo o comum pouco mais de dois anos nas firmas em geral. A rotatividade desses empregados também é relativamente menor. Entretanto, possivelmente por serem empresas de grande porte, a mediana de tempo de estudo médio é um pouco menor (10,73), se comparada às outras firmas (11,00).

TABELA 1
Características principais das firmas da amostra completa de empresas na Rais que não receberam financiamento do BNDES, das selecionadas de acordo com similaridade a empresas com financiamento e das remanescentes que não foram classificadas como prioritárias

	Amostra completa	Firmas remanescentes	Firmas selecionadas
Número de observações	29.233.896	27.793.376	1.440.520
Número de firmas únicas	5.773.178	5.678.484	498.314
Mediana do número de empregados	2,50	2,25	47,66
Mediana do tempo de emprego médio	25,90	25,40	35,29
Mediana da rotatividade	0,29	0,29	0,27
Mediana da massa salarial (R$)	2.814,59	2.587,70	90.385,74
Mediana do tempo de estudo médio	11,00	11,03	10,73
Mediana da idade da empresa	6,92	6,75	14,33

Elaboração dos autores.

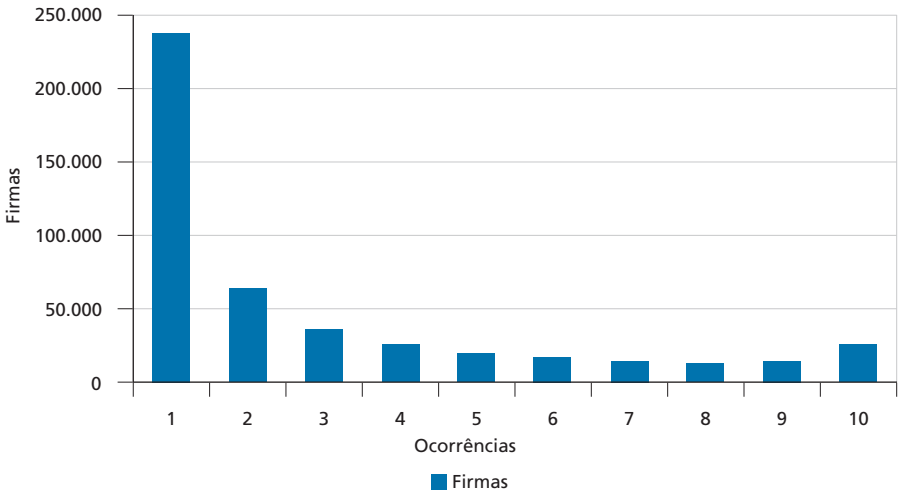
Algumas outras características adicionais das firmas selecionadas são bastante restritivas e não aparecem na análise da mediana da amostra e das firmas remanescentes. O número de filiais, por exemplo, é majoritariamente 1 para a grande maioria das empresas. Mesmo na amostra completa, a média do número de filiais ainda é de 1,1 por firma. Entre as selecionadas, embora mantenham a mediana em 1, o número de filiais médio sobe para 2,27.

Comportamento similar apresenta a variável calculada pelo Ipea *PO_TEC*, cuja mediana é 0 para todos os grupos, porém tem a média se elevando de 0,24 na amostra completa para 4,21 entre as firmas selecionadas.

Um segundo recorte possível é classificar apenas as empresas que foram selecionadas em todos os dez anos (gráfico 1). Nesse caso, observam-se, claramente, empresas de porte ainda maior, com cerca de 137 empregados, que trabalham na

empresa em média por período mais longo (há quase dez anos), e em empresas com aproximadamente três décadas de existência.

GRÁFICO 1
Número de firmas selecionadas e frequência com que são listadas



Elaboração dos autores.
Obs.: A maioria das firmas (255.458) é selecionada pelas suas características de apenas um ano, e 26.704 firmas aparecem em todos os dez anos da amostra.

A distribuição espacial é pequena, com grande concentração no município de São Paulo (tabela 2). Logo depois das capitais maiores, aparecem municípios da região Sul (Londrina, Maringá, Joinville e Caxias do Sul), seguidos de outros da região metropolitana de São Paulo.

TABELA 2
Lista de municípios e Unidades da Federação (UFs) com maior número de empresas selecionadas

UF	Município	Número de firmas selecionadas
São Paulo	São Paulo	60.860
Rio de Janeiro	Rio de Janeiro	21.721
Paraná	Curitiba	15.239
Rio Grande do Sul	Porto Alegre	11.691
Minas Gerais	Belo Horizonte	10.393
Goiás	Goiânia	6.593
Bahia	Salvador	6.160
Ceará	Fortaleza	5.888
Pernambuco	Recife	5.233

(Continua)

(Continuação)

UF	Município	Número de firmas selecionadas
São Paulo	Campinas	4.895
Santa Catarina	Florianópolis	4.293
Amazonas	Manaus	4.201
Paraná	Londrina	3.802
Paraná	Maringá	3.673
Santa Catarina	Joinville	3.649
Rio Grande do Sul	Caxias do Sul	3.544
São Paulo	Guarulhos	3.298
Santa Catarina	Blumenau	2.926
Pará	Belém	2.896
São Paulo	Ribeirão Preto	2.701
Mato Grosso	Cuiabá	2.679
São Paulo	Barueri	2.673
São Paulo	São Bernardo do Campo	2.556
Mato Grosso do Sul	Campo Grande	2.535
São Paulo	Sorocaba	2.272
Rio Grande do Norte	Natal	2.129
São Paulo	Santos	2.121
São Paulo	São José dos Campos	1.998
Rio Grande do Sul	Novo Hamburgo	1.982
São Paulo	Santo André	1.960
Paraná	Cascavel	1.936
Minas Gerais	Uberlândia	1.922
Maranhão	São Luís	1.887
Alagoas	Maceió	1.885
Santa Catarina	Itajaí	1.883
São Paulo	São José do Rio Preto	1.780
São Paulo	Jundiaí	1.734
Santa Catarina	São José	1.714
Paraíba	João Pessoa	1.705
Espírito Santo	Vitória	1.682
Rio de Janeiro	Niterói	1.679
Minas Gerais	Contagem	1.670
Paraná	São José dos Pinhais	1.656
Paraná	Ponta Grossa	1.635
Piauí	Teresina	1.585
Santa Catarina	Balneário Camboriú	1.572
São Paulo	Piracicaba	1.558
Rio Grande do Sul	Canoas	1.501
São Paulo	Osasco	1.440

Elaboração dos autores.

Em termos de natureza jurídica, de acordo com a classificação do IBGE, a preponderância é claramente de empresas de capital fechado (tabela 3). Todavia, o empresário ou empresa individual, juntos, representam mais de 17% do total e aparecem em segundo e terceiro lugares.

TABELA 3
Natureza jurídica das firmas selecionadas

Código	Descrição	Número de firmas selecionadas
2062	Sociedade empresária limitada	323.098
2135	Empresário (individual)	58.660
2305	Empresa individual de responsabilidade limitada (de natureza empresária)	28.268
3999	Associação privada	17.074
3085	Condomínio edilício	15.260
2054	Sociedade anônima fechada	12.103
2240	Sociedade simples limitada	11.570
1031	Órgão público do Poder Executivo municipal	7.840
2143	Cooperativa	2.458
3131	Entidade sindical	2.252
2232	Sociedade simples pura	2.215
1066	Órgão público do Poder Legislativo municipal	1.937
4081	Contribuinte individual	1.691
2151	Consórcio de sociedades	1.689
3069	Fundação privada	1.324
1120	Autarquia municipal	973
3034	Serviço notarial e registral (cartório)	960
1023	Órgão público do Poder Executivo estadual ou do Distrito Federal	917
2046	Sociedade anônima aberta	891
1015	Órgão público do Poder Executivo federal	610
2313	Empresa individual de responsabilidade limitada (de natureza simples)	514
1104	Autarquia federal	506
3220	Organização religiosa	499
2011	Empresa pública	445
2038	Sociedade de economia mista	441
1112	Autarquia estadual ou do Distrito Federal	343
1155	Fundação pública de direito público municipal	332
4014	Empresa individual imobiliária	311
4120	Produtor rural (pessoa física)	272
1210	Consórcio público de direito público (associação pública)	243
3077	Serviço social autônomo	216
2283	Consórcio de empregadores	179

(Continua)

(Continuação)

Código	Descrição	Número de firmas selecionadas
1244	Município	167
2178	Estabelecimento, no Brasil, de sociedade estrangeira	145
1082	Órgão público do Poder Judiciário estadual	135
1147	Fundação pública de direito público estadual ou do Distrito Federal	134
5029	Representação diplomática estrangeira	133
2127	Sociedade em conta de participação	98
1074	Órgão público do Poder Judiciário federal	93
2321	Sociedade unipessoal de advogados	86
2070	Sociedade empresária em nome coletivo	65
4022	Segurado especial	57
1058	Órgão público do Poder Legislativo estadual ou do Distrito Federal	54
2089	Sociedade empresária em comandita simples	53
3204	Estabelecimento, no Brasil, de fundação ou associação estrangeiras	47
2160	Grupo de sociedades	47
1139	Fundação pública de direito público federal	45
3263	Órgão de direção regional de partido político	42
1333	Fundo público da administração direta municipal	42

Elaboração dos autores.

4 CONSIDERAÇÕES FINAIS

Este capítulo apresenta resultados de uma classificação simples de empresas presentes na Rais, a partir da sua semelhança com firmas que foram financiadas pelo BNDES. A separação do topo da classificação sugere quais empresas poderiam se tornar alvo de financiamento, se mantidos a lógica e o procedimento observados na década de 2010. A caracterização indica claramente que o conjunto selecionado é composto por empresas grandes e tradicionais, localizadas em São Paulo, nas grandes capitais, no Sul e no Sudeste. Todavia, a distinção por natureza jurídica também aponta para um grupo relevante de empresários individuais.

Este foi um exercício simples. É possível detalhar e expandir a análise, de modo a experimentar outros cortes de classificação no topo, outros métodos e parâmetros de aprendizado de máquinas ou outras variáveis disponíveis para o conjunto das empresas e de interesse do BNDES. Também é possível, com acesso a ambientes computacionais mais robustos, a expansão da base amostral para maior número de anos.

REFERÊNCIAS

BREIMAN, L. Rejoinder: arcing classifiers. **The Annals of Statistics**, v. 26, n. 3, p. 841-849, 1998.

_____. **Manual on setting up, using, and understanding random forests V3.1**. Berkeley: University of California Berkeley, 2002.

LAVIN, A. *et al*. Simulation intelligence: towards a new generation of scientific methods. **Arxiv**, 6 Dec. 2021. Disponível em: <https://arxiv.org/abs/2112.03235>.

MISHINA, Y. *et al*. Boosted random forest. **Ieice Transactions on Information and Systems**, v. E98-D, n. 9, p. 1630-1636, 2015.

REN, Q.; CHENG, H.; HAN, H. Research on machine learning framework based on random forest algorithm. **AIP Conference Proceedings**, v. 1820, n. 1, 2017.