LUCIENE MOREIRA LUCAS

FISCALIZAÇÃO DA ARRECADAÇÃO DE CFEM

Classificador para seleção de processos

Brasília

2019

LUCIENE MOREIRA LUCAS

FISCALIZAÇÃO DA ARRECARDAÇÃO DE CFEM

Classificador para seleção de processos

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.
Orientador: Prof. Dr. Eduardo Chaves Ferreira

Brasília

2019

REFERÊNCIA BIBLIOGRÁFICA

LUCAS, Luciene M. **Fiscalização da arrecadação CFEM: classificador para seleção de processos**. 2019. Trabalho de Conclusão de Curso (Especialização em Análise de dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF. 59 fl.

CESSÃO DE DIREITOS

NOME DO AUTOR: Luciene Moreira Lucas
TÍTULO: Fiscalização da arrecadação de CFEM: Classificador para seleção de processos-
GRAU/ANO: Especialista/2019

_____

Luciene Moreira Lucas
luciene.lucas@cgu.gov.br

Lucas, Luciene Moreira
Fiscalização da arrecadação de CFEM: classificador para seleção das mineradoras/ Luciene Moreira Lucas; orientador, Eduardo Chaves Ferreira, 2019.
59.

Trabalho de Conclusão de Curso (especialização) – Instituto Serzedello Corrêa, Curso de Análise de dados para o controle, Brasília, 2019.

1.Ciência de dados. 3. Royalties da Mineração. I. Ferreira, Eduardo Chaves. II. Instituto Serzedello Corrêa. Análise de dados para o controle. III. Título

LUCIENE MOREIRA LUCAS

FISCALIZAÇÃO DA ARRECADAÇÃO DE CFEM:

Classificador para seleção dos processos

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de dados para o controle, realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.

Brasília, 23 de março de 2020.

Banca Examinadora:

_____

Eduardo Chaves Ferreira, Dr.

Orientador

Tribunal de Contas da União

_____

Eduardo Álvaro Pinto de Freitas Neto

Agência Nacional de Mineração

AGRADECIMENTOS

Minha gratidão Karen pelo apoio e ajuda sem a qual não teria conseguido esse título.

Carlos, obrigado por todo o suporte, dedicação e paciência.

Agradeço aos meus colegas e professores do curso que me deram apoio e ajuda sempre que necessário.

A todos os amigos, familiares e colegas de trabalho que tiveram que conviver com minha ausência e estresse durante esse período.

“Importante não é ver o que ninguém nunca viu, mas sim, pensar o que ninguém nunca pensou sobre algo que todo mundo vê.”

Arthur Schopenhauer

RESUMO

O setor mineral possui grande importância na economia brasileira e, apesar disso, várias dificuldades e falhas vem ocorrendo na arrecadação dos royalties ligados à área. O trabalho teve por objetivo apresentar esse cenário e analisar as informações e dados disponíveis que pudessem auxiliar na otimização da fiscalização da arrecadação da CFEM. Foi levantando sistemas e bases de dados que contivessem informações que influenciassem direta ou indiretamente no comportamento fiscal das empresas mineradoras. Partindo dos dados referentes as fiscalizações realizadas pela ANM e de posse das outras informações levantadas, realizou-se uma tentativa de classificar, por meio de técnicas de mineração de dados, os processos minerários de acordo com a proporção entre o que se deixa de arrecadar e o valor corretamente apurado. Dificuldade relacionadas à qualidade dos dados e sigilo fiscal foram impostas ao trabalho. Devido a esses e outros fatores o classificador apresentou um baixo desempenho. servindo esse trabalho de ponto de partido para melhora e desenvolvimento de ferramentas eficazes para auxílio na fiscalização da arrecadação da CFEM.

Palavras-chave: Royalties. CFEM. Mineração. Mineração de dados. Classificação.

ABSTRACT

The mineral sector has a great importance in the Brazilian economy and, despite of this, several difficulties and failures occur in the tax collection of mineral royalties. The object of the work is to present this scenario and analyze what information and data are available and how they could allow the optimization of CFEM collection. It was search for systems and databases that contain information that directly or indirectly influences the mining company's fiscal behavior. Based on inspections data carried out by the ANM and combined with other collected information, an attempt was made to classify, using data mining techniques, the mining processes according to the proportion between the tax evasion and the amount correctly calculated. Difficulty related to data quality and fiscal secrecy were imposed on the work. Due to these and to other factors, the classifier had a bad performance. However, this work could inspire and be a start point to improve e development effective tools to help the inspection of the collection of CFEM.

Keywords: Royalties. CFEM. Mining. Data mining. Classification.

LISTA DE ILUSTRAÇÕES (opcional)

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT Associação Brasileira de Normas Técnicas

AGU Advocacia Nacional da União

ANATEL Agência Nacional de Telecomunicações

ANEEL Agência Nacional de Energia Elétrica

ANM Agência Nacional de Mineração

CAPAG Capacidade de Pagamento

CETEM Centro de Tecnologia Mineral

CFEM Compensação Financeira pela Exploração de Recursos Minerais

CFOP Código Fiscal de Operações e Prestações

CGU Controladoria-Geral da União

CNAE Classificação Nacional de Atividades Econômicas

CNPJ Cadastro Nacional de Pessoa Jurídica

CBO Classificação Brasileira de Ocupações

CONFAZ Conselho Nacional de Política Fazendária

CPF Cadastro de Pessoa Física

CRISP-DM Cross Industry Standart Process for Data Mining

DNPM Departamento Nacional de Produção Mineral

FNDCT Fundo Nacional de Desenvolvimento Científico e Tecnológico

GED Gerenciamento Eletrônico de Documentos

IBAMA Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis

IBGE Instituto Brasileiro de Geografia e Estatística

ICMS Imposto sobre Circulação de Mercadoria e Serviços

ISS Imposto sobre Serviços

MIT Instituto de Tecnologia de Massachusetts

NCM Nomenclatura Comum do Mercosul

NF-e Nota Fiscal Eletrônica

OCDE Organização para a Cooperação e Desenvolvimento Econômico

PIB Produto Interno Bruto

PLG Permissão de Lavra Garimpeira

RAIS Relatório Anual de Informações Sociais

RAL Relatório Anual de Lavra

RENAVAM Registro Nacional de Veículos Automotores

RFB Receita Federal do Brasil

RRF Regime de Recuperação Fiscal

SAPIENS Sistema de Apoio à Procuradoria Inteligente

SCM Sistema de Cadastro Mineiro

SEI Sistema Eletrônico de Informações

SICONFI Sistema de Informações Contábeis e Fiscais

STN Secretária do Tesouro Nacional

TCC Trabalho de Conclusão de Curso

TCU Tribunal de Contas da União

TI Tecnologia da Informação

TRF4 Tribunal Regional Federal da 4ª Região

SUMÁRIO

1 INTRODUÇÃO

O objetivo do trabalho é criar um modelo de classificação utilizando técnicas de mineração de dados no qual, de acordo com características das mineradoras e/ou dos títulos minerários, se possa definir o grau de impacto do comportamento fiscal da empresa em relação à arrecadação de Compensação Financeira pela Exploração de Recursos Minerais - CFEM a fim de servir de subsidio para o planejamento das fiscalizações em relação a esta Compensação.

Considerando que o pagamento é realizado com base em informações autodeclaratórias pretende-se mapear e analisar sistemas, bases de dados e fontes de informações existentes na Agência Nacional de Mineração - ANM e em outros Órgãos e Esferas do Governo que contenham informações sobre:

- as empresas mineradoras;
- os títulos minerários;
- comercialização dos bens minerários.

A fim de alcançar tal objetivo, será dado inicialmente um panorama do setor mineral brasileiro, mostrando sua importância para a economia, uma vez que a atividade de extração mineral é responsável por quase 5% do PIB (Produto Interno Bruto) do Brasil e por cerca de 20% do valor de exportações. Entrando em detalhes também sobre a CFEM, mais conhecida como royalties da mineração, de modo a possibilitar a compreensão de quais informações podem influenciar, direta ou indiretamente, na sua arrecadação.

Diante das várias falhas e fragilidades encontradas pelo Órgão de Controle na arrecadação de CFEM, será explanado sobre os desafios, necessidades e potencial de arrecadação da mesma, bem como os possíveis benefícios da utilização da Tecnologia da Informação para otimizar a fiscalização da arrecadação.

Nos capítulos seguintes será detalhada uma proposta para otimização da seleção de processos a serem fiscalizados utilizando os dados referentes às fiscalizações já realizadas pela ANM. Será explicado quais informações foram utilizadas para a construção do classificador, suas fontes, motivações da escolha e restrição impostas ao trabalho, bem como todo o processo de limpeza e processamento dos dados.

Por fim, será tratado sobre os algoritmos testados e resultados obtidos na aplicação destes, bem como possíveis trabalhos futuros decorrentes deste TCC.

2 PANORAMA GERAL

Quando se pensa em produto em mineração, a primeira coisa que se costuma vir à mente são pedras e metais preciosos, contudo, os produtos provenientes do extrativismo mineral também estão presentes no nosso dia a dia por meio da água que bebemos, dos adubos e fertilizantes utilizamos na agricultura, nos materiais da construção civil como areia, brita, pedras ornamentais, etc.

Responsável por quase 5% do PIB brasileiro, o setor mineral fechou o ano de 2018 com superávit de US$ 23,3 bilhões (US$ 49,8 bilhões em exportações e US$ 26,4 bilhões em importações). O valor exportado representou 20,8% do total das exportações brasileiras no ano, de US$ 239,9 bilhões. Isoladamente, as exportações da mineração foram de US$ 25,2 bilhões, representando 10,5% das exportações brasileiras e 50,6% das exportações do setor mineral.

O Brasil produz mais de oitenta diferentes substâncias minerais e o setor mineral é responsável por cerca de 750.000 empregos formais diretos, sendo 22% deles na mineração e 78% na indústria da transformação mineral.

Mais conhecida como os royalties da mineração, a Compensação Financeira pela Exploração de Recursos Minerais – CFEM é uma receita originária patrimonial da União devida pelo aproveitamento econômico dos recursos minerais e cujo regime jurídico encontra-se delimitado pela Constituição Federal, nos seus artigos 5 , II; 20, IX, e § 1º ; 176; 37; 155, X, “b” e 225, § 2º.

O fato gerador da obrigação de pagar é o ato da venda ou transferência para utilização do bem mineral, sua arrematação em hasta pública ou consumo pelo minerador e sua base de cálculo, conforme art. 2º da Lei nº 8.001/1990, alterada pelo Lei nº 13.540/2017, é:

- a receita bruta da venda, deduzidos os tributos incidentes sobre sua comercialização;
- a receita bruta calculada, considerado o preço corrente do bem mineral, ou de seu similar, no mercado local, regional, nacional ou internacional, conforme o caso, ou o valor de referência, definido a partir do valor do produto final obtido após a conclusão do respectivo processo de beneficiamento em caso de consumo;
- nas exportações, sobre a receita calculada, considerada como base de cálculo, no mínimo, o preço parâmetro definido pela Receita Federal do Brasil - RFB do Ministério da Economia, com fundamento no art. 19-A da Lei nº 9.430, de 27 de dezembro de 1996, e na legislação complementar, ou, na hipótese de inexistência do preço

parâmetro, será considerado o valor de referência, observado o disposto nos §§ 10 e 14 deste artigo;

- na hipótese de bem mineral adquirido em hasta pública, sobre o valor de arrematação;
- na hipótese de extração sob o regime de permissão de lavra garimpeira, sobre o valor da primeira aquisição do bem mineral.

A alíquota sobre a base de cálculo vária de 1,0% a 3,5%, de acordo com a substância extraída, e o lançamento se dá por homologação, ou seja, a legislação atribui ao sujeito passivo o dever de antecipar o pagamento sem prévio exame da autoridade administrativa e opera-se pelo ato em que a referida autoridade, tomando conhecimento da atividade assim exercida pelo obrigado, expressamente a homologa.

Conforme Lei nº 13.575/2017, é da Agência Nacional de Mineração – ANM a competência para regular, fiscalizar, arrecadar, constituir e cobrar os créditos decorrentes da CFEM, de que trata a Lei 7.7990/1989, e é da Procuradoria Federal junto à autarquia, a inscrição do débito em dívida ativa e o ajuizamento da competente execução.

O prazo de decadência para que a ANM constitua créditos de CFEM é de 10 anos, sendo de 5 o prazo prescricional, de acordo com o art. 47 da Lei nº 9.636/1998. Em que pese a CFEM constituir obrigação pecuniária de natureza não tributária, sua cobrança obedecerá ao rito estabelecido na Lei nº 6.830/80, por se tratar de receita financeira de autarquia federal.

Para o regular aproveitamento desses bens, mediante pesquisa, lavra, beneficiamento, distribuição, consumo ou utilização, sua exploração deve se adequar a um destes regimes a ser concedido pela ANM:

- Autorização de pesquisa: visa à realização dos trabalhos necessários à definição da jazida, sua avaliação e a determinação da exequibilidade do seu aproveitamento;

- Guia de utilização: título provisório e precário o qual admite, em caráter excepcional, a extração de substâncias minerais em área titulada, antes da outorga da concessão de lavra, fundamentado em critérios técnicos, ambientais e mercadológicos, mediante prévia autorização da ANM;

- Concessão: visa à realização do conjunto de operações coordenadas objetivando o aproveitamento industrial da jazida, da extração do minério até seu beneficiamento;
- Licenciamento: visa ao aproveitamento das substâncias minerais de emprego imediato na construção civil ou como corretivo de solos na agricultura;
- Permissão de Lavra Garimpeira (PLG): visa à lavra e aproveitamento imediatos de substâncias minerais que, em razão da sua dimensão, natureza, localização e utilização econômica, independem de prévios trabalhos de pesquisa;
- Monopolização: depende de lei especial e os trabalhos são executados direta ou indiretamente pelo Poder Executivo Federal; e
- Registro de Extração: permite aos órgãos da administração direta e autárquica da União, dos Estados, do DF e dos Municípios extrair substâncias minerais de emprego imediato na construção civil para uso exclusivo em obras públicas por eles executadas diretamente, respeitados os direitos minerários em vigor nas áreas onde devam ser executadas as obras e vedada sua comercialização.

Nos termos do art. 20, § 1º, da Constituição de 1988, é assegurada aos Estados, ao Distrito Federal, aos Municípios e aos órgãos da administração da União, a participação no resultado da exploração de recursos minerais no respectivo território, a título de compensação financeira. Conforme art. 2º, § 2º, da Lei 8.001/1990, com redação dada pela Lei 13.540/2017, a distribuição entre os entes federativos se dá da seguinte maneira:

- **7%** para a entidade reguladora do setor de mineração, no caso a ANM;
- **1%** para o Fundo Nacional de Desenvolvimento Científico e Tecnológico - FNDCT, instituído pelo Decreto-Lei no 719, de 31 de julho de 1969, e restabelecido pela Lei no 8.172, de 18 de janeiro de 1991, destinado ao desenvolvimento científico e tecnológico do setor mineral;
- **1,8%** para o Centro de Tecnologia Mineral - Cetem, vinculado ao Ministério da Ciência, Tecnologia, Inovações e Comunicações, criado pela Lei no 7.677, de 21

de outubro de 1988, para a realização de pesquisas, estudos e projetos de tratamento, beneficiamento e industrialização de bens minerais;

- **0,2%** para o Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis - Ibama, para atividades de proteção ambiental em regiões impactadas pela mineração;
- **15%** para o Distrito Federal e os Estados onde ocorrer a produção;
- **60%** para o Distrito Federal e os Municípios onde ocorrer a produção;
- **15%** para o Distrito Federal e os Municípios, quando afetados pela atividade de mineração e a produção não ocorrer em seus territórios, [...]:

Os maiores Estados arrecadadores são Minas Gerais e Pará, os quais tiveram uma arrecadação que representou 85,83% do total de recolhimento de CFEM em 2018, sendo o principal produto o minério de ferro.

A arrecadação nos últimos 5 anos se deu conforme quadro abaixo:

Tabela 1 – Arrecadação CFEM de 2014 a 2018

Exercício	2014	2015	2016	2017	2018
Arrecadação	1.578.310.171,91	1.406.521.316,13	1.469.653.013,09	1.892.922.428,89	3.172.357.212,48

Fonte: Site ANM

3 DESAFIOS DA FISCALIZAÇÃO DA ARRECADAÇÃO DA CFEM

Em que pese o aumento de arrecadação constatada ao longo dos anos, principalmente a partir de 2018, quando se mudou a base de cálculo, a qual antes era sobre o faturamento líquido e passou a ser sobre o bruto deduzidos os impostos incidentes sobre a comercialização do produto, foi verificado em auditorias anteriores que a receita decorrente da CFEM está abaixo do verdadeiro potencial de arrecadação da mesma.

No relatório do Acórdão 1979/2014 – Plenário, cujo objeto, dentre outros, foi a arrecadação da CFEM, o Tribunal de Contas da União - TCU identificou vários problemas, destacando-se os seguintes:

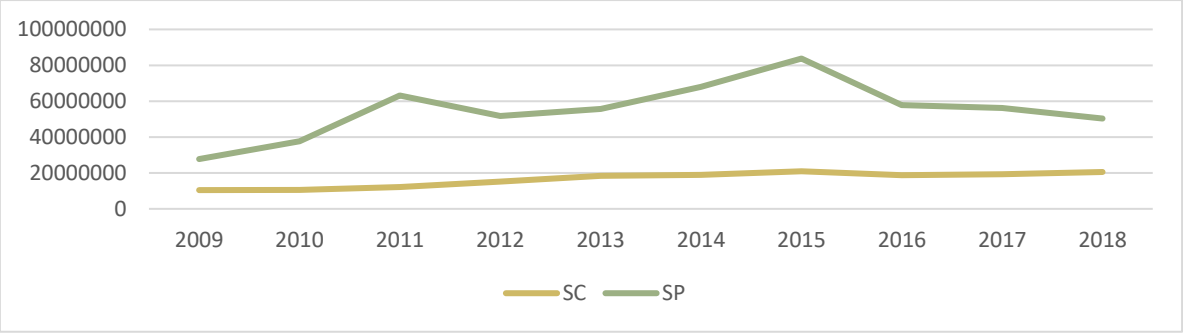
- Os dados coletados revelam que apenas 22,7% do valor da CFEM devida pelas mineradoras fiscalizadas in loco pelas superintendências[1] do ANM entre 2009 e 2011 foram recolhidos espontaneamente;
- Das 101 empresas cujos processos de fiscalização foram analisados, apenas uma encontrava-se quite com suas obrigações, sendo que 59 (29% do universo fiscalizado) não efetuaram nenhum recolhimento espontâneo, muitas por mais de 10 anos;
- Entre os anos de 2009 e 2012 (até setembro), os créditos de 1.931 processos sofreram decadência, sendo 1.024 nas procuradorias junto ao Departamento Nacional de Produção Mineral - DNPM[2] e 907 nos órgãos de execução da Procuradoria-Geral Federal;
- Montante da CFEM: entre os anos de 2009 a 2012, R$ 5.831.741,13 não foram destinados aos Estados e municípios. Muitos mineradores recolhem a mencionada compensação sem o registro de todos os dados necessários à destinação dos recursos (número do título minerário, substância extraída, município beneficiário etc.).

Considerando que em 2019 a arrecadação de CFEM foi de R$ 4.504.238.668,90 e aplicando o percentual levantado pelo TCU, temos que esse valor poderia chegar a quase 20 bilhões de reais por ano, sendo preferível que as empresas o fizessem espontaneamente uma vez que, devido aos infindáveis debates, no Judiciário, acerca da base de cálculo da CFEM e da pertinência das deduções, a recuperação de passivos após a fiscalização é de apenas 2% em média.

Apesar das dificuldades de recuperação do passivo, é importante ressaltar que os reflexos das fiscalizações também podem ser observados indiretamente. Por exemplo: a arrecadação da CFEM no Estado de São Paulo, que vinha crescendo continuamente desde 2006, atingiu seu ápice em 2014 e passou a apresentar ligeira queda, ano exato em que se deixou de realizar fiscalizações por falta de pessoal.

[1] Superintendência era o nome utilizado na época do DNMP. Após a criação da ANM as unidades localizadas no Estados passaram a ser denominadas Gerências Regionais e Unidades Avançadas. Como serão utilizados dados até 2018, os quais foram produzidos pelo antigo DNPM, manteremos a nomenclatura Superintendência e sua respectiva estrutura a época, ao longo do TCC.
[2] Com a publicação, em 28/11/2018, do Decreto nº 9.587, foi criada a ANM em substituição ao DNPM.

Figura1: Arrecadação CFEM em SC e SP nos últimos 10 anos



Fonte: Página http://dados.gov.br/dataset/sistema-arrecadacao, em 19/09/2019.

Essa verificação se torna mais preocupante quando se nota que a diminuição no número de fiscalizações ou mesmo suspensão total, como ocorreu em São Paulo (desde 2014) e Santa Catarina (desde 2016), é um fenômeno nacional.

Logo, efetivar esse potencial de arrecadação também se faz bastante necessário quando vislumbramos a atual conjuntura econômica adversa que atinge todos os entes federados, sendo Estados e Municípios os mais fortemente afetados, mas também os principais beneficiados com os royalties da mineração, pois 90% da receita da CFEM são destinadas a eles.

Conforme Boletim de Finanças dos Entes Subnacionais, publicado em agosto de 2019 pela Secretaria do Tesouro Nacional - STN, o qual faz um retrato da situação fiscal de Estados e Municípios e analisa os principais fatores que influenciaram o desempenho desses números, Minas Gerais, o maior arrecadador de CFEM e responsável por 43% do total foi avaliado na análise da capacidade de pagamento (CAPAG[3]) como D, nota mais baixa do ranking.

Ainda de acordo com a CAPAG, São Paulo e Pará, que estão entre os principais arrecadadores de CFEM, possuem nota de capacidade de pagamento B, a qual permite que o Ente receba garantia da União para novos empréstimos, entretanto Bahia e Goiás, que também estão entre os principais arrecadadores, ganharam nota C.

[3] A análise da capacidade de pagamento apura a situação fiscal dos Entes Subnacionais que querem contrair novos empréstimos com garantia da União. O intuito da CAPAG é apresentar de forma simples e transparente se um novo endividamento representa risco de crédito para o Tesouro Nacional. A metodologia do cálculo, dada pela Portaria MF nº 501/2017, é composta por três indicadores: endividamento, poupança corrente e índice de liquidez. Logo, avaliando o grau de solvência, a relação entre receitas e despesa correntes e a situação de caixa, faz-se diagnóstico da saúde fiscal do Estado ou Município. Os conceitos e variáveis utilizadas e os procedimentos a serem adotados na análise da CAPAG foram definidos na Portaria STN nº 882/2018. A nota do CAPAG vai de A à D, sendo que entes avaliados como C ou D são considerados sem capacidade de pagamento.

Entre os 5.569 municípios brasileiros, 868 possuem nota A, 774 a nota B, 2.400 a nota C e 1.527 não apresentaram dados suficientes para o cálculo da CAPAG neste ano, sendo São Paulo a capital mais endividada, mas também é que a obtém a maior parcela de sua receita total com arrecadação própria.

Dessa forma esse trabalho poderia apontar uma fonte de recursos para a recuperação fiscal destes entes federados sem a necessidade de empréstimos ou criação de novos impostos/contribuições, sem aumento de alíquota dos já existentes e sem submetê-los às exigências e vedações previstas no Regime de Recuperação Fiscal – RRF, Lei Complementar nº 159/2019.

Para se ter ideia do impacto dessa receita no orçamento dos Municípios apresento tabela comparando a receita municipal total do exercício 2018 e a receita decorrente da CFEM no ano correspondente e o percentual que esta representa no orçamento do Município em relação aos vinte maiores Municípios extratores de substâncias minerais:

Tabela 2 – Proporção da receita da CFEM em relação ao Orçamento do Município

Munícipio	**Receitas Municipais**	**CFEM**	**%CFEM**
Parauapebas – PA	1.347.757.449,53	400.551.963,90	29,7%
Canaã dos Carajás – PA	381.743.190,60	177.274.396,30	46,4%
Nova Lima – MG	628.089.061,00	98.920.170,73	15,7%
Itabira – MG	591.958.220,51	98.044.283,56	16,6%
Congonhas – MG	454.075.937,02	96.378.003,66	21,2%
Marabá – PA	1.004.974.485,80	76.420.659,38	7,6%
Itabirito – MG	312.151.678,81	74.410.644,16	23,8%
São Gonçalo do Rio Abaixo	164.978.649,90	60.418.316,80	36,62
Mariana – MG	284.830.467,00	61.757.485,18	21,7%
Brumadinho – MG	159.000.752,28	35.680.083,39	22,4%
Conceição do Mato Dentro – MG	99.303.074,58	23.567.489,68	23,7%
Itatiaiuçu – MG	76.724.743,10	21.665.425,31	28,2%
Paracatu – MG	297.239.652,67	21.483.498,76	7,2%
Ouro Preto – MG	297.284.136,34	21.302.507,09	7,2%
Alto Horizonte – GO	101.046.684,20	19.409.780,82	19,2%
Paragominas – PA	387.078.530,23	19.137.574,82	4,9%
Resende – RJ	635.725.819,04	17.000.463,41	2,7%
Belo Vale – MG	57.858.538,91	16.025.406,64	27,7%
Oriximiná – PA	244.311.972,46	15.765.726,03	6,5%
Curionópolis – PA	95.853.041,33	14.069.081,77	14,7%
Rio Piracicaba – MG	52.999.613,16	11.956.805,83	22,6%

Fonte: Siconfi

Além da questão econômica, temos a situação da ANM, cujos trabalhos da Controladoria Geral da União - CGU e TCU demonstraram, reiteradamente as deficiências e carências estruturais do Órgão, em que pese este ser superavitário (R$ 334.129.606,81 em despesa realizada em 2018 e receita de R$ 344.924.693,63 relativa a sua parcela de CFEM, além de outras taxas e receitas arrecadadas) tem margem para melhorias e aumento da arrecadação.

Dentre as principais fragilidades identificadas podemos citar:

- Carência crônica de pessoal (nos últimos vinte anos foram realizados apenas dois concursos públicos para provimento de cargos – o último foi há mais de seis anos);
- Quadro de pessoal com elevada idade média dos servidores (nos próximos cinco anos há potencial redução de metade da força de trabalho em razão de aposentadorias);
- Carência de recursos logísticos, materiais e orçamentário-financeiros necessários ao desempenho de suas atividades (os contingenciamentos dos recursos e incertezas quanto às datas de efetiva disponibilidade afetam negativamente seu desempenho);
- Sistemas corporativos em uso (Tecnologia da Informação-TI) insuficientes e falhos.

Como uma das consequências dessas fragilidades temos a diminuição na quantidade de fiscalizações realizadas (2183 em 2014 para 387 em 2018), sendo que o resultado das fiscalizações costuma ser aumento no valor da CFEM devida pela empresa mineradora, mesmo os critérios de seleção sendo limitados a dois fatores principais: prazo decadencial (10 anos) e materialidade (histórico de arrecadação da empresa).

A eficácia desses critérios também é questionável, pois focar em débitos de quase uma década ao invés de ilícitos recentes pode afetar a eficiência e eficácia da cobrança. Conforme relatório de auditoria da CGU nº 201801505[4] somente 2% dos débitos de CFEM apurados nas fiscalizações da ANM em Santa Catarina são pagos ou parcelados pelas empresas notificadas.

O mesmo relatório aponta que:

> "...embora não haja esta formalização do processo de planejamento e avaliação formal dos riscos para a escolha dos mineradores, a partir do cotejamento dos dados das fiscalizações in loco realizadas no período de 2011 a 2016 com as informações dispo-

[4] https://auditoria.cgu.gov.br/download/13010.pdf

> níveis nas bases de dados da ANM, foi possível inferir que tais fiscalizações priorizaram substâncias e/ou empresas/mineradores com maior arrecadação de CFEM no exercício em análise.

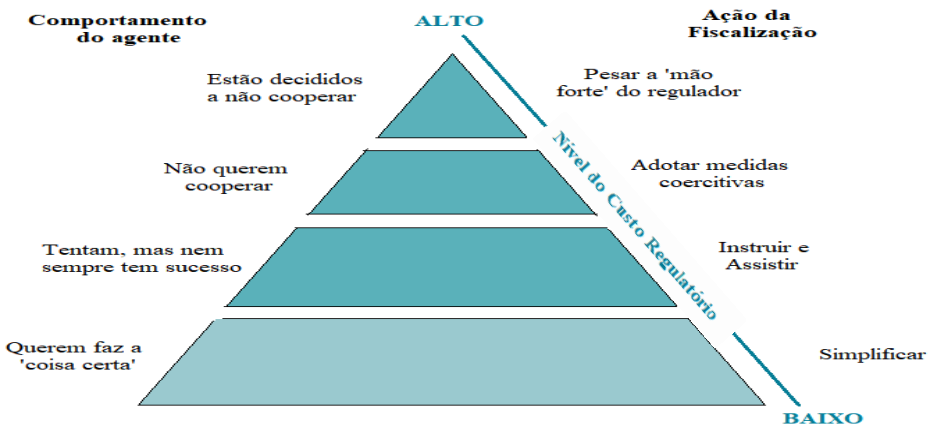
Outrossim, nos casos em que se voltaram para fiscalizar mineradores/substâncias com menor arrecadação, como é o caso das empresas que atuam com a substância água, no geral, os resultados da fiscalização evidenciaram diferença de CFEM significativa."

Portanto, diante de um quadro de escassez de recursos que dificilmente se reverterá em um curto prazo e das fragilidades no planejamento das fiscalizações realizadas pela ANM é necessário que esta reveja seus métodos de modo a otimizar o trabalho por ela desenvolvido e o uso de recursos tecnológicos poderia auxiliar no planejamento, seleção de amostra e execução das fiscalizações.

Agências regulatórias como Aneel e Anatel vêm implementando em suas fiscalizações a teoria da regulação responsiva, a qual tem como essência a modulação de sua atuação de acordo com o risco e com o comportamento cooperativo ou não do fiscalizado.

Pensado inicialmente por Ian Ayres e Jonh Braithwate, o conceito normalmente é representado por meio de uma pirâmide de fiscalização, similar a apresentada abaixo durante o IX Congresso Brasileiro de Regulação, e na qual de um lado está representado o comportamento do agente fiscalizado e do outro as respectivas medidas a serem tomadas pelo agente fiscalizador em resposta ao primeiro, de forma que quanto mais inadequado o comportamento mais rígida é a ação e quanto melhor o comportamento mais simples é o processo de monitoramento:

Figura 2: Regulação responsiva



Fonte: Aneel

Além disso, ele está em consonância com os princípios promovidos pela Organização para a Cooperação e Desenvolvimento Econômico – OCDE, dentre os quais:

- Foco no risco e proporcionalidade: a frequência das inspeções e dos recursos utilizados devem ser proporcionais ao nível de risco e as ações de execução devem ter por objetivo reduzir o risco real colocado pelas infrações;
- Atuação baseada em evidências: decidir o que inspecionar e como deve ser fundamentada em dados e evidências, e os resultados devem ser avaliados regularmente.
- Integração de informação: as tecnologias da informação e da comunicação devem ser utilizadas para maximizar a concentração de riscos, a coordenação e a partilha de informações, bem como otimização da utilização recursos.

4 DESENVOLVENDO O CLASSIFICADOR

Utilizarei como guia para desenvolvimento do modelo de classificação o método Cross-Industry Standart Process for Data Mining - CRISP-DM (CHAPMAN et al, 2000), o qual consiste em seis etapas principais:

- Entendimento do negócio: fase inicial que visa obter conhecimento sobre os objetivos do negócio e seus requisitos, e então converter esse conhecimento em uma definição de um problema de mineração de dados;
- Compreensão dos dados: essa fase se inicia com uma coleta inicial de dados e com procedimentos e atividades visando a familiarização com os dados, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formar hipóteses.
- Limpeza dos dados: o objetivo é a limpeza, transformação, integração e formatação dos dados da etapa anterior. É a atividade peça qual os ruídos, dados estranhos ou inconsistentes são tratados, sendo a fase que exige mais esforço, correspondente geralmente a mais de 50% do trabalho de mineração de dados;
- Modelagem de dados: aplicação de técnicas de modelagem sobre o conjunto de dados preparados na etapa anterior. Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para se obter valores otimizados;

- Avaliação do processo: avaliar e rever os passos executados de modo que o modelo permita o alcance dos objetivos e se necessário retornar a qualquer um dos estágios anteriores.
- Execução: uma vez que o modelo tenha sido construído e validado, é preciso colocado em execução. Deve-se documentar e apresentar os resultados de maneira compreensível ao usuário, sendo importante monitorar os resultados e adaptar o modelo sempre que necessário.

Considerando que uma das características da regulação responsiva é utilizar o comportamento do fiscalizado como balizador da atuação futura corretiva e de monitoramento, para se alcançar o objetivo proposto para o TCC servirão de ponto de partida os dados sobre as fiscalizações efetuadas pela ANM no período entre 2014 e 2018.

Seguindo os extratos da pirâmide de regulação responsiva, dividirei os processos em três níveis de subarrecadação esperados após a fiscalização:

- **Baixo**: abaixo de 30% de diferença entre o valor recolhido espontaneamente e o valor apurado após a fiscalização;
- **Médio**: entre 30% e 90% de diferença entre o pago e o apurado;
- **Alto**: mais de 90% de diferença.

Em seguida pretende-se mapear e analisar sistemas, bases de dados e fontes de informações existentes na ANM e em outros Órgãos e Esferas do Governo, que contenham informações/características das empresas mineradoras, dos títulos minerários e da comercialização dos bens minerários.

Com essas características/informações catalogadas e organizadas serão feitos cruzamentos das informações com as transformações e limpezas que se fizerem necessárias para, só então, aplicar alguns algoritmos de classificação de dados.

O modelo de classificação que apresentar o melhor resultado poderá ser aplicado no universo de detentores de títulos minerários ativos. Dessa forma, associado com outros indicadores, como materialidade, tempo sem fiscalização, divergências entre Relatório Anual de Lavra - RAL e informações dos boletos mensais, já usados pela ANM, pode ser traçada uma estratégia de fiscalização baseada em riscos e evidências, permitindo à Agência concentrar maiores esforços fiscalizatórios nos processos em que haja uma maior probabilidade de subarrecadação.

Portanto, o classificador poderá servir de ferramenta para atender aos princípios de Foco no Risco e Proporcionalidade e Atuação Baseada em Evidências aconselhados pelo OCDE, utilizando dessa forma de inteligência fiscal e regulação responsiva de forma similar ao que ocorre em outros órgãos do Governo Federal, como a RFB, Agência Nacional de Telecomunicação - ANATEL e Agência Nacional de Energia Elétrica - ANEEL.

4.1 SELEÇÃO E LIMPEZA DOS DADOS

O trabalho se concentrará em pessoas jurídicas, uma vez que o recolhimento por pessoa física é uma exceção que ocorre quando esta possui apenas uma licença de pesquisa, mas, durante o processo, extrai pequenas quantidades de minerais decorrentes dos estudos do solo. O recolhimento, nesse caso, se dá por meio de guia de utilização.

Não será objeto de análise os recolhimentos realizados por titulares de Permissão de Lavra Garimpeira - PLG, pois, segundo as normas, o responsável pelo recolhimento aos cofres públicos é o primeiro adquirente, não sendo obrigatório a emissão de nota fiscal de compra para todos os adquirentes, tornando o controle dos minerais extraídos sob esse regime mais difícil.

Sendo o órgão competente pela cobrança e fiscalização da CFEM, o trabalho iniciou com o levantamento dos sistemas/dados disponíveis na ANM.

O Sistema de Cadastro Mineiro - SCM[5] é a fonte primaria de informações sobre os títulos minerários e é onde os primeiros dados sobre o título minerário e as mineradoras são registrados. O Sistema de Arrecadação[6] e o RALWEB[7] estão relacionados aos dados sobre a extração de bens minerais e a cobrança e arrecadação de CFEM. O primeiro possui ainda infor-

[5] Sistema de Cadastro Mineiro (SCM): é responsável pela automação do processo de outorga, controlando requerentes (pessoas físicas ou jurídicas), solicitações, prioridades, prazos, fases do processo, ações necessárias e emissão/publicação dos documentos necessários à formalização dos atos previstos nos Código de Mineração. Este sistema, além de possibilitar o controle do ciclo de vida dos processos minerários, fornece informações essenciais aos sistemas das demais áreas finalísticas, sendo base para tomada das ações de arrecadação e fiscalização.

[6] Sistema de Arrecadação - Sistema desenvolvido para controlar o processo de arrecadação da Autarquia, contemplando funcionalidades que controlam prazos, valores devidos, valores recolhidos, conciliação e distribuição dos recursos arrecadados, conforme previsto na Legislação. Produz também os relatórios gerenciais necessários a uma efetiva gestão dos recursos oriundos da pesquisa e exploração mineral.

[7] Relatório Anual de Lavra (RALWEB): Trata-se de sistema responsável por receber as informações declaradas pelos mineradores a respeito das atividades realizadas no processo de lavra no ano referência, vinculadas a cada processo minerário de sua responsabilidade, contemplando dados da lavra, reserva, custos, valor de venda e outros.

mações sobre fiscalizações. Como apoio, também são utilizados o AMBWEB[8], SEI[9], SAPIENS[10] e Comex[11].

Foi fornecido à CGU um backup dos bancos de dados relacionados ao sistema de Arrecadação e ao SCM, contudo, foram utilizadas informações divulgadas no sítio www.dados.gov.br como fonte de informações de referência por ser um dado público e oficial.

Essa decisão foi tomada devido aos problemas verificados nas bases de dados fornecidas. Dentre algumas das principais características recomendadas pela Gestão das Informação e Qualidade de Dados[12](MATTIODA, 2006), usadas como referência para análise, podemos citar as seguintes falhas:

- **uniformidade**[13]: uso de dados com padrão de unidade de medidas e/ou nomenclaturas distintas dentro do mesmo processo ou em sistemas distintos no âmbito da Agência, sendo utilizada múltiplas nomenclaturas para substâncias iguais ou com definição científica ou comercial bastantes similares;
- **completude**[14]: em relação aos 6 bancos de dados encaminhados vinculados ao sistema de Arrecadação e ao SCM havia 597 tabelas das quais 98 delas estavam em branco e outras 9 possuíam apenas 1 linha de registro. As tabelas que continham mais linhas de informação possuíam uma grande proporção de registros faltantes. A tabela com os dados dos processos de fiscalização continha 237 registros sem número de processo e 12 sem identificação do fiscalizado. Não há dados sobre fiscalizações cuja conclusão foi pela regularidade;

---

[8] Anuário Mineral Brasileiro (AMBWEB): aplicação responsável pelo tratamento das informações recebidas através da declaração do RAL, juntamente com dados do Sistema SCM e Arrecadação, de forma a possibilitar a geração de dados estatísticos do setor mineral brasileiro e sua publicação para o público interno e externo.

[9] Sistema Eletrônico de Informações (SEI): desenvolvido pelo Tribunal Regional Federal da 4ª Região (TRF4), é uma ferramenta de gestão de documentos e processos eletrônicos, e tem como objetivo promover a eficiência administrativa.

[10] Sistema AGU de Inteligência Jurídica (SAPIENS): Gerenciador Eletrônico de Documentos (GED) que possui avançados recursos de apoio à produção de conteúdo jurídico e de controle de fluxos administrativos, focado na integração com os sistemas informatizados do Poder Judiciário e do Poder Executivo. Procura simplificar rotinas e expedientes, além de auxiliar, com suas ferramentas de inteligência, no Processo de tomada de decisão e na elaboração de documentos.

[11]. Comércio Exterior (Comex): responsável por gerar informações sobre o desempenho do setor mineral na balança comercial brasileira, incluídos dados de importação e exportação por substância

[12] *Total Data Quality Management – TDQM, desenvolvido pelo MIT*

[13] Grau de utilização de uma mesma escala, formato, padrão ou referência para representar um dado

[14] Proporção entre os dados armazenados em relação ao universo potencial de informações.

- **unicidade**[15]: 40 tabelas com nomes similares cuja diferença única nestes era a numeração nos três últimos caracteres e cujo período a que se referiam as informações eram sobrepostos não havendo documentação que explicasse a diferença entre as referidas tabelas. Existência mineradoras com mais de um número de CNPJ/CPF;
- **validade**[16]: identificados 19 mil CNPJs/CPFs iniciando ou finalizando com 9999 e 138 problemas nos dígitos verificadores
- **consistência**[17]: a fiscalização de mesa baseia-se nas divergências encontradas entre os valores declarados ao se emitir o boleto de CFEM e os dados do RAL, sendo o batimento feito de forma manual por meio de planilhas eletrônicas. Devido a transferência de direitos se dar de forma que fique registrado apenas no processo físico, foram identificados realização de pagamentos por uma determinada empresa cujo a titularidade do direito minerário estava registrada em nome de outra;
- **acurácia**[18]: de 82 mil CNPJs cadastrados em diferentes tabelas da base de dados da Agência 36 mil não foram localizados na base de CNPJs da RFB, sendo que em apenas 15 casos foi possível localizar o código correto apenas utilizando o nome constante dos cadastros. Inserção de números de CEP inexistentes. Das 2.985 empresas cujos dados no RAIS 2017 contém empregados que executam funções com CBO vinculado a atividades minerárias, apenas 411 recolheram CFEM. Das 60.740 (fonte RBF) empresas ativas cujo CNAE principal está vinculado a atividades de extração mineral, apenas 3.577 recolheram CFEM em 2018.

---

[15] É o conjunto dos vocábulos ou dos termos utilizados na descrição dos objetos modelados para o banco de dados. Os termos são dispostos com o seu respectivo significado para apresentar uma descrição textual da estrutura lógica e física do banco de dados. Quando adequadamente documentado, o dicionário de dados é uma importante ferramenta de resolução de problemas. Ele identifica para os usuários finais e para os especialistas empresariais quais dados existem no banco de dados, sua estrutura e formato, e sua utilização.

[16] Dados são válidos se estão em conformidade com a sintaxe (formato, tipo, alcance) da sua definição.

[17] A ausência de diferença, ao comparar duas ou mais representações de uma coisa contra uma definição. Grau de coerência dos dados se comparado com informações internas.

[18] O grau em que os dados descrevem corretamente o objeto ou evento do "mundo real" sendo descrito. Grau de coerência dos dados se comparado com informações externas

- **atualidade**[19] não há uma obrigatoriedade de atualização dos dados cadastrais, como por exemplo o endereço, por parte das mineradoras, conforme consta do Acórdão nº 2863/2015-TCU-Plenário. Além disso, o sistema não está atualizado para todas as ocorrências em um processo. Por exemplo, a cessão de direito não está inserida no sistema, ou o pagamento de multa feito pelo contribuinte não fica atualizado no cadastro do mineiro, muito embora no sistema de Arrecadação esteja correto; e
- **acessibilidade**[20]**:** ausência de manuais de utilização dos sistemas (exceto RAL-Web, para o qual existe o "Sistema Analisador do RAL - Manual do Usuário"), ausência de dicionário de dados para as bases de dados e sistemas.

Não obstante as falhas relatadas, utilizou-se como ponto de partida as informações contidas nos processos de cobrança iniciados entre 2014 e 2018, combinadas com informações contidas nas tabelas CFEM_Fiscalizacao_RAL, TB_CfemFiscalizacaoDetalhado e CFEM_emissao, contida no banco de dados DB_CREDITO, para extrair as seguintes informações: número do processo de cobrança, número do processo original, CPF/CNPJ, código da substância, código do Município, valor pago espontaneamente a título de CFEM, a diferença entre o valor pago e o apurado após a fiscalização.

Dos dados iniciais foram derivadas duas colunas: do número do processo original e de fiscalização calculou-se o tempo de entre a requisição do título e a realização da fiscalização e dos números de processos de cobrança derivou-se a quantidade de fiscalizações realizadas na mineradora entre 2014 e 2018.

Das tabelas de apoio vinculadas aos dados dos processos de cobrança, extraiu-se os nomes das substâncias, as classes das substâncias, nome dos Municípios e respectivas siglas dos Estado (da qual derivou a coluna Região).

As informações referentes ao número do processo original (criado quando da requisição do direito minerário), ao titular (CPF/CNPJ), à substância e ao Município combinadas é que individualizam, qualificam e identificam o título minerário. O processo de cobrança é o processo aberto após verificado alguma divergência entre o valor pago espontaneamente e o

[19] O grau em que os dados representam a realidade em relação a determinado ponto no tempo. O quanto a informação está suficientemente atualizada para a tarefa a ser realizada.
[20] O quanto a informação está disponível, ou fácil e rapidamente recuperável

valor apurado. Para cada título minerário, conforme caracterizado acima, é aberto um processo de cobrança específico, normalmente tem como digito inicial o número 9.

Dessas consultas e cruzamentos iniciais, obteve-se 4220 linhas contendo os dados básicos do título minerário, da mineradora, o valor pago espontaneamente e a diferença entre estes e o valor apurado pela fiscalização. Com os dois últimos dados é possível calcular o percentual que deixou de ser recolhido, o qual será utilizado para elaborar nosso alvo (impacto alto, médio e baixo):

$$\text{Índice de Subarrecadação} = \frac{(\text{Valor apurado} - \text{Valor pago espontaneamente})}{\text{Valor apurado}}$$

Após identificar e excluir as linhas de dados com CPF na coluna CPF/CNPJ e as que possuíam valores negativos na coluna índice de subarrecadação, restaram 3866 linhas de informação, sendo distribuídas, conforme classificação de subarrecadação após fiscalização, da seguinte forma:

- Alto: 1402;
- Médio: 1193;
- Baixo: 1271.

Algumas ideias iniciais foram abandonadas ao longo do trabalho devido ao baixo desempenho e a dificuldade de se trabalhar com valores nulos.

- Calcular o potencial de arrecadação, extrapolando os resultados das fiscalizações e características das empresas para o universo dos títulos minerários;
- Calcular o potencial de arrecadação após a classificação das empresas com base no ganho em percentual na arrecadação após a fiscalização e extrapolar o resultado do percentual médio do grupo para o todo;
- Classificar os processos em cinco grupos. Os dois a mais do que o utilizado para o trabalho seriam obtidos separando os processos que tiveram direito a restituição por terem recolhido a maior e os processos em que não houve recolhimento espontâneo.

Isto posto, o próximo passo foi rastrear possíveis fontes de informações que contivessem características vinculadas à empresa mineradora e ao título minerário em si, as quais, direta ou indiretamente, pudessem influenciar no comportamento fiscal do minerador.

Para as informações sobre a modalidade dos títulos minerários vinculados a cada um dos processos fiscalizados utilizou-se a linguagem python (por meio dos cadernos Jupyter) para baixar os arquivos CSV[21] do site dados.gov.br referentes a cada tipo (concessão de lavra, licenciamento, requerimento de lavra, disponibilidade, autorização de pesquisa, requerimento de licenciamento, requerimento de pesquisa) e consolidar a informação em uma única planilha.

Nos arquivos disponibilizados no site, são informados a superintendência em que o processo foi aberto (uma para cada Estado, exceto DF que está vinculado à Regional de Goiás e Acre, acompanhado pela de Rondônia), o número do processo minerário, tipo de requerimento do processo minerário, a fase em que se encontra o processo minerário, CPF/CNPJ e nome do titular do processo minerário, localidade (pode conter mais de um Município), substância comercializada (pode conter mais de uma substância), tipo de utilização desta e se o processo está ativo ou inativo (todos os processos divulgados estavam ativos).

Para realizar os cruzamentos de dados foi necessário eliminar os pontos e hifens dos CNPJs a fim de que as formatações dos dados baixados e do banco de dados fiquem compatíveis. As células de dados que retornaram valores nulos foram preenchidas com o texto "NaoInformado". Esse último procedimento foi realizado em todo conjunto de dados em que se mostrou necessário.

Utilizando as informações que individualizam o título minerário, citadas anteriormente, relacionou-se os processos de cobrança aos respectivos tipos de direitos minerários, conforme divulgado no site dados.gov.br. Contudo, por motivos desconhecidos, mas que podem estar vinculados, dentre outros, à mudança de titularidade, inatividade do título, etc, nem todos os processos fiscalizados estão divulgados no site. Os tipos de títulos minerários não identificados nos dados públicos foram relacionados aos processos minerários constantes nos bancos de dados fornecido pela ANM. Como havia mais de uma informação por número de processo, foram priorizados os dados com registro de datas de atualização no sistema mais recente.

[21] "CSV" significa Comma Separated Values, ou seja, um arquivo CSV é um arquivo de valores separados por vírgula. Esse formato de armazenamento é simples e agrupa informações de arquivos de texto em planilhas, usado para trocas de dados com um banco de dados ou uma planilha entre aplicativos.

Da base de dados de apoio ao cadastro mineiro, obteve-se informações sobre o uso da substância, tipo de propriedade do solo, se a área está em zona de fronteira ou em área pertencente à Amazônia legal.

A consulta referente aos dados de propriedade do solo e ao de uso da substância retornaram 2501 e 2089 linhas de campos nulo ou não informado, respectivamente. Essa quantidade representa mais de 50% dos processos de cobrança e, apesar de terem potencial para integrar o classificador, não foram utilizadas. Situações similares ocorreram com a coluna valor de saída para mercado interno, valor de saída para o mercado externo, tipo de venda (interno, externo, consumo) e tipo de fiscalização (de mesa ou in loco), e por isso não foram utilizadas no classificador.

Quanto às características/informações das empresas mineradoras, procurou-se então sistemas que possuíssem dados que pudessem conter informações que influenciassem no comportamento fiscal destas.

As seguintes informações foram extraídas das bases da ANM, base de cadastro do CNPJ, da RAIS[22] e da base de veículos do CGUdata[23]:

- CNAE (Código Nacional de Atividade Econômica) principal da empresa está ou não relacionado à atividade de extração mineral;
- A empresa possuía algum funcionário, no período de 2014 a 2018, que execute funções cujo CBO (Classificação Brasileira de Ocupação) esteja vinculado à atividade de extração mineral;
- A empresa em 2017 teve quantos funcionários exercem funções cujo CBO esteja relacionado à atividade de extração mineral;
- Quantos veículos vinculados ao seu CNPJ a empresa possuía (ano de referência foi 2016);
- Há participação estrangeira na composição do capital da empresa;
- Quantos títulos minerários a empresa possui;
- Qual o porte da empresa (microempresa, empresa de pequeno porte, outras).

[22] Relação Anual de Informações Sociais (RAIS) base de dados enviadas anualmente pelas empresas ao Governo Federal com informações detalhadas sobre os empregadores e trabalhadores formais.
[23] Criado por meio da Portaria nº 1860/2019, CGUdata é o ambiente de gestão de dados institucionais da CGU

Mesmo após as formatações para padronizar os caracteres do CNPJ em 14 dígitos (na base ANM esse dado estava como número, logo, sem os zeros da esquerda, enquanto que na base da RFB os dados estavam caracterizados como texto), não foram localizados no Labcontas 390 CNPJs, de modo que foi utilizado os dados obtidos no sistema MACROS da CGU para completar as informações, sendo que para 2 casos foi necessário pesquisar pelo nome da empresa, uma vez que o código cadastrado possuía incorreções.

Situação semelhante ocorreu quando da contagem do número de processos de títulos minerários que cada empresa possuía. Não foi encontrado nos dados baixados do dados.gov.br informações referentes a 346. Após vários cruzamentos internos ao banco de dados e utilizando o CNPJ ou o número interno de cadastro da empresa junto à ANM restaram 20 empresas sem informação. Foram preenchidos então pela quantidade de 1, a fim de evitar células de informação em branco ou retirar mais linhas de dados.

Para se obter as informações relativas às participações estrangeiras optou-se por verificar na tabela PJ_Composisao_Empresa, do banco de dados fisicajuridica da ANM, quais as empresas cujo país do capital não fosse Brasil ou desconhecido, a extrair a informação da tabela Empresa_mineração do mesmo banco de dados, porque a primeira opção tinha mais dados positivos (2345 registros de CPNJ) do que a segunda opção (1011 empresas).

Tentativas de utilizar os dados relativos à quantidade de bens minerais comercializados restaram frustradas devido à falta de padronização da unidade de medida utilizada nos diversos bancos de dados governamentais.

A falta de padronização também impediu a utilização do preço médio praticado pela mineradora como atributo do classificador, pois seria necessário dividir o faturamento da operação com a quantidade comercializada. Essa informação também poderia ser utilizada para criar uma variável categórica identificando as empresas que comercializaram muito acima ou muito abaixo da média calculada com base nas informações da ANM ou utilizando algum preço de referência oficial.

Por fim, tendo que a obrigação relativa à CFEM nasce a partir da primeira destinação do bem mineral (por venda ou beneficiamento), temos como fonte de dados auxiliar e externa

à ANM o sistema SPED[24] e SISCOMEX[25], os quais contêm dados sobre a comercialização de bens em território nacional dados de exportação. Além da comercialização dos bens minerais, poder-se-ia, de forma indireta, utilizar a comercialização de outros produtos e de maquinários necessários para a extração destas. Por exemplo, para produção de concreto existe uma proporção de areia e cimento, então a quantidade utilizada de cimento poderia ratificar, ou não, a quantidade declarada de extração de areia. Contudo, não foi possível obter os dados registrados em ambos os sistemas devido ao sigilo fiscal.

A planilha final com todas as características reunidas e após as limpezas e transformações necessárias ficou com a estrutura descrita abaixo:

Tabela 3 – Estrutura do data set final

Informação da coluna	**Tipo de variável**	**Conteúdo/Descrição**
Número de processo de cobrança	categórica	A informação foi utilizada para extrair dados e informações que farão parte do classificador, mas não será usada nele.
Número do processo original	categórica	A informação foi utilizada para extrair dados e informações que farão parte do classificador, mas não será usada nele.
Valor recolhido espontaneamente	numérica	Variável numérica contínua com valores positivos. Dado utilizado para criação do atributo alvo.
Diferença entre o valor pago e o apurado	numérica	Variável numérica contínua com valores positivos. Dado utilizado para criação do atributo alvo.
Índice de subarrecadação	numérica	Variável numérica contínua com valores entre 0 e 1. Será utilizada como atributo alvo.
Impacto	categórica	Baixo, médio, alto.
Substância	categórica	142 diferentes bens minerais, como por exemplo areia, argila, granito, calcário, água mineral, ferro, quartzito, ametista, etc)
Classe da Substância	categórica	Construção civil, industriais, metalíferas, águas minerais, combustíveis fosseis sólidos, gemas e pedras ornamentais, fertilizantes
Superintendência	categórica	SP, MG, TO, PA, GO, RO, PR, RS, MT, SE, BA, PE, ES, AM, CE, MS, RN, PI, AP, RJ, MA, AL, PB, RR
Unidade da Federação	categórica	SP, MG, TO, PA, GO, RO, PR, RS, MT, SE, BA, PE, ES, AM, CE, MS, RN, PI, AP, RJ, MA, AL, PB, DF, AC, RR
Região	categórica	N, NE, S, SE e CO
Localizada na Amazônia Legal	categórica	Informação binária. 1 para sim e 0 para não.
Localizado em área de Fronteira	categórica	Informação binária. 1 para sim e 0 para não.
Tempo do processo	numérica	Valores inteiros positivos obtidos usando como referência 2018.
CPF/CNPJ	numérica	A informação foi utilizada para extrair dados e informações que farão parte do classificador, mas não será usada nele.

[24] Sistema Público de Escrituração Digital (SPED) - tem como objetivo unificar a recepção, validação, armazenamento e autenticação de livros e documentos integrantes das escriturações contábil e fiscal das pessoas jurídicas, através de um fluxo computadorizado de informações. é composto por três projetos: Escrituração Contábil Digital, Escrituração Fiscal Digital e a NF-e - Ambiente Nacional.

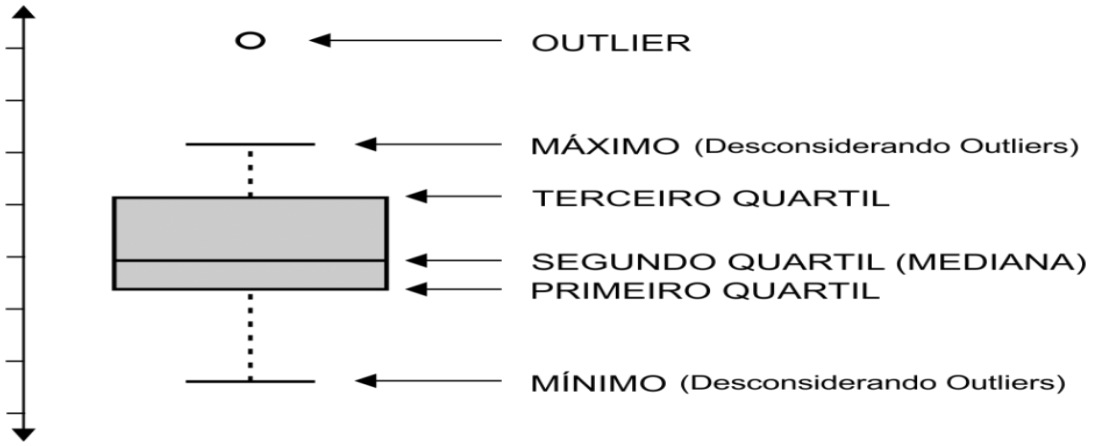
[25] O SISCOMEX (Sistema Integrado de Comércio Exterior) é um instrumento que integra as atividades de registro, acompanhamento e controle das operações de comércio exterior, através de um fluxo único, computadorizado, de informações, cujo processamento é efetuado exclusiva e obrigatoriamente pelo sistema.

Informação da coluna	Tipo de variável	Conteúdo/Descrição
CPF/CNPJ1	categórica	A informação foi utilizada para extrair dados e informações que farão parte do classificador, mas não será usada nele.
Título Minerário	categórica	Concessão de lavra, licenciamento, requerimento de lavra, disponibilidade, autorização de pesquisa, requerimento de licenciamento, requerimento de pesquisa
Nº de processos fiscalizados	numérica	Variável numérica discreta. Número inteiro positivo.
Quantidade de Títulos Minerários da empresa	numérica	Variável numérica discreta. Número inteiro positivo.
Porte da empresa	categórica	Demais, microempresa, empresa de pequeno porte
CNAE ligado ao extrativista mineral	categórica	Informação binária. 1 para sim e 0 para não.
Funcionários com CBO ligado à extração mineral	categórica	Informação binária. 1 para sim e 0 para não. Ano de referência da informação: 2014 a 2018.
Participação Estrangeira	categórica	Informação binária. 1 para sim e 0 para não.
Nº de funcionários com CBO ligado à extração mineral	numérica	Número inteiro positivo. Ano de referência 2017.
Nº de veículos.	numérica	Número inteiro positivo. Ano de referência 2016.

Finalizando a limpeza dos dados, por meio do gráfico de boxplot (diagrama de caixa), foi analisado a existência de outliers[26] que pudessem comprometer os resultados a fim de retirá-los da amostra.

O boxplot permite que se visualize a distribuição e valores discrepantes (outliers) de dados utilizando o conceito estatístico de quartil[27], conforme figura abaixo:

Figura 3: explicação gráfica do boxplot



Fonte: Oliveira, 2019.

[26] Outlier é um valor ou observação muito diferente das demais ocorrências

[27] Um quartil é qualquer um dos três valores que divide o conjunto ordenado de dados em quatro partes iguais, e assim cada parte representa 1/4 da amostra ou população.

Em relação ao atributo porte da empresa e região do Brasil onde está localizada a área do título, não foram detectados outliers.

Figura 4: Boxplot do porte da empresa

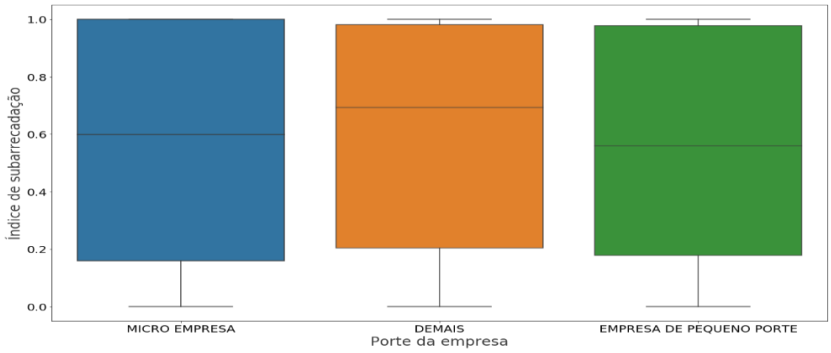
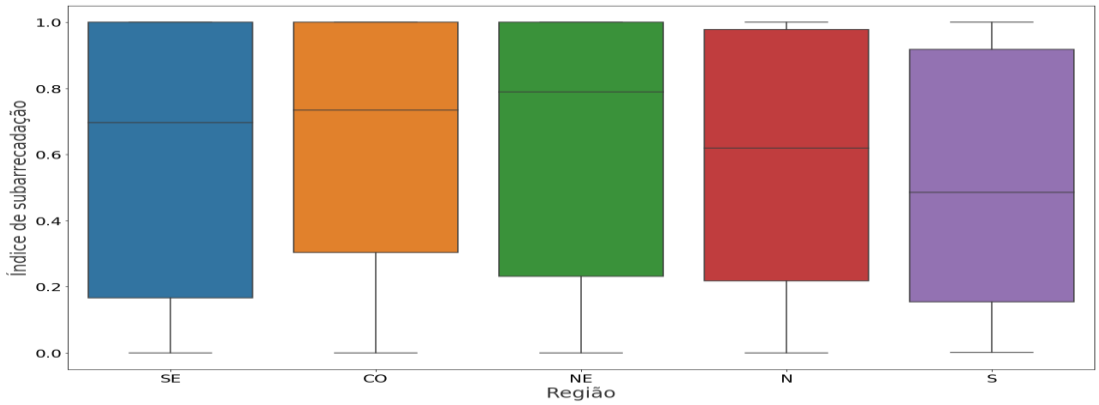
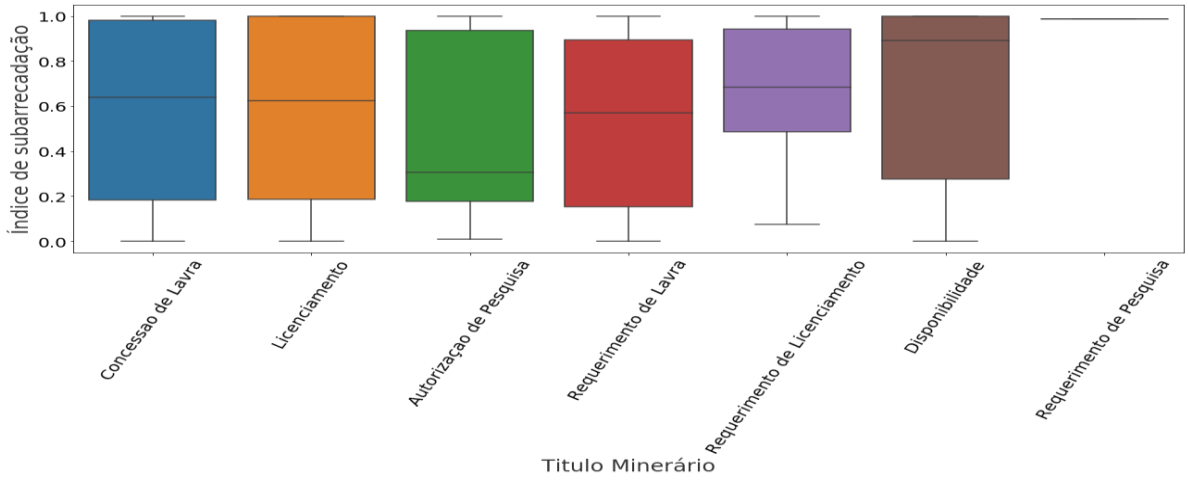


Figura:6 Boxplot relativo à região do Brasil



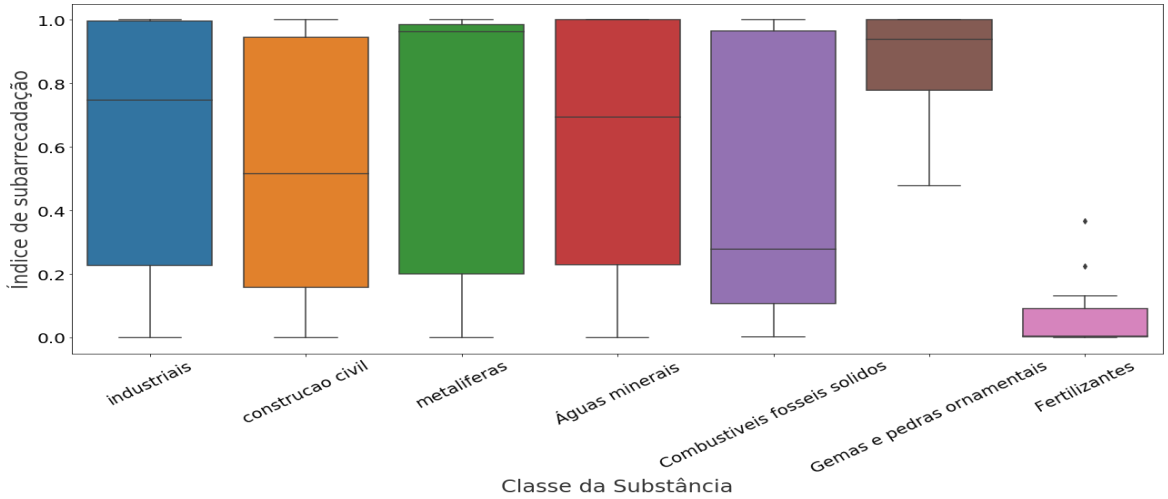
Em relação ao tipo de título minerário, não foram detectados outliers, contudo, notou-se que há apenas um processo de cobrança vinculado um título de requerimento de pesquisa.

Figura 5: Boxplot do atributo tipo de título minerário



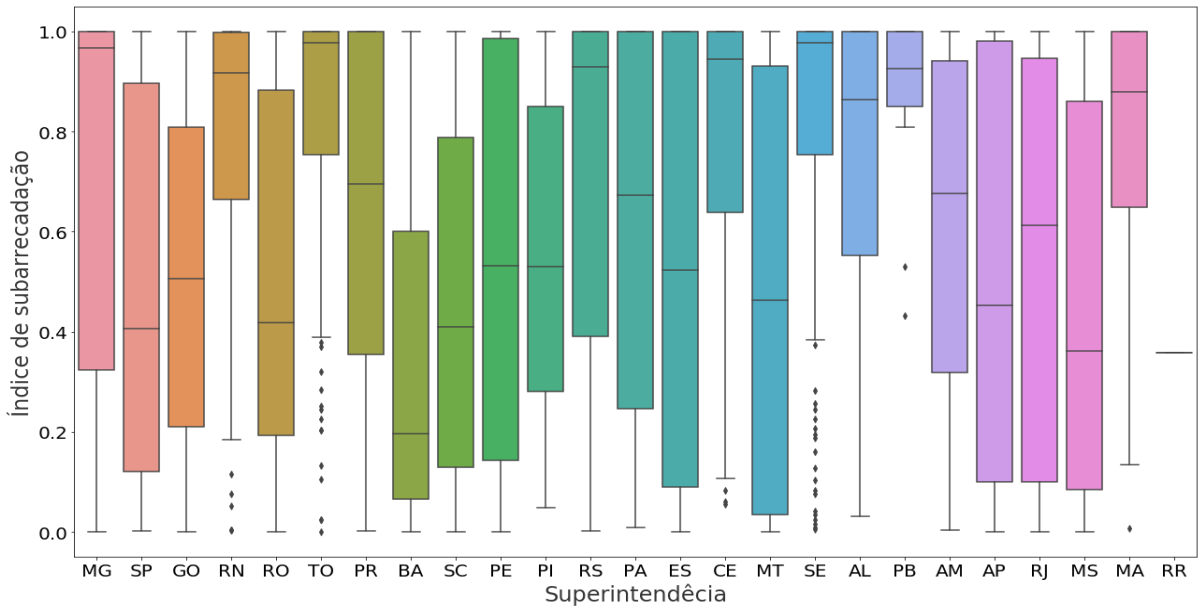
Em relação a classe de substância dos fertilizantes foi retirado apenas o processo cujo índice era maior que 0,3. O outro outlier que aparece no gráfico permaneceu, pois, ele classifica o processo na mesma categoria dos demais.

Figura 7: Boxplot da classe de substâncias.



Foram excluídas as linhas referentes ao outliers das Superintendências no Rio Grande do Norte, no Ceará, na Paraíba e no Maranhão, entretanto, foi mantido aqueles referentes aos Estados de Sergipe e de Tocantins devido à grande quantidade deles, 18 e 11, respectivamente. A decisão foi reforçada pela leve piora nos resultados do classificador quando se testou retirar os outliers desses elementos. Cabe ressaltar que, no período analisado, houve apenas uma fiscalização em Roraima.

Figura 8: Boxplot das Superintendências Regionais



Quantos aos dados sobre as substâncias, apesar da existência de outliers nos gráficos (o qual não foi representado aqui devido ao excesso de elementos, 142), optou-se por não os retirar do data set, seja pela quantidades de elementos distintos do conjunto de dados, seja pela quantidade de outliers em cada um desses elementos.

Conforme foi visto nos gráficos acima, a variável Roraima do atributo Superintendência e a variável requerimento de pesquisa, do Título Minerário, possuem apenas 1 linha de informação e, apesar de não permitirem o modelo treinar adequadamente, as linhas correspondentes não foram excluídas do data set por conterem informações úteis de outros atributos.

Para análise de outlier nas variáveis numéricas, a biblioteca matplotlib foi utilizada para criar o gráfico de dispersão (também chamados scatter), o qual mostra como uma variável afeta a outra, de modo que é possível visualizar o relacionamento e tendência dos dados. Em geral consideramos outliers os pontos distantes da parte mais densa do gráfico.

Analisando os dados não foram detectados padrões fortes do qual se pudesse inferir a existência de outliers capazes de interferir no desempenho dos modelos de classificação.

Figura 9 – Gráfico de dispersão do tempo de existência do processo minerário

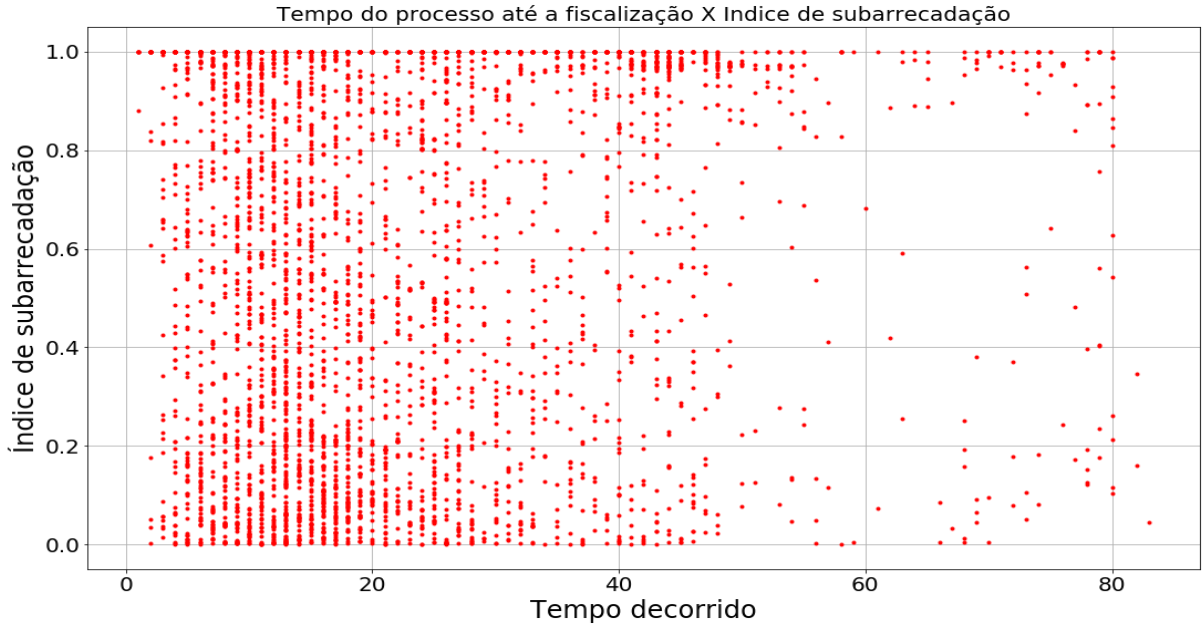


Figura 10: - Gráfico de dispersão da quantidade de fiscalizações

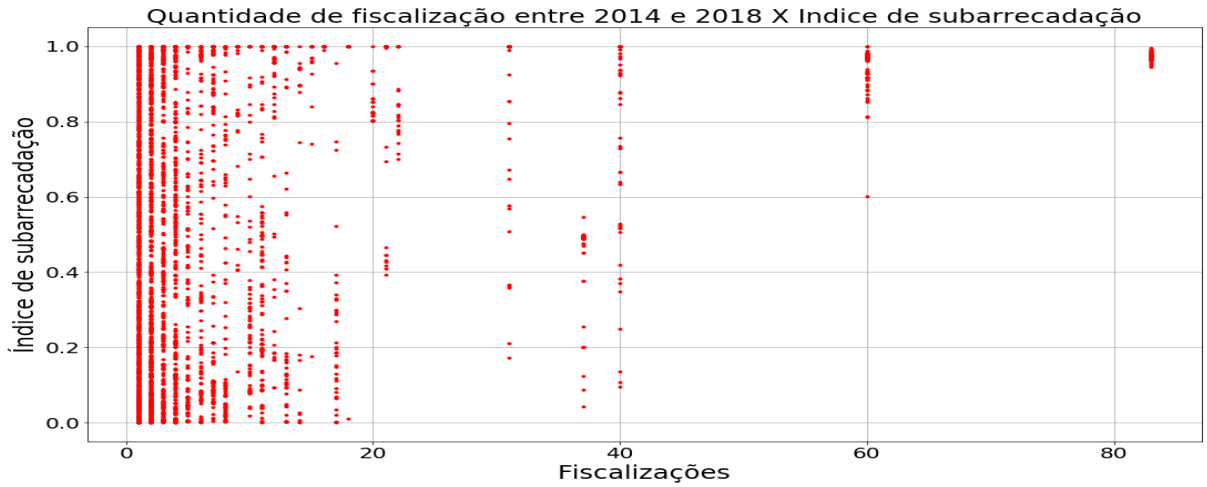
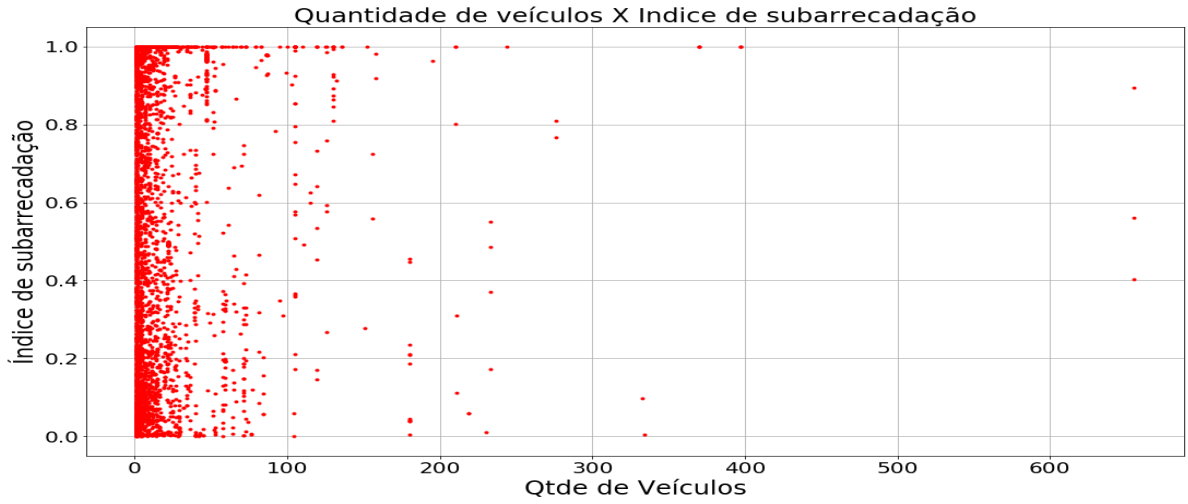


Figura 11: Gráfico de dispersão dos funcionários em atividades de extração mineral

Figura 12: Gráfico de dispersão relativo à frota de veículos da mineradora

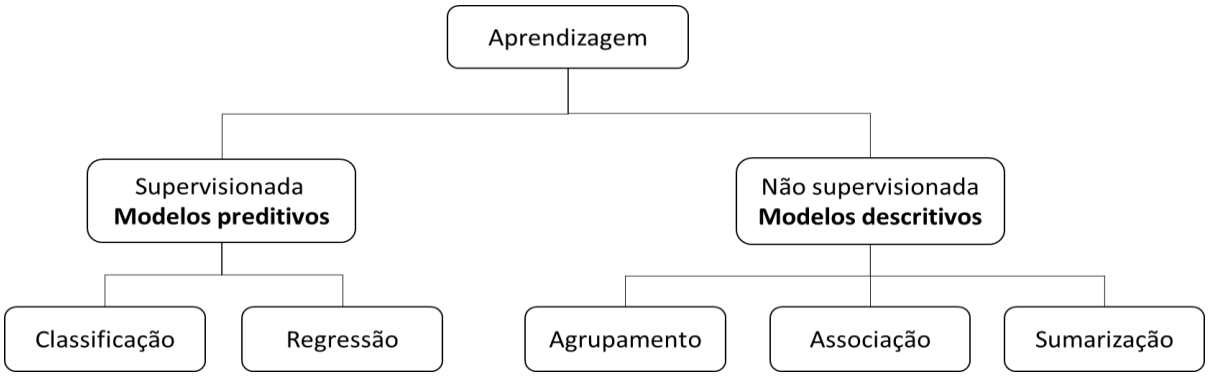


4.2 MINERAÇÃO DE DADOS

Para que dados e informações se transformem em conhecimento é necessário identificar padrões, correlações e relacionamentos entre os diferentes conteúdos de modo a torná-los úteis para tomada de decisões estratégicas. Para se alcançar esse fim, serão definidas a seguir as técnicas e algoritmos de mineração de dados que serão utilizados para alcançar o objetivo proposto.

Podemos agrupar as várias tarefas possíveis de um algoritmo de mineração de dados em atividades preditivas, utilizadas para, a partir de um histórico anterior, classificar eventos futuros ou estimar resultados desconhecidos, e descritivas, a qual identifica tendências, padrões, relacionamentos e correlação revelados pelos dados.

Figura 13: Tipos de tarefas de mineração de dados



Uma vez que a intenção é, com base nas fiscalizações já realizadas, prever o resultado de fiscalizações futuras, as atividades preditivas são as mais adequadas.

Assim estão definidas as tarefas mais comumente associadas a modelos preditivos (GOLDSCHMIDT et al, 2015):

> "• Classificação – nesta tarefa, os atributos do conjunto de dados são divididos em dois grupos. Cada atributo do primeiro tipo é denominado **atributo previsor**. O segundo tipo é denominado **atributo-alvo**. Para cada valor distinto do atributo-alvo tem-se uma classe que normalmente corresponde a um rótulo categórico pertencente a um conjunto pré-definido. A tarefa de Classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que esses registros se enquadram. (...)

> •Regressão – compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores reais. Esta tarefa é similar à tarefa de Classificação, com a diferença que o atributo-alvo assume valores numéricos."

Em que pese inicialmente ter planejado realizar uma regressão para quantificar o potencial real de arrecadação da CFEM, considerando as sonegações identificadas durante as fiscalizações, a tarefa se mostrou demasiadamente complexa. Em mais de um quarto dos processos analisados a arrecadação espontânea foi nula, de modo que não seria possível um cálculo preditivo do valor de aumento potencial na arrecadação após a fiscalização, pois não teríamos o dado referente ao valor correto apurado no momento da aplicação prática do modelo. Diante do exposto, a classificação pareceu uma técnica mais adequada ao cenário encontrado pelas equipes da ANM.

O passo seguinte seria escolher os modelos de classificação para, após testá-los, escolher o que apresentasse melhor resultado. Para esse fim, foi utilizada a biblioteca de aprendizagem de máquina de código aberto para a linguagem python Scikit-learn, a qual possui vários algoritmos de classificação, regressão, agrupamento, entre outros, sendo projetada para interagir com as bibliotecas numéricas dessa linguagem e as científicas NumPy e SciPy.

Os seguintes algoritmos[28](PEDREGOSA et al, 2011) foram escolhidos para serem utilizados nesse trabalho:

4.2.1 K-Nearest Neighbors

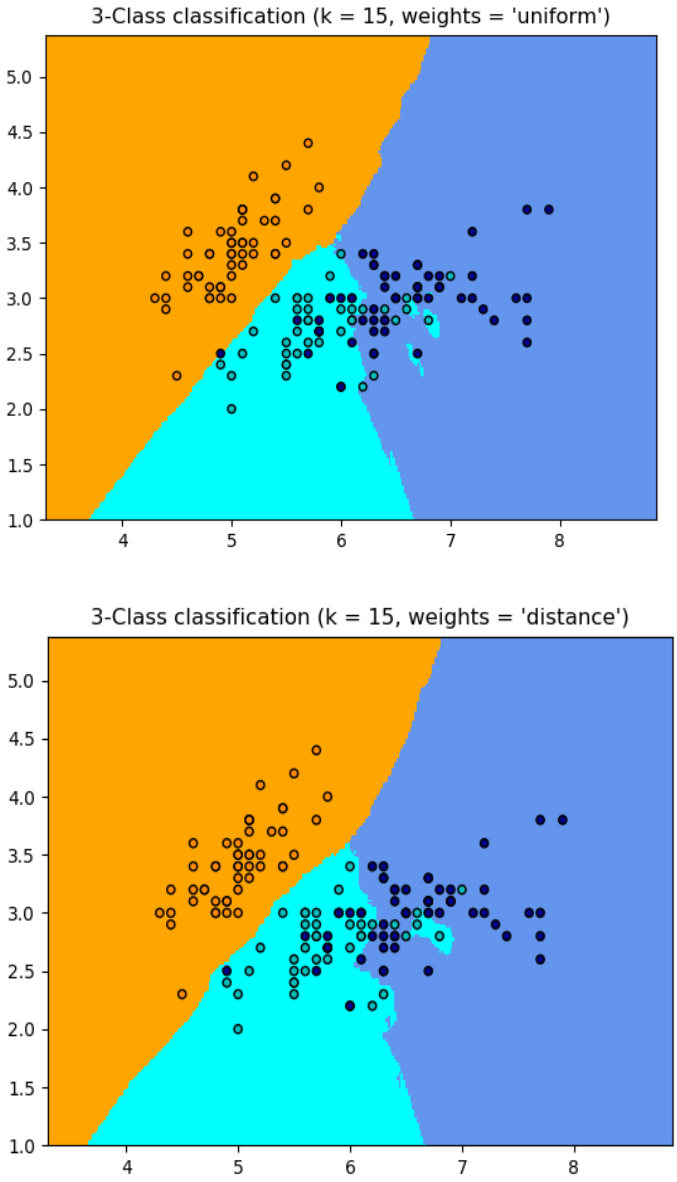
O algoritmo KNN (k-ésimo vizinho mais próximo) implementa a aprendizagem baseada nos K vizinhos mais próximos de cada ponto, onde K é um valor inteiro especificado pelo usuário. É como se a classificação funcionasse a partir de um voto majoritário simples dos vizinhos mais próximos de cada ponto: um ponto de consulta recebe a classe dados que tem mais representantes nos vizinhos mais próximos do ponto. Essa é um tipo de aprendizagem baseado em instância ou não generalista, ou seja, ele não tenta construir um modelo interno geral, mas simplesmente armazena as instâncias dos dados de treinamento.

O método mais básico usa pesos uniformes: ou seja, o valor atribuído a um ponto de consulta é calculado a partir de uma votação majoritária simples dos vizinhos mais próximos.

[28] As descrições de cada um dos algoritmos são uma tradução livre daquelas contidas no site oficial sckit-learn.org

Sob algumas circunstâncias, é melhor ponderar os vizinhos de modo que os vizinhos mais próximos contribuam mais para o ajuste. Isso pode ser realizado através da palavra-chave pesos. O valor padrão, pesos = 'uniforme', atribui pesos uniformes a cada vizinho. pesos = 'distância' atribui pesos proporcionais ao inverso da distância do ponto de consulta. Como alternativa, uma função da distância definida pelo usuário pode ser fornecida para calcular os pesos.

Figura 14: Exemplo do funcionamento do KNN

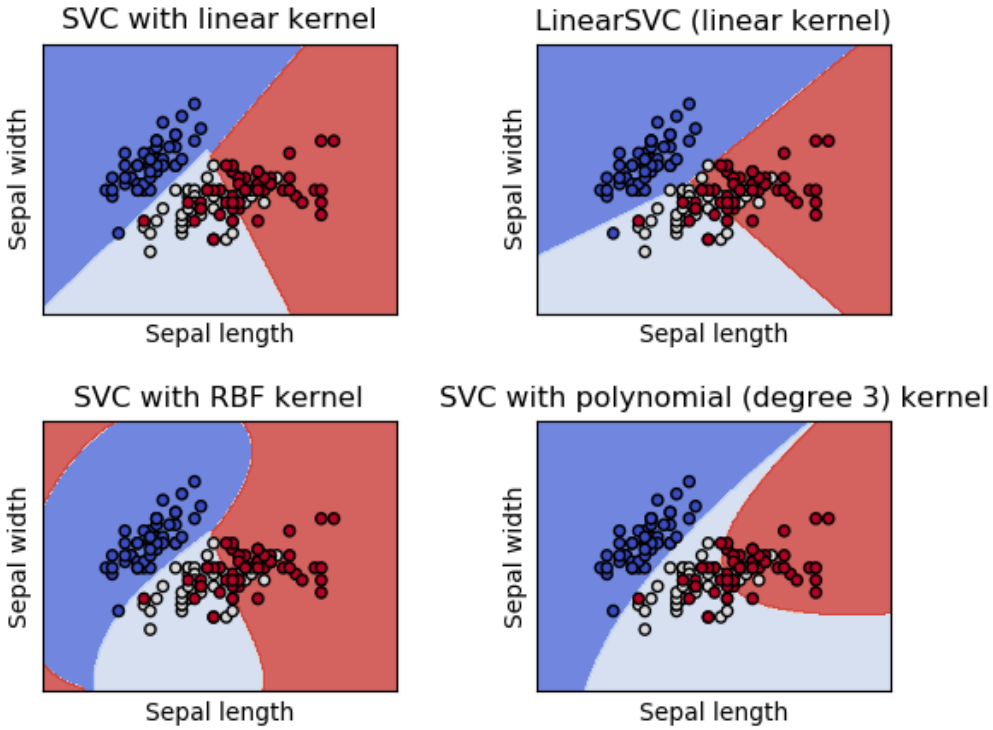


Fonte: PEDREGOSA et al, 2011

4.2.2 Support Vector Classification

O SVC é baseado na ideia de que, dados dois conjuntos de pontos linearmente separáveis no espaço, ao se encontrar um hiperplano que separe esses conjuntos, a classificação de um novo ponto torna-se trivial, pois basta verificar o sinal do resultado ao se inserir esse ponto na equação do hiperplano encontrado. A SVC é um uma classe capaz de executar a classificação multiclasse em um conjunto de dados. Ela recebe como entrada duas matrizes: uma matriz X de tamanho [n_amostras, n_atributos] contendo as amostras de treinamento e uma matriz Y de rótulos de classe (cadeias ou números inteiros), tamanho [n_amostras]. Após ser ajustado, o modelo pode ser usado para prever a classe de amostras. A função de decisão do algoritmo depende de um subconjunto dos dados de treinamento chamado vetores de suporte.

Figura 15: Exemplo do funcionamento do método SCV



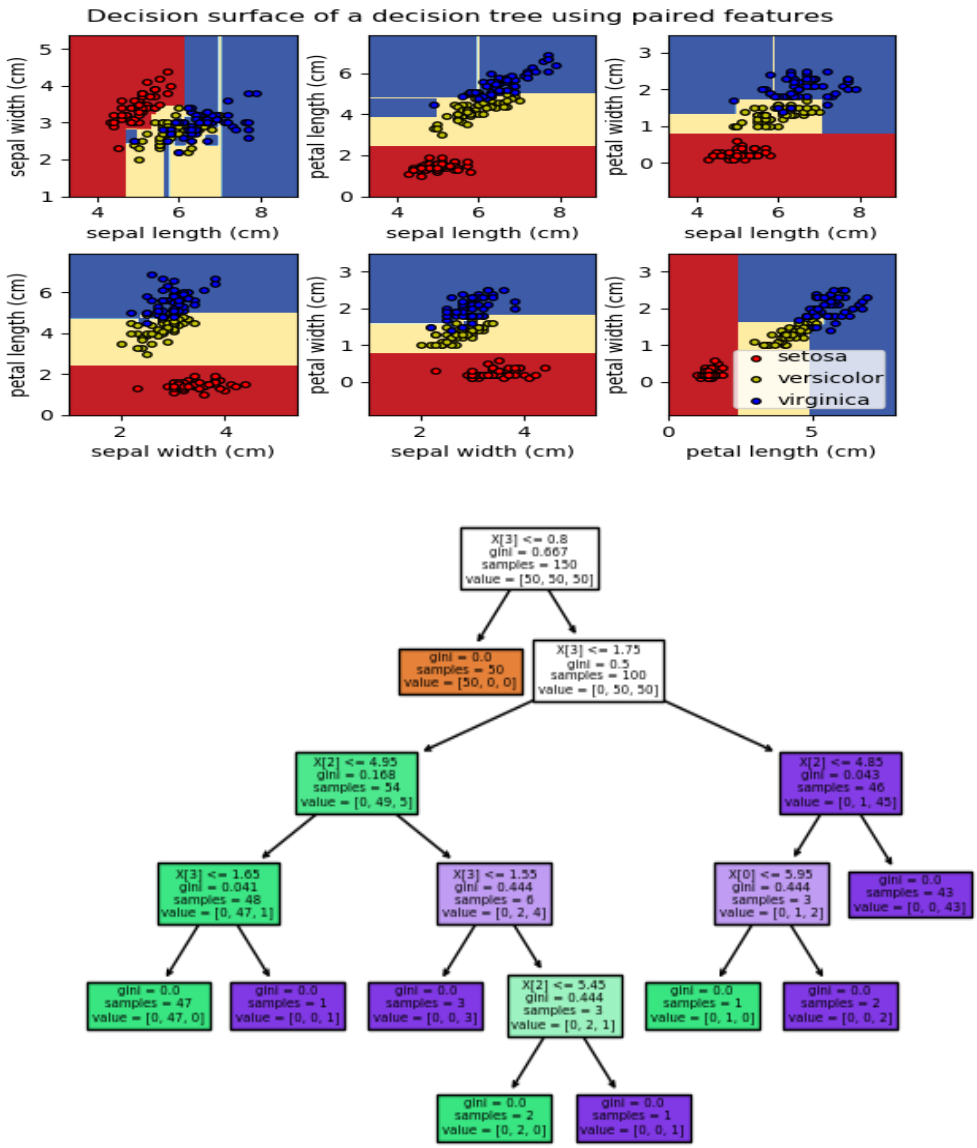
Fonte: PEDREGOSA et al, 2011

4.2.3 Árvore de decisão

Árvore de decisão: método de aprendizagem supervisionada não-paramétrico cujo objetivo é criar um modelo que preveja o valor da variável alvo por meio de aprendizagem de regras de decisão simples inferidas por meio dos atributos dos dados. A DecisionTreeClassifier é uma classe capaz de executar a classificação multiclasse em um conjunto de dados. Como em outros classificadores, ele recebe como entrada duas matrizes: uma matriz X, esparsa ou densa,

de tamanho [n_amostras, n_atributos] contendo as amostras de treinamento e uma matriz Y de valores inteiros, tamanho [n_amostras], mantendo os rótulos de classe para as amostras de treinamento. Após ser ajustado, o modelo pode ser usado para prever a classe de amostras. Em alguns casos a profundidade máxima da árvore pode gerar um sistema superespecializado nos dados de treinamento (overfititing) razão pelo qual podemos limitar a profundidade máxima usando o parâmetro max_depth ao usar esse algoritmo.

Figura 16: Exemplo do funcionamento da arvore de decisão



Fonte: PEDREGOSA et al, 2011

4.2.4 Radom Forest

Radom Forest: tem esse nome por criar uma combinação de árvores de decisão para classificar um data set. Cada árvore do conjunto é criada a partir de uma amostra projetada com reposição do conjunto de treinamento. Além disso, ao dividir cada nó durante a construção de uma árvore de decisão, a melhor divisão é encontrada em todos os atributos de entrada ou em um subconjunto aleatório de tamanho max_features.

O objetivo dessas duas fontes de aleatoriedade é diminuir a variação do estimador Random Forest. De fato, as árvores de decisão individuais geralmente exibem alta variação e tendem a se superestimar. A aleatoriedade injetada no Random Forest produz árvores de decisão com erros de previsão um pouco dissociados. Ao fazer uma média dessas previsões, alguns erros podem ser cancelados. O Random Forest alcança uma variação reduzida combinando diversas árvores, às vezes ao custo de um ligeiro aumento no viés. Na prática, a redução de variância geralmente é significativa, resultando em um modelo geral melhor.

4.2.5 Regressão Logística

Regressão Logística: apesar de seu nome, é um modelo linear de classificação, em vez de regressão. É usada para estimar valores discretos baseado em um grupo de variáveis independentes prevendo a probabilidade de ocorrência de um evento, ajustando os dados a uma função logística. Ela é uma classificação de entropia máxima (MaxEnt) ou classificador log-linear. Nesse modelo, as probabilidades que descrevem os possíveis resultados de um único estudo são modeladas usando uma função logística.

A regressão logística é implementada em LogisticRegression. Essa implementação pode ajustar a regressão logística binária, One-vs-Rest ou multinomial com opções opcionais L1, L2 ou regularização Elastic-Net.

4.3 APLICAÇÃO E RESULTADOS DO MODELO

Antes de aplicar os modelos de classificação selecionados, alguns passos têm que ser realizados. Como as funções escolhidas costumam ter como pré-requisito a utilização de variáveis numéricas, foi necessário transformar as variáveis categóricas em numéricas. Com esse objetivo foi utilizada a função get_Dummies.

Variáveis do tipo dummy são variáveis binárias (0 ou 1) criada para representar uma variável com uma ou mais categorias. O que a função get_Dummies faz é criar uma coluna para

cada categoria do atributo e atribuir valor 1 quando a variável pertencer ao atributo correspondente à coluna dele e 0 para os demais casos, criando assim, artificialmente, variáveis numéricas para dados categóricos.

Para evitar colunas com itens contendo poucas quantidades de dados, decidiu-se usar o atributo Superintendência, ao invés de Estado, pois, assim, duas colunas dummies poderiam ser suprimidas: DF (3 itens de informação) que passaria a ser a compor os dados de GO e AC integrante RO (1 item de informação).

Desse modo, para as colunas substância, classe da substância, Superintendência, região, título minerário e porte da empresa foram criadas colunas (142, 7, 25, 5, 8 e 3 respectivamente) para cada dado categórico com um valor binário associado.

Para todos os métodos é importante definir o alvo que, no caso, é o impacto/risco conforme o índice de subarrecadação encontrado em relação ao total apurado após a fiscalização, sendo os rótulos (labels) os níveis alto, médio ou baixo, conforme definido anteriormente. Os atributos seriam as características/informações em relação ao título minerário ou empresa mineradora.

Definidos o atributo alvo, é necessário dividir o restante em dois grupos: um para efetuar o treinamento e outro para teste o modelo. Foi utilizado a função train_test split da biblioteca SKlearn, sendo definido como parâmetro 15% do data set para teste, o que resultou em 3.276 linhas de dados para treino e 579 para teste, sendo 218 alto, 182 médio e 179 baixo. Somente então cada método foi testado e sua acurácia calculada.

Em relação aos atributos, foi observado que:

- Em relação à identificação do bem mineral, o atributo de classe da substância trouxe resultados ligeiramente melhores do que o uso das substâncias;
- Situação similar ocorreu em relação à localização do título minerário. Os resultados foram melhores quando se utilizou o atributo UF no lugar da informação sobre a região, de forma que se descartou a utilização dessa última.
- As informações sobre pertencer ou não a região de fronteira ou área pertencente à Amazônia Legal não fizeram diferença nos resultados dos modelos de classificação e, portanto, foram descartadas;

- Quanto aos dados vinculados à empresa, as informações sobre participação de capital estrangeiro e a declaração nas RAIS de funcionários com atividades típicas de mineração, ou o resultado foi indiferente ou piorou ligeiramente.

Desse modo, as variáveis que foram utilizadas no classificador foram:

Tabela 4: Atributos utilizados para o classificador

Atributo	**Tipo de variável**	**Conteúdo/Descrição**
Impacto	categórica	Baixo, médio, alto. (atributo alvo)
Classe da Substância	Categórica (dummy)	Construção civil, industriais, metalíferas, águas minerais, combustíveis fosseis sólidos, gemas e pedras ornamentais, fertilizantes
Superintendência	Categórica (dummy)	SP, MG, TO, PA, GO, RO, PR, RS, MT, SE, BA, PE, ES, AM, CE, MS, RN, PI, AP, RJ, MA, AL, PB, RR
Título Minerário	Categórica (dummy)	Concessão de lavra, licenciamento, requerimento de lavra, disponibilidade, lavra garimpeira, autorização de pesquisa, requerimento de licenciamento, requerimento de pesquisa
Tempo do processo	numérica	Valores inteiros positivos obtidos usando como referência 2018.
Qtde de processos fiscalizados	numérica	Variável numérica discreta. Número inteiro positivo.
Qtde de Títulos minerário	numérica	Variável numérica discreta. Número inteiro positivo.
Porte da empresa	categórica	Demais, microempresa, empresa de pequeno porte
CNAE ligado ao extrativista mineral	categórica	Informação binária. 1 para sim e 0 para não.
Qtde de funcionários com CBO ligado à extração mineral	numérica	Número inteiro positivo. Ano de referência 2017.
Qtde de veículos.	numérica	Número inteiro positivo. Ano de referência 2016.

Os resultados possíveis em um classificador são:

- Verdadeiro positivo (VP): a classe procurada foi prevista corretamente;
- Falso positivo (FP): a classe procurada não foi prevista corretamente;
- Falso verdadeiro (FV): a classe não procurada foi prevista corretamente;
- Falso negativo (FN): a classe não procurada foi prevista corretamente.

A acurácia mede a proximidade entre a classe revista pelo classificador e o valor real, isto é:

$$\text{Acurácia} = \frac{VP + FV}{VP + FP + FV + FN} = \frac{\text{Predições corretas}}{\text{todas as predições}}$$

Já a matriz de confusão é uma tabela que mostra a frequência de classificação para cada classe do modelo, onde os valores na diagonal representam as classes previstas corretamente e em todo o restante as classes que o algoritmo errou:

Tabela 5: Modelo matriz de confusão

		Classes previstas		
	Classes	**A**	**B**	**C**
Classe Real	**A**	**VP**	FP	FP
	B	FP	**VP**	FP
	C	FP	FP	**VP**

Essas duas métricas foram as ferramentas utilizadas para analisar a qualidade do resultado de cada um dos algoritmos.

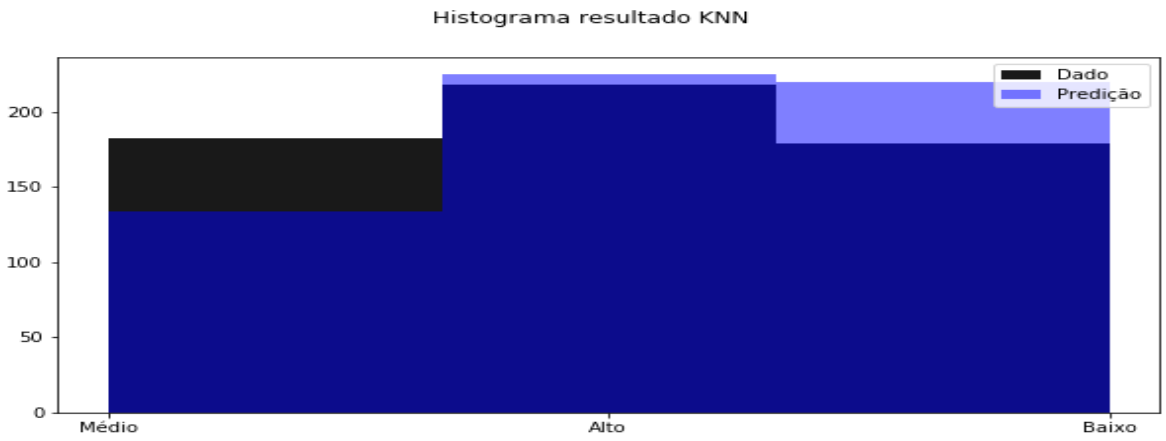
4.3.1 K-Nearest Neighbors

Como parâmetro do KNN para o cálculo dos vizinhos mais próximos foi utilizado a opção automática, a qual decide o melhor algoritmo (ball_tree, kd_tree ou bruto) baseado nos valores passados para ajustar o modelo. Outros parâmetros também permaneceram com seus valores padrões: número de vizinhos = 5, leaf_size = 30, parâmetro de força da métrica Minkoski p=2, dentre outras. Contudo, mudar o parâmetro de peso conforme a distância (inversamente proporcional à distância) proporcionou a melhor acurácia para o modelo: 0,5544.

Uma análise conjunta do histograma do resultado versus o dado de teste e da matriz de confusão nota-se que o desempenho do modelo é melhor na hora de classificar os processos com alto índice de subarrecadação. Nos de índice baixo há classificações a menos e nos baixos há processos classificados a mais.

[**142,** 36, 40]
[55, **63,** 64]
[28, 35, **116]**

Figura 17: Histograma resultado do KNN



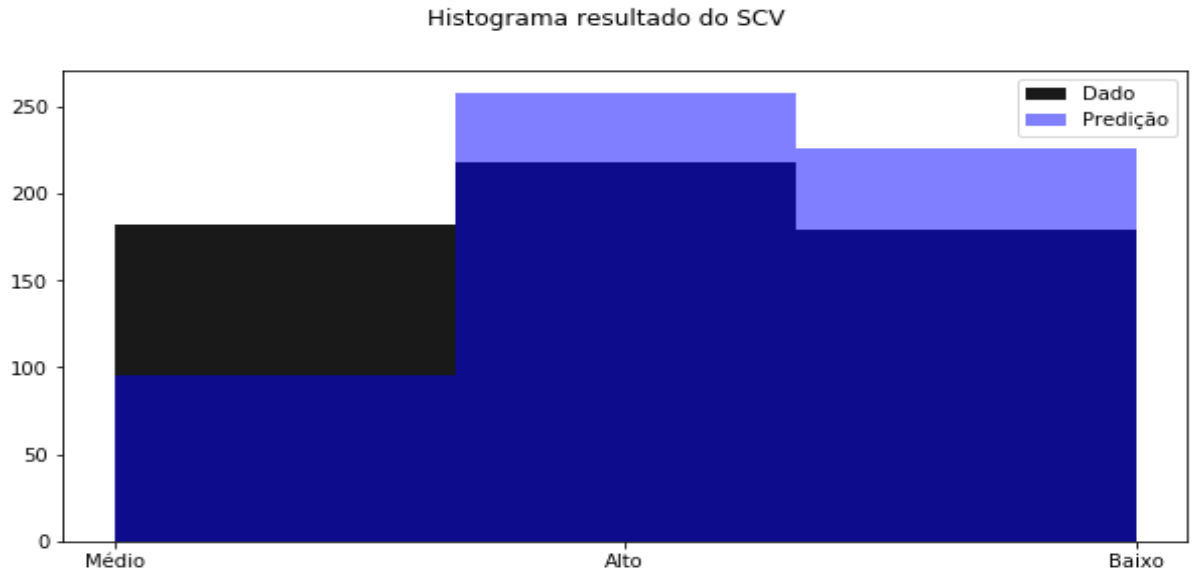
4.3.2 Support Vector Classification

A combinação de parâmetros que apresentou a melhor acurácia, 0,5595, foi a padrão do algoritmo: C=1.0, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='auto', kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False (PEDREGOSA et al, 2011).

A matriz de confusão apresentou resultado similar ao modelo KNN, contudo, qualitativamente, o resultado para índice médio foi bem abaixo apresentado quantidades superestimadas de classificações nos índices alto e baixo:

[**149,** 18, 51]
[66, **58,** 58]
[43, 19, **117]**

Figura 18: Histograma resultado do SVC



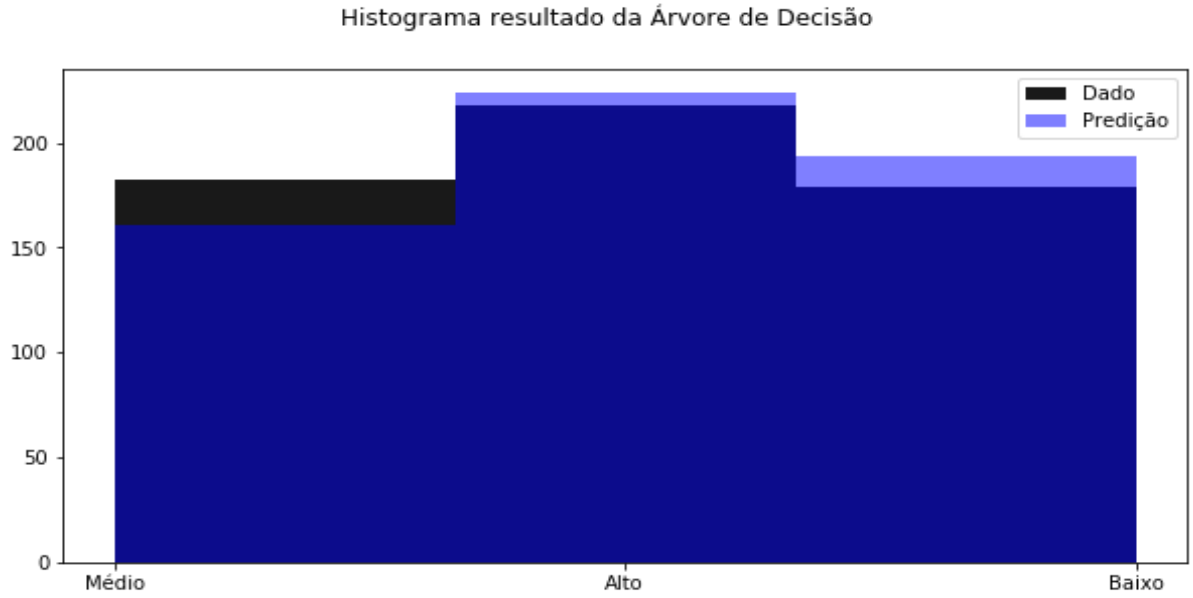
4.3.3 Árvore de decisão

A melhor acurácia, 0,5906, ocorreu com a mudança de valores padrões em dois parâmetros (PEDREGOSA et al, 2011): limitou-se as ramificações da árvore em 300 e a escolha dos números de atributos a se considerar quando se está procurando a divisão foi automatizada, max_features=’auto’.

Analisando a matriz de confusão e o histograma do resultado do classificador vê-se uma melhora nas classificações corretas quanto na distribuição quantitativa das classes, principalmente em relação ao índice médio.

[**149,** 41, 28]
[55, **77,** 50]
[20, 43, **116]**

Figura 19: Histograma resultado da Árvore de decisão



4.3.4 Random Forest

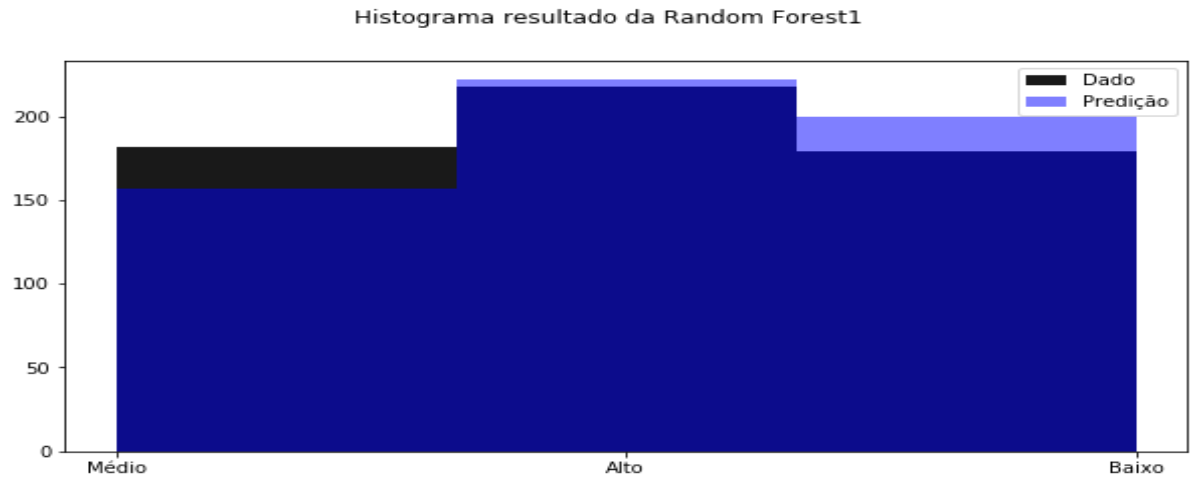
Uma vez que o algoritmo Random Forest é uma composição de Árvores de Decisão é esperado que ele apresente melhores resultados, tanto que a definição do número de árvores, parâmetro n_estimators, em 1250, fez com que o classificador apresentasse a maior acurácia, 0,6234.

Limitar a profundidade da árvore não mostrou alterações significativas de resultado, bem como, outras alterações dos valores padrões dos parâmetros do classificador (PEDREGOSA et al, 2011).

Em termos quantitativos, o histograma do resultado do classificador mostra uma distribuição similar à da Árvore de decisão, contudo, a matriz de confusão mostra maior número de acertos em todas as classes.

[**152,** 36, 30]
[53, **84,** 37]
[17, 37, **125]**

Figura 20: Histograma resultado da Random Forest



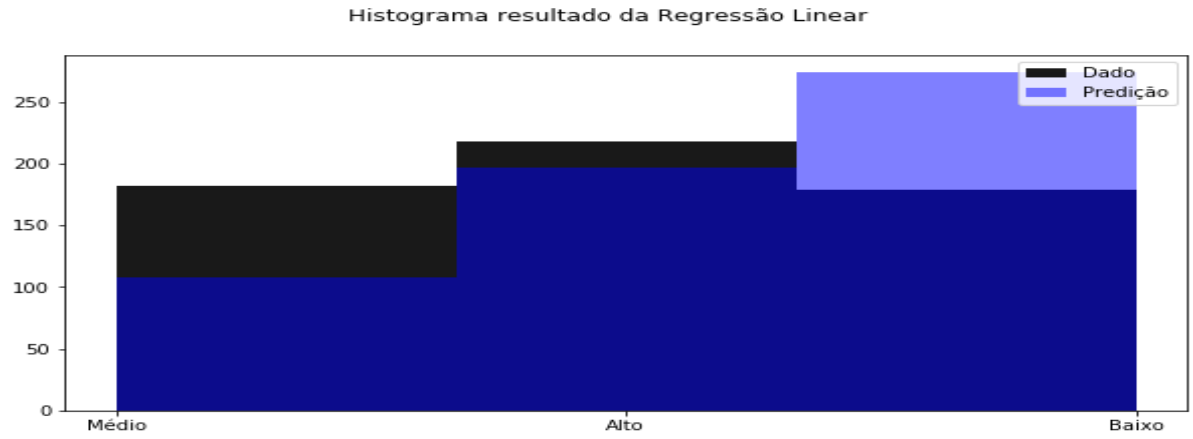
4.3.5 Regressão Logística

Dos valores padrões dos parâmetros (PEDREGOSA et al, 2011), usar o algoritmo ‘newton-cg’ na otimização do problema foi a única mudança para se alcançar a maior acurácia de 0,4905.

Esse modelo foi o que apresentou o pior desempenho em todas as métricas e análises realizadas. Ele superestima a quantidade de itens classificados como índice de sonegação baixo e subestima a quantidade dos médios. A matriz de confusão ainda mostra um menor números de processos classificados de maneira correta.

[**116,** 28, 74]
[48, **51,** 83]
[33, 29, **117]**

Figura 21: Histograma resultado da Regressão Logística



4.3.6 Comparativo dos modelos

Além das métricas já apresentadas, a precisão, que é a capacidade do classificador não rotular um dado em uma classe a qual ele não pertence, pode ajudar a compreender melhor os resultados apresentados por cada um dos modelos de classificação.

$$\text{Precisão} = \frac{VP}{VP + FP}.$$

Tabela 6: Comparação dos resultados dos modelos

Métrica		Random Forest	Árvore de Decisão	SVC	KNN	Regressão Logística
Precisão	Alto	0,68	0,67	0,58	0,63	0,58
	Médio	0,54	0,48	0,61	0,47	0,48
	Baixo	0,62	0,60	0,52	0,53	0,42
Acurácia		0,6234	0,5906	0,5595	0,5544	0,4905

De forma geral, o classificador não apresentou bom desempenho, apesar de ao longo do desenvolvimento do trabalho, inserindo novos atributos, trabalhando mais os dados e alterando parâmetros dos algoritmos, a acurácia ter subido de 0,35 para 0,6234.

Todos os classificadores apresentaram um desempenho melhor para os processos que apresentaram um alto índice de subarrecadação, provavelmente ocasionado pela maior proporção de dados pertencente a essa classe (1402, contra 1263 para médio e 1190 para baixo). Por outro lado, a classe de índice médio apresentou resultados piores, mesmo tendo mais dados para treinamento que o índice baixa subarrecadação.

Ainda que o índice de acurácia do melhor classificador (Random Forest = 0,6234) não tenha apresentado satisfatório, seu índice de acerto é melhor do que uma escolha aleatório, a qual estaria na média de 33%. Contudo, para utilização prática na escolha das empresas a serem fiscalizadas seria necessária uma melhora: seja no aumento da quantidade de dados de treinamento, seja na qualidade dos dados utilizados, seja na aplicação de técnicas mais elaborada para tratamento dos dados.

5 CONCLUSÃO

As dificuldades encontradas para acesso a algumas bases de dados e compreensão de outras podem servir de alerta e levar aos órgão envolvidos o aprimoramento do seu banco de dados, e a busca por convênios e acordos de cooperação e compartilhamento de dados que possam enriquecer o classificador, ou mesmo ser usado diretamente para ratificação dos dados fornecidos pelo mineradores.

Acredita-se também que com alguns ajustes no data set, que não foram realizados nesse trabalho, como retirada de outlier, balanceamento da amostra entre outros, o desempenho do classificador possa apresentar uma melhora significativa.

E, ainda que com acurácia baixa, o classificador pode ser usado como uma ferramenta auxiliar na definição da amostra de fiscalização da arrecadação de CFEM, uma vez que são 174.964 títulos minerários ativos, conforme dados do dados.gov.br, logo, se a proporção de mineradoras que não recolhem CFEM, conforme dados da fiscalização da ANM, se replicar no universo de títulos minerários, e o classificador conseguisse identifica-las com 100 % de certeza, ainda teríamos mais de 40 mil processos para fiscalizar. Com o atual quadro de pessoal e método de trabalho da ANM isso é quase impossível.

Diante disso e considerando que esse é um trabalho pioneiro no qual se olhou para a CFEM do ponto de vista da eficácia da arrecadação e não da sua essência fiscal ou jurídica, tendo esse trabalho como guia pode-se melhorar o classificador com inserção de novos atributos ou refinamento das técnicas aqui utilizadas e pode-se também desenvolver um modelo preditivo utilizando modelos de regressão e assim calcular o valor que está deixando de ser arrecadado pelas mineradoras.

REFERÊNCIAS

ANUÁRIO MINERAL BRASILEIRO: Principais substâncias Metálicas 2018 ano base 2017. Brasília: Agência Nacional de Mineração, 2019. 34 p.

BOLETIM DE FINANÇAS DOS ENTES SUBNACIONAIS. Brasília: Secretaria do Tesouro Nacional, ago. 2019.

BRASIL. Tribunal de Contas da União. Acordão nº 1979/2014. Plenário. Relator: Ministro Raimundo Carreiro. Sessão de 30/07/2014. **Diário Oficial da União**, Brasília, DF, 2014.

BRASIL. Tribunal de Contas da União. Acordão nº 2029/2016. Plenário. Relator: Ministro Raimundo Carreiro. Sessão de 10/08/2016. **Diário Oficial da União**, Brasília, DF, 2016.

BRASIL. Tribunal de Contas da União. Acordão nº 2604/2018. Plenário. Relator: Ministro Ana Arraes. Sessão de 14/11/2018. **Diário Oficial da União**, Brasília, DF, 2018.

BRASIL. Tribunal de Contas da União. Acordão nº 343/2019. Plenário. Relator: Ministro Aroldo Cedraz. Sessão de 22/02/2019. **Diário Oficial da União**, Brasília, DF, 2019.

CANDÃO, Jhonatan et al. A mineração de dados educacionais como apoio na análise e compreensão do processo de aprendizagem. Santiago de Chile: Nuevas Ideas em Informática Educativa, Volumen 14. 2018. 5p.

CHAPMAN, Pete et al. CRISP-DM 1.0: Step-by-step data mining guide. Disponível em https://www.the-modeling-agency.com/crisp-dm.pdf. 2000. Acesso em 11 de março de 2020.

FERNANDES, Camilla. A reforma do modelo de fiscalização do Setor Elétrico Brasileiro. Brasília: Enap, aril/2018. 30 p.

FERNANDES, Camilla et al. Fiscalização em 3 níveis – Aplicando o conceito de "Diferenciação de risco regulatório" na fiscalização de empreendimentos de geração de energia. IX Congresso Brasileiro de Regulação, Brasília, agosto de 2015. 14 p.

GOLDSCHMIDT, Ronaldo et al. **Data Mining:** Conceitos, técnicas, algoritmos, orientações e aplicações. 2ed. Rio de Janeiro: Elsevier, 2015. 276 p.

INFORME MINERAL: janeiro-junho de 2018. 1ed. Brasília: Agência Nacional de Mineração, 1º semestre/2018, dez. 2018. 16 p.

MATTIODA, Rosana et al. Qualidade da informação em duas empresas que utilizam DataWareHouse na perspectiva do consumidor de informação – um estudo de caso. Pontifícia Universidade Católica do Paraná, Curitiba, 2006.

OLIVEIRA, Bruno. Boxplot: como interpretar? Brasil, 2019.Disponível em https://operdata.com.br/blog/como-interpretar-um-boxplot/. Acesso em 22 de março de 2020.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. Regulatory Enforcement and Inspections: OECD Best Practice Principles for Regulatory Policy. Paris: OCDE Publishing, 2014.

PEDREGOSA, Fabian et al. Scikit-learn:Machine Learning in Python. EUA, 2011. Disponível em https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. Acesso em 22/03/2020.

PROVOST, Foster et al. **Data Science Para Negócios:** O que você precisa saber sobre mineração de dados e pensamento analítico de dados.1 ed. Rio de Janeiro: Alta Books, 2016. 408 p.

ZAKI, Mohammed et al. **Data Mining and Analysis**: Fundamental Concepts and Algorithms. 1 ed. New York: Cambridge University Press, 2014. 604 p.

ANEXO A - CBOs vinculados à atividade minerária

CODIGO CBO	DS_CBO_OCUPACAO	TIPO CARGO
003161	Técnicos em Geologia	Técnico Científico
316105	Técnicos em Geofísica	Técnico Científico
316110	Técnico em Geologia	Técnico Científico
316115	Técnico em Geoquímica	Técnico Científico
003163	Técnicos em Mineração	Técnico Científico
316305	Técnico de Mineração	Técnico Científico
316315	Técnico em Processamento Mineral (Exceto Petróleo)	Técnico Científico
316320	Técnico em Pesquisa Mineral	Técnico Científico
002134	Geólogos Oceanógrafos Geofísicos e Afins	Técnico Científico
213405	Geólogo	Técnico Científico
213410	Geólogo de Engenharia	Técnico Científico
213415	Geofísico	Técnico Científico
213420	Geoquímico	Técnico Científico
213425	Hidro geólogo	Técnico Científico
214515	Eng.Químico (Mineração Metalúrgica Siderúrgica Cimenteira Cerâmica)	Técnico Científico
214725	Engenheiro de Minas (Pesquisa Mineral)	Técnico Científico
007111	Trabalhadores da extração de Minerais Sólidos	Não Téc. Científico
711130	Mineiro	Não Téc. Científico
007112	Trab.de Ext. de Minerais Sólidos(Operadores de Máquina)	Não Téc. Científico
715225	Pedreiro (Mineração)	Não Téc. Científico
715520	Carpinteiro (Mineração)	Não Téc. Científico
811315	Operador de Filtro de Secagem (Mineração)	Não Téc. Científico
811325	Operador de Filtro-Esteira (Mineração)	Não Téc. Científico

ANEXO B – CNAEs vinculados à extração minerária

CNAE	Descrição SubClasse CNAE
810001	Extração de ardósia e beneficiamento associado
810006	Extração de areia, cascalho ou pedregulho e beneficiamento associado
891600	Extração de minerais para fabricação de adubo, fertilizante e outros prod.químicos
810099	Ext. e britamento de pedras e outros materiais para construção e benefic.associado
810007	Extração de argila e beneficiamento associado
893200	Extração de gemas (pedras preciosas e semipreciosas)
723501	Extração de minério de manganês
710301	Extração de minério de ferro
899199	Extração de outros minerais não-metálicos não especificados anteriormente
810003	Extração de mármore e beneficiamento associado
810002	Extração de granito e beneficiamento associado
729401	Extração de minérios de nióbio e titânio
724301	Extração de minério de metais preciosos
810009	Extração de basalto e beneficiamento associado
500301	Extração de carvão mineral
810004	Extração de calcário e dolomita e beneficiamento associado
892401	Extração de sal marinho
729404	Ext.de minérios de CU,PB,ZN e outros min. metálicos não-ferrosos não espec. ant.
725100	Extração de minerais radioativos
899102	Extração de quartzo
810008	Extração de saibro e beneficiamento associado
721901	Extração de minério de alumínio
722701	Extração de minério de estanho
810005	Extração de gesso e caulim
810010	Beneficiamento de gesso e caulim associado à extração
892402	Extração de sal-gema
899103	Extração de amianto

600003	Extração e beneficiamento de areias betuminosas
899101	Extração de grafita
729403	Extração de minério de níquel
729402	Extração de minério de tungstênio
600002	Extração e beneficiamento de xisto