Classificação dos Serviços de Saneamento para Valoração de Ativos utilizando *Random Forest*

Classification of Sanitation Services for Asset Valuation Using Random Forest

Álysson Soares[1]
orcid.org/0000-0001-7696-3290

Isabela Bulhões[1]
orcid.org/0000-0002-6171-3085

Vitor Silva[1]
orcid.org/0000-0002-6171-3085

Victor Trajano[1]
orcid.org/0000-0002-5358-747x

Alexandre M. A. Maciel[1]
orcid.org/0000-0003-4348-9291

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.
E-mail: alss@ecomp.poli.br

DOI: 10.25286/repa.v6i5.2148

Como citar este artigo pela NBR 6023/2018: SOARES, A.; BULHÕES, I; TRAJANO, V.; SILVA, V. Classificação dos Serviços de Saneamento para Valoração de Ativos utilizando Random Forest. Revista de Engenharia e Pesquisa Aplicada, Recife, v.6, n. 5, p. 90-99, Novembro, 2021.

RESUMO

No processo de revisão tarifária de uma agência reguladora a consistência dos dados é de fundamental importância para uma melhor assertividade. Para esta análise, grande parte dos dados de suma relevância não são informados, o que leva a um processo manual dos analistas responsáveis pela revisão. Visando auxiliar o trabalho, foi realizado um estudo de caso com abordagem qualitativa e quantitativa dos dados visando à extração de informações relevantes a partir de uma base disponibilizada com ativos de esgoto e de abastecimento hídrico, algoritmos de classificação baseado em Aprendizado de Máquina foram implementados e validados. Como resultado, um modelo de Random Forest capaz de classificar o tipo de serviço no qual os ativos estão inseridos foi desenvolvido, atingindo uma acurácia de aproximadamente 80%. Deste modo, o presente trabalho viabiliza predizer parte das informações faltantes nas revisões, o que diminuirá o tempo de análise dos agentes, além de reduzir os possíveis erros humanos no processo como um todo.

PALAVRAS-CHAVE: Serviços de Saneamento; Valoração de Ativos, Random Forest.

ABSTACT

In a regulatory agency's tariff review process, data consistency is of fundamental importance for better assertiveness. For this analysis, a large part of the highly relevant data is not informed, which leads to a manual process by the analysts responsible for the review. Aiming to assist the work, a case study was carried out with a qualitative and quantitative approach of the data aiming at extracting relevant information from a database made available with sewage and water supply assets, classification algorithms based on Machine Learning were implemented and validated. As a result, a Random Forest model capable of classifying the type of service in which the assets are inserted was developed, reaching an accuracy of approximately 80%. Thus, this work makes it possible to predict part of the missing information in reviews, which will reduce the agents' analysis time, in addition to reducing possible human errors in the process as a whole.

KEY-WORDS: Sanitation Services; Asset Valuation; Random Forest.

1 INTRODUÇÃO

Os sistemas de abastecimento hídrico e de tratamento de esgoto representam parcela significativa do processo de saneamento básico na sociedade. Tais sistemas são compostos por ativos tangíveis que demandam manutenção e revisão de forma periódica com os investimentos necessários.

Tendo em vista a demanda e importância de tais ativos e dos serviços a eles relacionados, existem órgãos públicos que atuam como agências reguladoras e desempenham um papel essencial no processo de gerência dos recursos disponíveis. Dessa forma, visando regular, fiscalizar e zelar pela qualidade dos serviços públicos delegados pelo estado, como por exemplo, energia elétrica, água, esgoto e gás natural canalizado.

As agências reguladoras possuem, como uma das suas responsabilidades, realizar as revisões tarifárias, onde se calcula o índice de reajuste a ser aplicado na tarifa para proporcionar ao prestador de serviço a receita anual necessária para recuperar os custos operacionais considerados eficientes e remunerar os investimentos realizados com prudência. Nessa ocasião, são verificadas todas as condições da prestação dos serviços, seus custos, receitas, remuneração de investimentos, de acordo com metodologia previamente definida por cada um dos órgãos.

1.1 DESCRIÇÃO DO PROBLEMA

A consistência dos dados é de fundamental importância para uma melhor assertividade nas revisões. No entanto, como a maior parte dos dados é encaminhada por outros agentes que não são da própria agência reguladora, informações de suma relevância não são informadas, o que leva a um processo manual e analítico dos analistas responsáveis pela revisão.

Deste modo, o processo de revisão tarifária descrito, atualmente estabelecido nas organizações para tomada de decisões é lento, pouco eficaz, executado manualmente e suscetível a falhas humanas, em decorrência da grande quantidade e variedade de dados disponíveis nas bases de ativos existentes para análise.

Sendo assim, os órgãos necessitam de uma solução refinada com a implantação de um sistema analítico mais eficiente, capaz de extrair maiores e mais relevantes informações sobre os componentes dos sistemas pertencentes às companhias reguladas. Transformando, desta forma, os dados atuais em informações importantes que agreguem mais valor ao trabalho das organizações.

1.2 OBJETIVOS

Este projeto tem como objetivo classificar, com o auxílio de técnicas e ferramentas de Mineração de Dados (MD), a qual serviço um determinado ativo está integrado, sendo ele de água ou esgoto. São objetivos específicos:

- Classificar os ativos de acordo com o tipo de serviço do qual o mesmo faz parte.
- Filtrar os dados existentes na base de dados para adquirir apenas as informações úteis para o processamento.
- Auxiliar e otimizar o trabalho dos funcionários das agências reguladoras.

1.3 JUSTIFICATIVA

Como o processo de revisão tarifária exige uma divisão primária entre ativos e suas regiões, bem como o tipo de serviço do qual está associado. Dentre a grande quantidade de dados disponível, muitos ativos não possuem essa divisão básica informada, sendo esse problema contornado com análises que demandam tempo demasiadamente.

Dado o potencial da problemática levantada e a enorme quantidade de dados disponíveis, identificou-se uma ótima oportunidade de desenvolver de forma prática os conhecimentos teóricos adquiridos ao longo da disciplina.

O lado social também foi um ponto decisivo para a escolha, tendo em vista a oportunidade de desempenhar um papel importante em um sistema que afeta diretamente a economia. Como por exemplo, impactar os mais de 9.5 milhões de residentes do estado de Pernambuco, de acordo com as projeções do Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano de 2020.

1.4 ESCOPO NEGATIVO

A ideia principal de implantar um sistema que extrai informações mais relevantes dos dados disponibilizados não contempla a identificação e/ou desenvolvimento de técnicas para a obtenção dos dados mencionados. Assim como, o objetivo final do trabalho desempenhado não visa o reajuste da equação final, mas sim auxiliar no processo de revisão.

DOI: 10.25286/repa.v6i5.2148

2 FUNDAMENTAÇÃO TEÓRICA

2.1 GESTÃO DE ATIVOS

Para Perez e Famá **[1]**, a contabilidade básica define ativos os bens e os direitos de uma entidade, expressos em moeda e à disposição da administração. Em uma ótica econômica, ativos são recursos que são controlados pela empresa e que são dotados da capacidade de gerar fluxo de caixa para os seus detentores. Para a Associação Brasileira de Manutenção de Gestão de Ativos (ABRAMAN), define ativo como algo que tenha valor real, ou potencial, para uma organização.

Segundo Pereira **[2]**, não há um consenso para definir "valor", sendo então um conceito que varia entre as organizações e seus públicos de interesse, podendo ser classificados em tangível ou intangível, financeiro ou não financeiro. Campbell *et al.* **[3]**. reforça em seu trabalho que há um grande desafio nas organizações, para determinar o impacto do ativo, e a identificação e categorização do que realmente é considerado ativo para os mesmos.

Em 2014, uma série ISO 55000 foi lançada **[4]**, definindo um padrão global que permite o gerenciamento de ativos de forma consistente e sustentável ao longo do tempo, devidos custos com manutenções, períodos de inatividade e eficiência operacional. Cantelli *et al*. **[5]**, por sua vez, define a gestão de ativos como o tratamento dado sobre os ativos físicos, utilizado para suportar a tomada de decisões, priorização de investimentos, determinação de manutenção ideal dos ativos e a frequência de renovação.

É evidente a importância que os ativos representam dentro das suas organizações, sendo, então, necessário realizar a gerência e gestão desse tipo de recurso de forma eficiente e eficaz para todos os agentes envolvidos no processo.

2.2 MINERAÇÃO DE DADOS

Segundo Fayyad **[6]**, durante os últimos anos, o aumento substancial na quantidade de dados tem mostrado a inviabilidade de tratar os dados apenas com planilhas ou com a análise humana. Portanto, é necessário encontrar ferramentas para a extração de informações relevantes dos dados é essencial, sendo essa extração, para Fayyad, chamada de *Knowledge Discovery in Databases* (KDD).

Para Zhang **[7]**, um problema de classificação ocorre quando objetos precisam ser atribuídos a um grupo ou uma classe previamente definida, com base em um número de atributos observados para o objeto em questão.

Árvores de decisão ou *Decision Tree* são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados, realizando uma partição dos seus dados de forma recursiva baseado nos valores de seus atributos.

Uma árvore de decisão é formada por três componentes: nós de decisão, os quais representam atributos que são usados para prever o resultado, nós de folha, que representam a resposta final do algoritmo e um nó raiz, o qual contém a informação com maior ganho da árvore.

Uma métrica utilizada para calcular a incerteza é a entropia. O ganho de informação é uma medida de como a incerteza na variável alvo é reduzida, dado um conjunto de variáveis independentes

As árvores de decisão têm sido aplicadas em várias áreas como, por exemplo, um sistema de admissão em universidades **[8]**, análise de risco de crédito **[9]**, análise de notas esportivas **[10]** e etc. De acordo com Patel **[11]**, uma árvore é similar ao processo de tomada de decisão humana, o que a torna mais facilmente explicável.

Random Forest ou Floresta Aleatória é um algoritmo de aprendizado supervisionado que combina árvores de decisão, de forma que cada árvore depende dos valores de um vetor aleatório amostrado independente e com igual distribuição para todas as árvores que compõem a floresta **[12]**.

A floresta aleatória emprega o método de Bagging para gerar a previsão necessária **[13]**. Esse método realiza diferentes amostragens de dados no treinamento, em vez de apenas uma amostra. As árvores de decisão individuais produzem resultados diferentes, dependendo dos dados de treinamento fornecidos ao algoritmo de floresta aleatório.

Em um processo de classificação com o algoritmo de *Random Forest*, cada nó folha de cada árvore é a saída final produzida por aquela árvore de decisão específica. O produto escolhido pela maioria das árvores de decisão torna-se o produto final do sistema. Dentre as aplicações do algoritmo de floresta aleatória, é possível citar trabalhos de classificação no setor industrial e elétrico **[14]**, no setor agrícola **[15]** e etc.

2.3 TRABALHOS RELACIONADOS

O artigo de Campos *et al*. **[16]** destaca as características de análise de dados na área de manufatura com o foco em gestão de ativos. Nele é proposto um modelo de arquitetura analítico baseado em três camadas. A primeira destaca o uso de tecnologias baseadas nos 3 V 's do Big Data. A segunda camada trataria os dados de forma analítica com técnicas de MD, usando abordagens como o CRISP-DM. E a última camada se trata de uma camada de visualização, o que facilitaria a extração de conhecimento por parte dos cientistas e engenheiros de dados.

Para Mathew *et al.* **[17]**, a indústria de serviços hídricos é um dos ramos que mais apresenta dificuldades para a gestão dos seus ativos, devido aos seus grandes estoques de diferentes tipos de ativos, sendo eles equipamentos mecânicos, elétricos, civis e etc. A partir disso, os autores realizaram uma revisão da literatura e propuseram um modelo conceitual de Data Warehouse (DW) baseado em estudos de casos.

Em Babovic *et al.* **[18]** são utilizadas técnicas de MD para determinar os riscos de rompimento de tubulações, os quais, de uma forma geral, fazem parte dos ativos da rede de água e esgoto das cidades. Com os dados obtidos em um estudo de caso, os autores elaboraram um modelo de classificação, agrupando as tubulações em classes indicando qual a probabilidade de determinada tubulação se romper baseado em parâmetros como idade da tubulação, data da última manutenção, diâmetro e etc.

Já em Li e Gao **[19]** é proposto um modelo de MD baseado na técnica multiagente para realizar a exploração de descrédito no serviço público de oferta de água. Um sistema multiagente (MAS) é um sistema composto de entidades independentes e fracamente acopladas que trabalham de forma conjunta para a tomada de decisões. Desta forma, os autores propõem um sistema de MD que inclui os agentes exploratórios, os agentes de mineração, os agentes de identificação ou avaliação e um agente de interação humano-máquina, passando por árvores de decisão, redes neurais multicamadas e protocolos de comunicação Knowledge Query and Manipulation Language (KQML).

3 MATERIAIS E MÉTODOS

3.1 STAKEHOLDERS ENVOLVIDOS

Quadro 1 – Stakeholders do projeto.

STAKEHOLDER	DESCRIÇÃO
Analistas e técnicos de regulação	Os únicos apoiadores diretos. São funcionários da ARPE que estão em constante contato com a equipe para auxiliar nas etapas do desenvolvimento do projeto.
Concessionária do serviço público	COMPESA faz parte dos apoiadores indiretos.
Poder concedente	Estado de Pernambuco
Usuários do serviço	Sociedade pernambucana que usufrui do serviço.

Fonte: Os Autores.

3.2 DESCRIÇÃO DA BASE DE DADOS

Quadro 2 – Dicionário de dados da base de ativos.

NOME	DESCRIÇÃO	TIPO	TAM.
sistema	Nome do sistema que o ativo está inserido.	String	255
município	Nome do município onde o ativo está situado.	String	255
grupo_ativo	Grupo no qual o ativo é classificado.	String	255
valor_aquisicao	Valor de compra do ativo em Reais.	Float	20
valo_depreciacao_acumulada	Valor total da depreciação acumulada do ativo.	Float	20
valor_mercado	Valor atual do ativo.	Float	20
situação_imovel	Status de operação do imóvel.	String	255
tipo	Indicativo de qual tipo de ativo (Esgoto ou água).	String	255
regiao	Região onde o ativo está localizado (RMR ou interior)	String	255
microrregião	Nome da microrregião que subdivide a região.	String	255
mesoregião	Nome da mesoregião que congrega diversos municípios.	String	255

Fonte: Os Autores.

DOI: 10.25286/repa.v6i5.2148

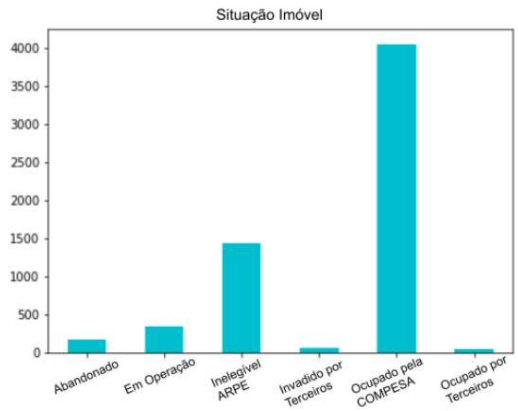
Os dados disponibilizados pela ARPE, a fim de serem utilizados no desenvolvimento deste trabalho e possibilitar a extração de informações mais relevantes, consistem em aproximadamente 63 mil linhas de registros. A base de dados representa os ativos, que são um componente essencial da equação tarifária nos processos de revisões ordinárias da COMPESA.

A base é, originalmente, constituída por 29 colunas que contém informações sobre cada sistema que compõem as redes de tratamento de esgoto e de água encanada do estado de Pernambuco. Após realizar um tratamento inicial da base, foram obtidas 11 colunas que teriam suma relevância para a análise do problema, como mostra o Quadro 2.

3.3 ANÁLISE DESCRITIVA DOS DADOS

A partir do tratamento dos dados, foi realizada uma análise gráfica das informações disponíveis. De acordo com a Figura 1, é possível observar uma maior concentração de imóveis ocupados pela COMPESA para a prestação do serviço de água e tratamento de esgoto. Esses imóveis são elegíveis e juntamente com os classificados como: Em operação, equivale a 61% dos ativos totais. Em contrapartida, os imóveis inelegíveis são classificados como: Abandonado; Inelegível Arpe; Invadido por terceiros e Ocupado por terceiros, somando um total de 1.707 terrenos, o que representa cerca de 39% do total de ativos imóveis.

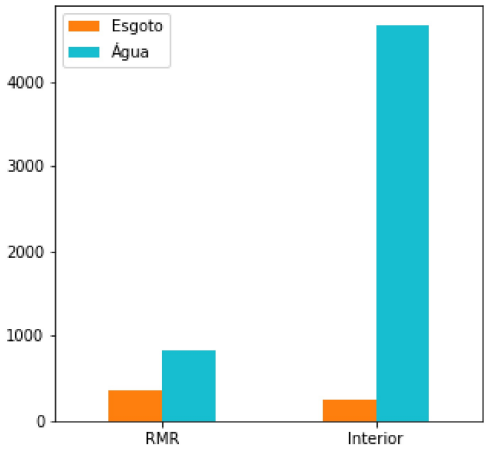
Figura 1 – Análise gráfica da situação do imóvel.



Fonte: Os Autores.

Com base na análise da Figura 2, fica notória a discrepância na alocação de recursos para a distribuição de água tratada entre a Região Metropolitana do Recife (RMR) e o interior do Estado.

Figura 2 – Distribuição dos ativos de Esgoto e Água por região do Estado.



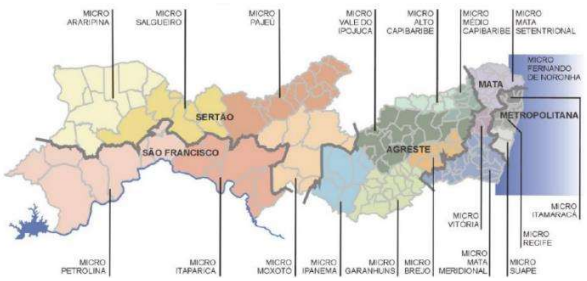
Fonte: Os Autores.

O esgoto da RMR é composto por 359 ativos, que representam 58,6% do total do sistema de esgoto do Estado. Observa-se que, para esses recursos, o quantitativo entre as regiões analisadas não difere tanto comparado aos ativos de água tratada, os quais representam apenas 15% dos ativos na RMR.

A partir da identificação da necessidade de observar os dados em um contexto regional, foram introduzidas duas novas colunas na base de ativos que facilitaram o entendimento dos valores disponíveis. A coluna de microrregião foi inserida com o objetivo de agrupar os 174 municípios presentes durante o tratamento de dados dos seus sistemas. O estado de Pernambuco é dividido em 19 microrregiões, das quais 18 foram utilizadas no mapeamento realizado, tendo em vista que Fernando de Noronha não participou da análise.

Para alcançarmos uma visualização ainda mais macro, a coluna de mesorregião foi adicionada, de forma que agrupou os municípios e microrregiões nas 5 mesorregiões presentes na divisão do estado pernambucano. A Figura 3 mostra a distribuição das micro e mesorregiões de Pernambuco.

Figura 3 – Distribuição das microrregiões e mesorregião no estado de Pernambuco.

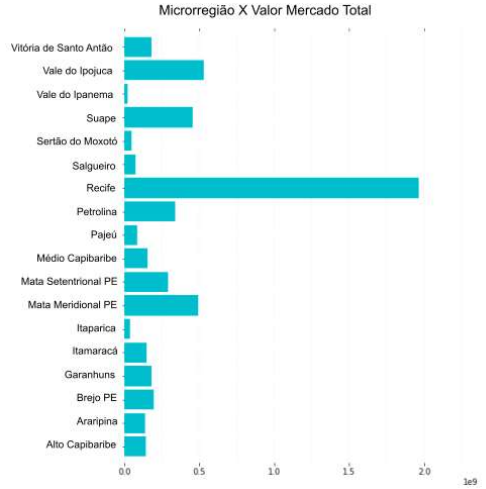


Fonte: Agência Condepe/Fidem **[20]**

A partir dos novos relacionamentos entre os municípios conhecidos e as regiões inseridas na tabela, foi possível realizar análises gráficas mais relevantes e com uma maior facilidade de se extrair as informações devido à simplicidade dos gráficos gerados.

Inicialmente foi criado um gráfico (Figura 4) onde é possível avaliar o valor de mercado dos ativos dos municípios que foram somados por sistemas contidos nas microrregiões. É notável a disparidade entre os valores da microrregião do Recife quando comparada às outras 17 analisadas. No entanto, deve-se levar em consideração que na base tratada, o município do Recife possui a maior quantidade de ativos totais somando todos os sistemas presentes na Região, o que pode comprometer a análise realizada.

Figura 4 – Análise dos valores de mercado total dos ativos por microrregiões do Estado.

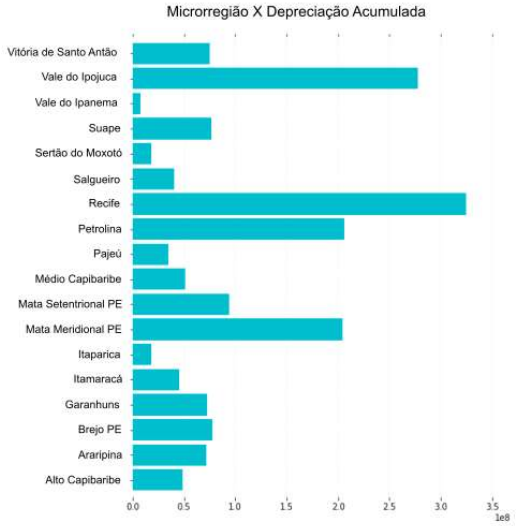


Fonte: Os Autores.

Em seguida, foi realizada uma análise acerca da depreciação acumulada sobre os ativos em questão, e para esse cálculo, leva-se em consideração o valor de aquisição do ativo subtraindo seu atual valor de mercado.

Observa-se que a microrregião do Recife ainda se destaca das demais, possuindo o maior valor em somatório da depreciação dos ativos dos sistemas contidos na região (Figura 5).

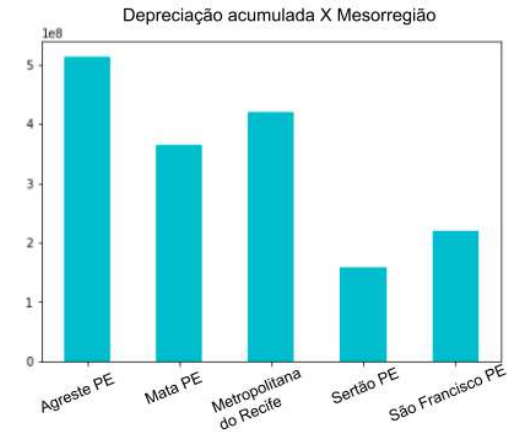
Figura 5 – Análise dos valores de depreciação acumulada dos ativos por microrregiões do Estado.



Fonte: Os Autores.

A mesma análise foi realizada a partir das mesorregiões de Pernambuco, com o objetivo principal de se obter uma visualização mais generalizada das depreciações acumuladas nas regiões, que reúnem diversos municípios de uma área geográfica com similaridades econômicas e sociais (Figura 6).

Figura 6 – Análise dos valores de depreciação acumulada dos ativos por mesorregiões do Estado.



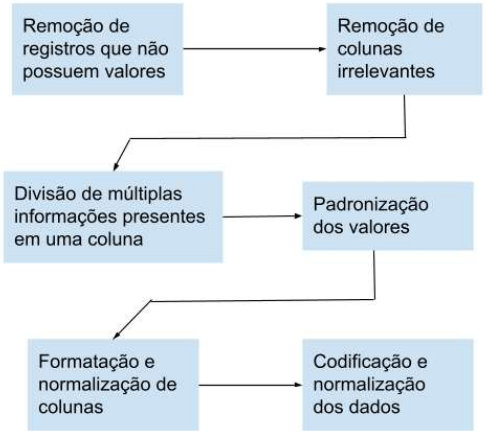
Fonte: Os Autores.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

DOI: 10.25286/repa.v6i5.2148

Durante o desenvolvimento do trabalho, uma das etapas que mais consumiu tempo foi a execução do pré-processamento dos dados originais gerados pela COMPESA e disponibilizados pela Arpe. Esse processo consistiu em um conjunto de atividades que envolveram preparação, organização e estruturação dos dados. A Figura 7 mostra o fluxograma do pré-processamento.

Figura 7 – Fluxograma do pré-processamento.



Fonte: Os Autores.

As etapas de pré-processamento dos dados consistiram em:

1 - A primeira ação realizada durante essa etapa foi a remoção de registros com valores faltantes para colunas relevantes da base. Os ativos que não possuíam valores para tais colunas não seriam importantes para realizar a classificação e valoração dos ativos.

2 - A partir do resultado obtido anteriormente, foram removidas colunas que não agregavam valor ao objetivo do trabalho, como, por exemplo, as colunas de datas, endereços dos ativos e identificadores.

3 - Ainda no processo de tratamento da base, foi realizada a divisão de uma coluna que continha duas informações relevantes, o "tipo" (esgoto ou água encanada) e a "regiao", onde o mesmo se localiza (RMR ou interior). A referida divisão gerou uma coluna a mais na base, o que resultou em um objeto de trabalho com maior granularidade e proporcionou uma análise crítica sobre os valores distribuídos nas regiões do estado de Pernambuco.

4 - Foi feita a transformação de palavras do campo "regiao" de letras maiúsculas para minúsculas, visando manter a consistência e singularidade dos registros.

5 - Colunas originalmente textuais que representavam valores (valor_aquisição, valor_depreciacao_acumulada e valor_mercado) foram formatadas e convertidas para colunas numéricas, através da substituição de vírgulas por pontos, remoção de caracteres inválidos (+ e -), substituição de valores nulos por zeros para a depreciação dos terrenos e finalmente a conversão para o tipo numérico float. Possibilitando, dessa forma, a realização de operações com seus registros e a análise gráfica dos dados.

6 – Por fim, para entrada do modelo, os valores categóricos foram codificados com a técnica de label encoding e os demais valores foram normalizados usando a técnica de mínimo-máximo, a qual segue a Equação (1).

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Onde, x_{scaled} é o valor normalizado, x é o valor atual, min(x) é o menor valor da coluna e max(x) é o maior valor presente na coluna.

Após o tratamento dos dados, foi possível notar a necessidade de observá-los através de diferentes perspectivas para um melhor entendimento dos registros disponíveis. Por conta disso, novas colunas foram adicionadas à tabela para auxiliar no relacionamento entre os valores existentes e a interpretação dos mesmos de forma mais relevante.

Portanto, a partir do pré-processamento realizado, é notável a importância da execução do tratamento dos dados antes da aplicação de modelos ou técnicas de MD.

3.3 METODOLOGIA EXPERIMENTAL

A partir do tratamento da base de dados original, foi observada a lacuna existente nos valores do tipo de serviço prestado. Foram, então, implementados algoritmos de classificação, como árvore de decisão e floresta aleatória, buscando alcançar um modelo capaz de prever a informação faltante, com uma alta acurácia, baseada nos dados das demais colunas disponíveis.

Inicialmente, se fez necessário realizar um balanceamento das classes a serem preditas para desenvolver a solução, para isso duas abordagens foram implementadas.

A primeira delas foi o *undersampling* **[21]**, que se refere a um grupo de técnicas projetadas para equilibrar a distribuição distorcida de classes para um conjunto de dados de classificação diminuindo o número de elementos da classe maioritária.

A segunda abordagem foi o SMOTE (*Synthetic Minority Oversampling Technique*) **[22]** que é um dos métodos de sobre-amostragem mais comumente usados para também resolver o problema de desequilíbrio. Seu objetivo é equilibrar a distribuição de classes, aumentando aleatoriamente os exemplos de classes minoritárias, replicando-os.

Na criação dos modelos, as *features* dadas como entradas para os algoritmos de classificação foram: região, valor de depreciação, valor de mercado, valor de aquisição, situação do imóvel e grupo ativo.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 RESULTADOS

Em relação as técnicas de balanceamento dos dados tanto para árvore de decisão quanto para a floresta aleatória, o melhor resultado obtido entre as duas abordagens foi *undersampling*, enquanto a técnica de SMOTE gerou o problema de *overfitting* em ambos os algoritmos.

Com a abordagem de *undersampling*, foram utilizadas 4196 instâncias dos dados, onde 2098 eram ativos com o tipo de serviço de água (classe 1) e os 2098 restantes, eram do tipo de serviço de esgoto (classe 0). Esses dados foram divididos em 75% para treino e 25% para testes para garantir uma melhor avaliação do modelo final.

A Tabela 1 mostra o comparativo entre os modelos avaliados para o trabalho. Como se pode observar, apesar dos valores próximos, o algoritmo *random forest* obteve melhores resultados em todas as taxas analisadas.

Tabela 1 – Comparativo dos modelos.

	Acuracy	ROC	Recall	Precision	F1
RF	79,9%	87,6%	81,5%	79,4%	80,4%
DT	77,9%	82,2%	76,6%	79,2%	77,8%

Fonte: Os Autores.

Como demonstrativo das taxas modelo construído com *random forest,* a Figura 8 mostra a curva ROC (*Receiver Operating Characteristic*).

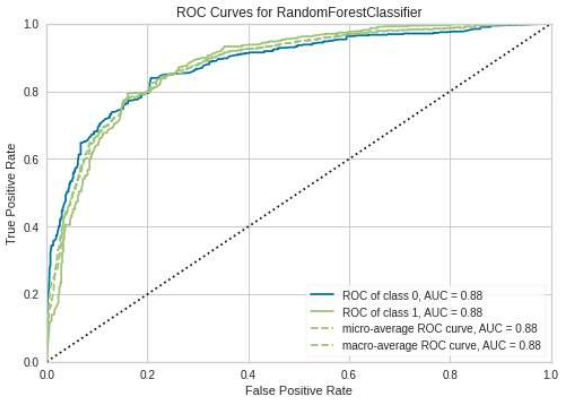
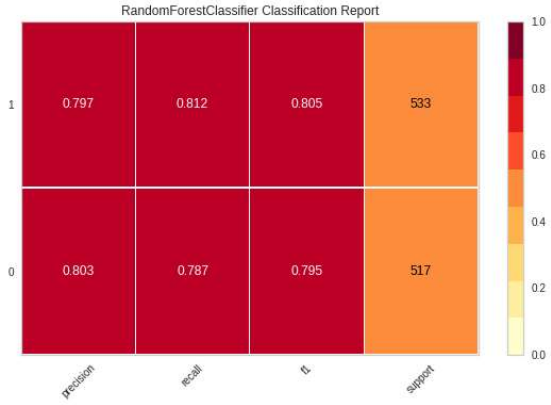


Figura 8 – Curva ROC do modelo com *Random Forest.*

Fonte: Os Autores.

Na Figura 9 tem-se uma explicitação da análise percentual por classe dos valores preditos com o modelo de melhor taxa.

Figura 9 – *Random Forest:* taxas por classe.



Fonte: Os Autores.

4.2 DISCUSSÃO

Como visto na seção anterior, o modelo baseado no algoritmo de floresta aleatória obteve melhor desempenho em todas as métricas analisadas.

Como explicitado na Figura 9, na base de testes composta por 533 ativos com tipo de serviço de água e 517 esgoto, observa-se que as taxas precision, recall e f1-score se mantêm com baixa variância em ambas as classes, o que nos mostra uma boa robustez para o modelo.

Quanto a curva ROC demostrada na Figura 8, para ambas as classes foi obtido um valor de 0.88, o que reforça ainda mais uma consistência na distinção entre as diferentes classes.

DOI: 10.25286/repa.v6i5.2148

5 CONCLUSÕES E TRABALHOS FUTUROS

O processo de revisão tarifária é uma tarefa prolongada para os analistas e um dos fatores que contribui para o aumento do tempo despendido é a falta de dados com alta relevância para a realização da análise.

Para sanar esse problema, a solução encontrada pela agência se resume em aplicar índices médios nas lacunas existentes na base de dados.

Por ser um trabalho manual e maçante, esta análise está amplamente sujeita a erro humano, podendo reduzir a precisão e influenciar negativamente nos resultados do processo de revisão tarifária.

Sendo assim, aplicar um modelo de classificação para predizer parte das informações faltantes poderá ter uma contribuição relevante para os envolvidos no processo.

Visando a continuidade do presente trabalho e consequente melhoria da base de dados original, identificou-se a possibilidade de realizar um agrupamento a partir da coluna de Sistema, de forma que seja possível unir todos os ativos que compartilham as mesmas características de endereço e tipo de serviço em sistemas reais. Desta forma, reduzindo a falta dessas informações e aumentando a precisão dos cálculos realizados pela agência reguladora.

Além disso, apesar dos resultados alcançados pelo modelo desenvolvido terem sido satisfatórios, ainda há possibilidade de melhoria em suas taxas, por meio de testes de outros algoritmos de classificação ou modelos mais robustos baseados em aprendizado profundo.

REFERÊNCIAS

[1] PEREZ, Marcelo Monteiro; FAMÁ, Rubens. Ativos intangíveis e o desempenho empresarial. **Revista de Contabilidade & Finanças**, São Paulo, v. 17, n. 40, Abr. 2006.

[2] PEREIRA, Luciana Maria Pinho. **Gestão de Ativos: Estudo de Caso em Empresa de Telecomunicações**. 2016. Dissertação em Mestrado em Engenharia de Produção, PUC-Rio.

[3] CAMPBELL, John; JARDINE, Andrew; McGLYNN, Joel. **Asset Maintenance Excellence: optimizing equipment life cycle decision**, CRC Press; 2nd ed, 2011.

[4] International Standards Organization. **ISO 55000: Asset management - Overview, principles and terminology**. ISO, 2014.

[5] CATELLI, Armando; PARISI, Cláudio; SANTOS, Edilene Santana. Gestão econômica de investimentos em ativos fixos. **Revista Contabilidade & Finanças**, v. 14, n. 31, Abr. 2003.

[6] FAYYAD, U.; SHAPIRO, G.P.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI magazine**, vol. 17, n. 3, 1996.

[7] ZHANG, G. P. Neural networks for classification: a survey. *In*: **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, 2000. vol. 30, pp. 451-462.

[8] MASHAT, A. F. *et al*. A Decision Tree Classification Model for University Admission System. *In:* **International Journal of Advanced Computer Science and Applications**, 2012. vol. 3, N. 10.

[9] KABARI, L. G; NWACHUKWU, Enoch. Credit Risk Evaluating System Using Decision Tree – Neuro Based Model. *In*: **Nigerian Journal of Tecnology**, 2009. vol. 28, no. 1.

[10] LINI, Z. Application of Research on Decision Tree Algorithm for Sports Grade Analysis. *In*: **The Open Automation and Control Systems Journa**l, 2012, vol. 7. pp. 2300-2305.

[11] PATEL, H. PRAJAPATI, P. Study and Analysis of Decision Tree Based Classification Algorithms. *In:* **International Journal of Computer Sciences and Engineering**, 2018. vol. 6, issue-10.

[12] BREIMAN, L; Random forest. In: **Machine Learning**, vol. 45 pp. 5-32, 2001.

[13] BÜHLMAN, P; Bagging, Boosting and Ensemble methods. *In*: **Handbook of Computational Statistics,** 2012.

[14] LOPES, T. D *et al*; Aplicação do Algoritmo Random Forest como Classificador de Padrões de Falhas em Rolamentos de Motores de Indução. *In*: **XIII Simpósio Brasileiro de Automação Inteligente**, ed. 1, 2017.

[15] YOU, J *et al*; Application of random forest classification to predict daily oviposition events in broiler breeders fed by precision

feeding system, *In*: **Computers and Electronics in Agriculture**, 2020, v. 175.

[16] CAMPOS, Jaime *et al*. A big data analytical architecture for the Asset Management. In: **The 9th CIRP IPSS Conference: Circular Perspectives on Product / Service-Systems**. 2017. pp. 369 – 374 .

[17] MATHEW, Avin *et al*. A Water Utility Industry Conceptual Asset Management Data Warehouse Model. *In*: **Proceeding of the 36th International Conference on Computers and Industrial Engineering**. 2006. National Tsing Hua University, CD Rom, pp. 1-11.

[18] BABOVIC, Vladan *et al*. A data mining approach to modelling of water supply assets. *In*: **Urban Water Journal**. 2002. v. 4. p. 401-414.

[19] LI, Chunsheng; GAO, Yatian. Agent-Based Pattern Mining of Discredited Activities in Public Services. *In*: **International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops).** 2006.

[20] CONDEPE/FIDEM, Agência; Planos Regionais de Inclusão Social. *In*: **Programa Governo nos Municípios, Seplan/Governo do Estado de Pernambuco**, v. 12, 2004.

[21] BACH, M.; WERNER, A.; PALT, M.; The Proposal of Undersampling Method for Learning from Imbalanced Datasets. *In*: **Procedia Computer Science**, 2019, v. 159, pp. 125-134.

[22] CHAWLA, N. V et al; SMOTE: Synthetic Minority Over-sampling Technique. *In*: **Journal of Artificial Intelligence Research** 16. 2002, pp. 321-357.

DOI: 10.25286/repa.v6i5.2148