Mineração de Dados para Obtenção do Grau de Complexidade de Processos Judiciais

Data Mining for Obtaining the Degree of Complexity of Lawsuits

Maria Gabriely L. da Silva[1]
orcid.org/0000-0002-3056-3985

Marcos Pereira da Silva[1]
orcid.org/0000-0002-9956-3422

Matheus H. Marques da Silva[1]
orcid.org/0000-0002-7964-011X

Diego Andrade Teixeira[1]
orcid.org/0000-0002-4341-3986

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

DOI: 10.25286/repa.v6i5.1755

Como citar este artigo pela NBR 6023/2018: SILVA, M. G. L.; SILVA, M. P.; SILVA, M. H. M.; TEIXEIRA, D. A. Mineração de Dados para Obtenção do Grau de Complexidade de Processos Judiciais Revista de Engenharia e Pesquisa Aplicada, Recife, v.6, n. 5, p. 56-64, Novembro, 2021.

RESUMO

A aplicação de técnicas de mineração de dados vem sendo amplamente utilizada em diversos contextos da sociedade, entre eles, o de processos jurídicos. Esse projeto tem como objetivo analisar dados jurídicos, a fim de resolver questões de complexidade dos processos judiciais e sua distribuição. O atual sistema responsável por realizar essa distribuição é o Sistema de Automação da Justiça (SAJ), em que é determinado o grau de complexidade para os processos judiciais. Porém, na maioria das vezes, a definição da complexidade não é realizada corretamente e a parte responsável pelo processo necessita alterá-la após uma revisão. Dessa forma, esse projeto consiste em aplicar técnicas de classificação e agrupamento para obtenção do grau de complexidade dos processos judiciais. Os resultados foram bastante satisfatórios para os cenários determinados, obtendo métricas de avaliação com valor superior à 73%.

PALAVRAS-CHAVE: Mineração de dados; Complexidade de Processos; Árvore de decisão; K-Means;

ABSTACT

The application of data mining techniques has been widely used in various contexts of society, among them of legal cases. This project aims to analyze legal data to solve questions regarding the complexity of lawsuits and their distribution. The current system responsible for performing this distribution of lawsuits is the System for the Automation of Justice (SAJ) which the degree of complexity for the lawsuits is determined. However, most of the time this complexity is not correct and the party responsible for the lawsuit needs to change the complexity after a review. Thus, this project consists in applying classification and grouping techniques to obtain the degree of complexity of the lawsuits. The results were quite satisfactory for the determined scenarios obtaining evaluation metrics with values higher than 73%.

KEY-WORDS: Data Mining; Process Complexity; Decision Tree; K-Means;

REVISTA DE Engenharia e Pesquisa Aplicada

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

A mineração de dados em processos jurídicos é um campo que explora algoritmos estatísticos, de aprendizado de máquina e de mineração de dados sobre os diferentes tipos de dados. Nesse projeto o principal objetivo foi analisar os dados, a fim de resolver questões de complexidade dos processos judiciais e sua distribuição **[1]**. Essa análise foi direcionada para desenvolver métodos para explorar os tipos únicos de dados dos processos jurídicos e para entender melhor a complexidade de cada processo. Promovendo, portanto, um justo balanceamento de carga de trabalho entre os procuradores da justiça.

A desconsideração do grau de complexidade dos processos na distribuição tem criado uma sobrecarga de trabalho desbalanceada. Dessa forma, a aplicação da mineração de dados proporciona um novo contexto, que possibilita a utilização desses dados para prever a complexidade de processos "semelhantes" e os distribuir de forma mais igualitária entre os procuradores.

1.2 DESCRIÇÃO DO PROBLEMA

O órgão da Procuradoria Geral do Estado (PGE) que exerce a função de fiscalizar a eficiência e a execução das atividades funcionais dos Procuradores do Estado e dos demais órgãos integrantes da Procuradoria é a Corregedoria. Cabe a esse órgão instaurar procedimentos correcionais e, juntamente com os gestores, realizar a distribuição de processos para os demais núcleos. Atualmente, a PGE conta com um quadro total de 169 procuradores distribuídos entre as especializadas Fazenda, Contencioso, Consultiva, Apoio Jurídico e Legislativo ao Governador e Regionais. Tal número vem se mostrando insuficiente para atender à grande quantidade de demandas de processos judiciais.

O sistema pelo qual é realizada a distribuição de processos é o Sistema de Automação da Justiça (SAJ). Esse sistema determina o grau de complexidade para os processos judiciais, porém na maioria das vezes essa complexidade não é definida corretamente e a parte responsável pelo processo necessita alterar essa complexidade após uma revisão. Assim, alguns núcleos acabam trabalhando com um volume maior e mais complexo de processos, enquanto outros, podem ter uma carga menor e menos complexa, em um mesmo período. Consequentemente, gera-se um desequilíbrio na carga de trabalho e insatisfação permanente daqueles que se envolvem nesses processos arrojados.

Por isso, o correto grau de complexidade dos processos se torna uma variável essencial para o balanceamento da carga de trabalho, minimizando o desequilíbrio na distribuição dos processos e o acúmulo de complexidade em determinados núcleos.

1.3 OBJETIVO

A finalidade desse artigo é analisar o conjunto de dados provenientes da PGE, utilizando a abordagem de mineração de dados, a fim de encontrar um modelo de classificação para complexidade dos processos jurídicos, e por meio desse modelo recomendar uma melhor e mais equilibrada distribuição de processos.

1.4 JUSTIFICATIVA

Atualmente, a distribuição de processos jurídicos para os procuradores do estado depende de pesos que são calculados no sistema SAJ, não levando em consideração o fator de complexidade do processo, o que acarreta um desbalanceamento da carga de trabalho. Assim, a identificação da complexidade se torna um fator fundamental para a distribuição equilibrada da carga de trabalho entre os núcleos, a fim de que esses grupos não trabalhem com processos jurídicos mais complexos e de maior volume, enquanto outros tenham uma carga menos complexa e de menor volume.

1.5 ESCOPO NEGATIVO

Não é objetivo do estudo a realização da distribuição dos processos classificados de acordo com a complexidade aos procuradores. A escolha de um determinado procurador não faz parte do escopo apresentado, apenas a classificação de acordo com a complexidade. Além disso, não é finalidade desse artigo a classificação da complexidade de acordo com um atributo específico, seja a quantidade de páginas, duração do processo, valor do processo ou complexidades específicas de movimentação ou manifestação,

DOI: 10.25286/repa.v6i5.1755

pois a complexidade final de cada processo deve levar em conta um conjunto de fatores e atributos que precisam ser analisados através da mineração de dados.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 DESCRIÇÃO DO PROBLEMA

Apesar dos avanços recentes na aplicação de técnicas de aprendizagem de máquina e processamento de linguagem natural, problemas na área jurídica possuem características próprias por contextos geográficos e linguísticos, além de trabalhos que se empenharam na busca por soluções para tornar o processo jurídico mais eficaz e ágil. O trabalho desenvolvido por Amaral **[2]** se propõe a criar uma arquitetura de rede neural artificial para predição de movimentações de processos trabalhistas na esfera jurídica. Como estudo de caso, foi utilizado um banco de dados de processos do ano de 2015 da mesma vara da esfera trabalhista, em razão do volume de dados disponíveis. O modelo proposto conseguiu obter uma predição de razoável sucesso em nível de acurácia, delimitando uma base promissora para evolução em melhores modelos preditivos e estabelecendo um mínimo de precisão que, dado o ineditismo de tal tipo de ferramenta no contexto jurídico, apresentou um protagonismo necessário para a realidade da área. O resultado também demonstrou uma capacidade do modelo de entender e interpretar parcialmente as entrelinhas dos processos trabalhistas e suas movimentações, indicando uma base sólida no que tange em abstrair as diferentes relações que os processos trabalhistas apresentam em suas movimentações. Com isso, foi criado um progresso efetivo do modelo neural para entender como a evolução histórica de um processo trabalhista influencia no seu culminar, considerando como os diferentes agentes envolvidos se fazem presentes nesse procedimento **[2]**.

Outra aplicação da técnica de mineração de dados para análise de processos jurídicos foi realizada no Estado de São Paulo pela Universidade Fatec. O objetivo desse trabalho era resolver o problema de acúmulo de processos e a duração na resolução dos casos jurídicos. Para isso, foi utilizada a técnica de mineração de dados, a fim de analisar detalhadamente os processos jurídicos do Estado de São Paulo. Além de observar que o volume de dados se tornou um fator crucial para a escolha do algoritmo a ser utilizado, conclui-se também que a área tributária possuía maior probabilidade de ter processos com longa duração, ademais, verificou-se que a Comarca de Marília tem os processos mais demorados, seguida por Bauru e Santos **[3]**.

Utilizando Inteligência Artificial, a revista do Conselho Nacional de Justiça (CNJ) publicou um artigo que permitia identificar e unificar, automaticamente, volumes significativos de demandas judiciais em tramitação que possuíam o mesmo fato e tese jurídica **[4]**. A identificação e a unificação dos processos em agrupamentos, objetiva-se em criar pendências no Sistema de Processo Eletrônico com a finalidade de informar a possibilidade de ocorrência de conexão às diferentes unidades judiciais que receberam as causas por distribuição, alertando e facilitando a análise pelo Julgador. Nesse trabalho foram aplicadas técnicas de Processamento de Linguagem Natural, Aprendizagem por Similaridade e Redes Neurais Artificiais. A solução de Inteligência Artificial (IA) construída, chamada Berna, encontra-se em produção no Poder Judiciário Goiano. A precisão de 96% nos estudos de casos demonstra a efetividade do método **[4]**.

2.2 MINERAÇÃO DE DADOS

Mineração de dados consiste em um processo analítico para explorar grande quantidade de dados cuja finalidade é encontrar padrões consistentes entre variáveis e os aplicar em novos subconjuntos de dados. Para que boas ideias sejam geradas no final da mineração, existem várias etapas que devem ser seguidas:

- **Limpeza de dados:** processo realizado para remover ruídos, incoerências, dados inconsistentes;
- **Integração de dados:** em que várias fontes de dados podem ser combinadas;
- **Seleção de dados:** dados relevantes para a tarefa de análise são recuperados da base de dados;
- **Transformação de dados:** dados são transformados ou consolidados em formulários para mineração, gerando medidas de resumo ou agregação, por exemplo;
- **Mineração de dados:** processo essencial em que são aplicadas técnicas inteligentes para extrair padrões dos dados;

- **Avaliação dos padrões:** etapa para identificar os padrões relevantes que representam o conhecimento com base em medidas de interesse;
- **Apresentação do conhecimento:** em que as técnicas de visualização e de representação do conhecimento são utilizadas para apresentar o conhecimento minado ao usuário **[8]**.

Assim, as técnicas de mineração de dados utilizam o conceito de levantamento de dados para compor grandes quantidades de dados não relacionados, localizando correlações úteis e resgatando informações valiosas. As metodologias de mineração de dados são separadas em: estatísticas, classificação, agrupamento, regressão e associação. Redes neurais, árvores de decisão, clusterização são alguns exemplos dessas metodologias, cada uma com sua vantagem, desvantagem e foco de problema **[1]**.

2.2.1 Árvore de Decisão

Árvore de Decisão é um algoritmo de classificação determinado por meio de uma série de perguntas binárias. Cada pergunta pode levar a outras perguntas ou a uma decisão. O atributo mais relevante é representado pelo nó principal da árvore, e os outros menos relevantes pelos nós subsequentes **[5]**. Uma de suas principais vantagens está na tomada de decisão levando em consideração os atributos mais relevantes da base de dados. Além disso, é um dos classificadores mais simples e de fácil compreensão que não necessita de um grande conjunto de dados para a sua geração e pode ser usado muito bem com dados categóricos **[6]**.

2.2.2 Clusterização

A Clusterização de Dados ou Análise de Agrupamentos é uma técnica de mineração de dados multivariados que por meio de métodos numéricos e, a partir somente das informações das variáveis de cada caso, tem por objetivo agrupar automaticamente por aprendizado não supervisionado os n casos da base de dados em k grupos, geralmente disjuntos denominados clusters ou agrupamentos **[7]**.

3 MATERIAIS E MÉTODOS

3.1 DESCRIÇÃO DA BASE DE DADOS

A base de dados utilizada nesse projeto é composta pela junção das seguintes tabelas do banco de dados do Sistema SAJ: **Processo, Movimentação e Manifestação**. Um processo pode ter N movimentações e uma movimentação pode ter M manifestações.

A movimentação de um processo jurídico é o resultado da tomada de ações dentro de um processo. Após uma ação ser tomada, as partes envolvidas no processo podem interpretar essa ação e tomar uma atitude em relação à movimentação lançada. Essas atitudes são chamadas de manifestações.

Atualmente o SAJ, a partir de regras internas definidas, sugere N manifestações para cada movimentação lançada. E todo processo, com suas movimentações e manifestações, é distribuído para procuradores de um determinado núcleo, para que eles possam tomar as devidas atitudes em relação ao processo. Cabendo ao procurador decidir se vai realizar uma, algumas ou todas as manifestações sugeridas pelo sistema.

Os campos da base de dados como: o número do processo e nome das partes foram criptografados por questões de privacidade dessas informações. A base também traz informações que podem ser relevantes para a definição do grau de complexidade, como o tipo da ação, assunto do processo, valor da ação, situação atual do processo, tempo do processo, tribunal e vara onde o processo está sendo tramitado. Além desses, a base também possui três campos de complexidade: complexidade do processo, da movimentação e da manifestação. O campo de complexidade dos processos é determinado pelo procurador, enquanto as complexidades das movimentações e manifestações são definidas pelo próprio sistema a partir de regras pré-determinadas.

3.2 ANÁLISE DESCRITIVAS DOS DADOS

A análise dos dados foi realizada para detectar padrões ou correlações que poderiam sugerir indícios para determinação do grau de complexidade dos processos. As variáveis utilizadas para essa etapa foram: complexidade (os três níveis) e quantidade de páginas do processo.

DOI: 10.25286/repa.v6i5.1755

As Tabela 1 mostra as medidas de resumo calculadas para a variável numérica quantidade de páginas.

Tabela 1 - Medidas de Resumo da variável quantidade de páginas.

Variável: QNTPAGINASPROCESSOTOTAL	
Tipo	Float
Ocorrência	30469.00000
Média	278.99960
Desvio Padrão	534.64850
Valor Mínimo	2.00000
Quartil 25%	83.00000
Quartil 50%	165.00000
Quartil 75%	318.00000
Valor máximo	16603.00000

Fonte: Os Autores.

Para a variável nominal complexidade, Tabela 2 e Tabela 3, as medidas de resumo descrevem o número de ocorrências, a quantidade de tipos de complexidade, o valor de complexidade mais frequente e as frequências relativas e absolutas para cada complexidade.

Tabela 2 - Medidas de Resumo da variável complexidade do processo.

Variável: COMPLEXIDADEPROC	
Ocorrência	30469
Quantidade de tipos	5
Tipo mais frequente	Média

Fonte: Os Autores.

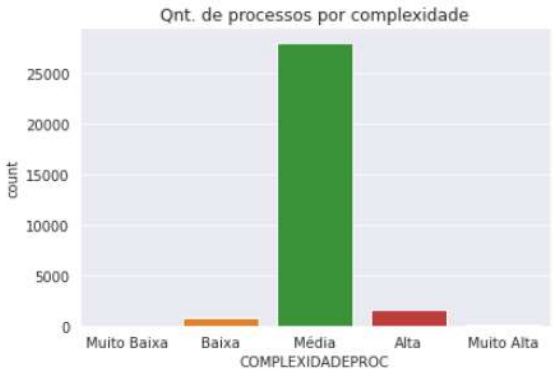
Tabela 3 - Frequência absoluta e relativa da variável complexidade do processo.

Complexidade Processo	**Frequência absoluta**	**Frequência relativa**
Muito baixa	8	0.00026
Baixa	821	0.02694
Média	27995	0.91880
Alta	1576	0.05172
Muito alta	69	0.00226

Fonte: Os Autores.

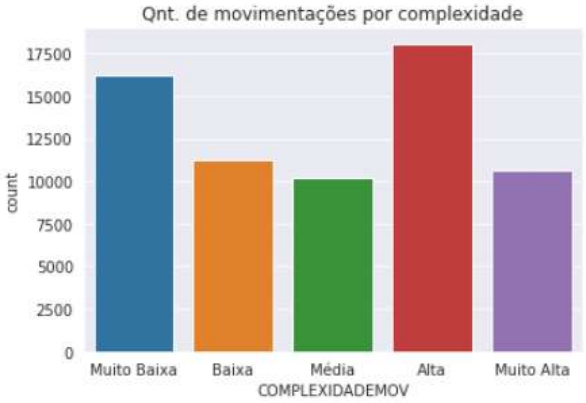
A partir dessas medidas de resumo, pode-se perceber que a maior parte da base de dados possui uma complexidade média, Figura 1. Foram analisados também os campos de complexidade da movimentação e manifestação, Figuras 2 e 3. Observa-se que em nível de movimentações, os processos são mais bem distribuídos entre os graus de complexidade. Contudo, em nível de manifestação, só existem as complexidades baixa, média e alta.

Figura 1 – Quantidade de processos por grau de complexidade



Fonte: Os Autores.

Figura 2 – Quantidade de movimentações por grau de complexidade



Fonte: Os Autores.

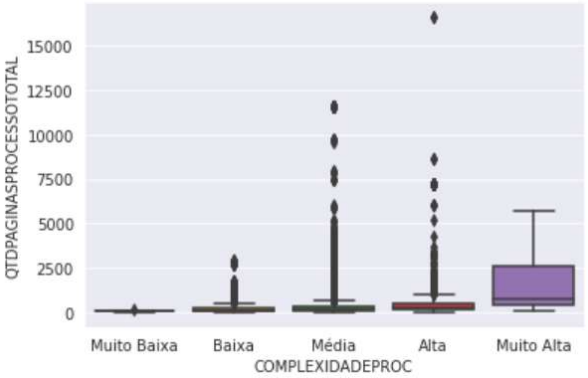
Figura 3 – Quantidade de manifestações por grau de complexidade

Fonte: Os Autores.

O *boxplot* (Figura 4) fornece uma análise visual da posição, dispersão e valores discrepantes (*outliers*) do conjunto de dados, porém existe uma dificuldade para analisar os dados, pois há alguns *outliers* muito distantes do limite superior dos *boxplots*.

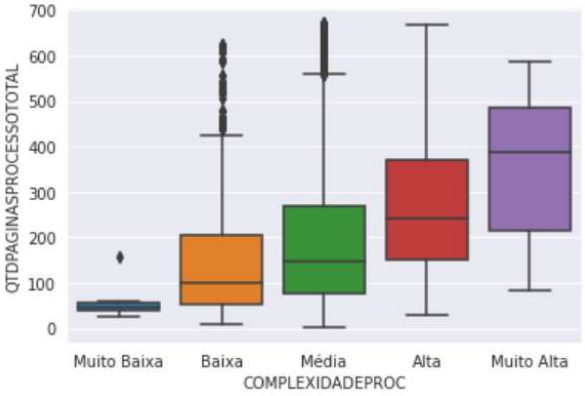
Para obter uma melhor visualização das informações fornecidas pelo *boxplot*, os dados que estão acima do limite superior do atributo "QTDPAGINASPROCESSOTOTAL" foram removidos. A partir dessa análise, foi identificada uma relação direta entre as variáveis de complexidade e quantidade de páginas, em que quanto maior a quantidade de páginas, maior a complexidade do processo (Figura 5).

Figura 4 – *Boxplot* do conjunto de dados



Fonte: Os Autores.

Figura 5 – Boxplot do conjunto de dados desconsiderando os outliers



Fonte: Os Autores.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

Dados ausentes, incompletos, inconsistentes ou ruidosos são incoerências frequentes em diversas bases de dados. O processo de limpeza dos dados auxilia na eliminação desses problemas, preparando a base de dados para a aplicação dos algoritmos de mineração.

Na base de dados desse artigo o campo numérico valor da ação possuía alguns valores nulos representados por hífen, esses valores foram substituídos pela mediana dessa coluna. A partir das análises dos gráficos *boxplot*, foi possível perceber que os campos valor da ação e quantidade de páginas possuíam bastante *outliers*, assim foi decidido pela remoção dos valores que estavam muito distantes dos limites do *boxplot*.

As colunas assunto e movimentação são compostas por classificações e subclassificações separadas por hífen. Dessa forma, foi criado mais quatro campos, dois que receberam as classificações das colunas assunto e movimentação e os outros dois com as informações das subclassificações dessas mesmas colunas.

Por fim, para os campos nominais classificação, subclassificação e complexidade do processo foi realizada uma codificação, atribuindo valores numéricos para esses campos. Já para os campos numéricos valor da ação e quantidade de páginas foi realizado a normalização dos dados.

3.4 METODOLOGIA EXPERIMENTAL

3.4.1 Classificação

Para realizar a classificação de processos jurídicos em relação a sua complexidade, foi aplicada à técnica de aprendizado supervisionado árvore de decisão. Foi definido realizar a classificação dos processos em três níveis e, não, em cinco, sendo eles complexidade baixa, média e alta. A base de dados pré-processada foi dividida em 75% para treinamento e 25% para teste e, para um melhor balanceamento dos dados, foi realizado o *under-sampling*. Os atributos de entrada da árvore de decisão foram os campos de classificação, subclassificação e valor da ação. As métricas de avaliação calculadas foram: matriz de confusão, acurácia, precisão e *recall*. Para a escolha dos parâmetros da árvore de decisão, foi utilizada a técnica *Grid Search*, variando o parâmetro de critério de divisão entre o índice Gini e entropia e o parâmetro que define o número de profundidade máxima.

DOI: 10.25286/repa.v6i5.1755

3.4.2 Agrupamento

Para realizar o agrupamento de processos jurídicos em relação a sua complexidade, foi aplicada a técnica de clusterização K-means. Adicionou-se à base de dados os atributos quantidade de movimentações por processo jurídico, média e mediana das complexidades das movimentações por processo. Esses atributos foram utilizados como entrada do algoritmo de agrupamento, além dos atributos valor da ação, quantidade de páginas e classificações dos processos. As métricas de avaliação calculadas foram *silhouette score* e *davies bouldin score*. O número de clusters definido foi igual a três, pois o objetivo é encontrar três níveis de complexidade (baixa, média e alta).

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 RESULTADOS - CLASSIFICAÇÃO

A base de dados possui um considerável desbalanceamento em relação à complexidade do processo, em que 92% dos dados possuem complexidade média, Tabela 4.

Tabela 4 - Frequências absoluta e relativa das complexidades baixa, média e alta

Complexidade Processo	Frequência absoluta	Frequência relativa
2.0	23165	0.92519
1.0	649	0.02592
3.0	1224	0.04888

Fonte: Os Autores.

Dessa forma, para obter um melhor balanceamento dos dados foi realizado o *under-sampling*, Tabela 5.

Tabela 5 - Frequências absoluta e relativa das complexidades baixa, média e alta após o under-sampling

Complexidade Processo	Frequência absoluta	Frequência relativa
2.0	2000	0.5163
1.0	649	0.16757
3.0	1224	0.31603

Fonte: Os Autores.

Realizado o balanceamento, escolheu-se os atributos de entrada da árvore de decisão, sendo eles a classificação, a subclassificação e o valor da ação. Em relação aos parâmetros da árvore, os parâmetros *criterion* (medida de qualidade da divisão) e *max_depth* (profundidade máxima da árvore) foram variados. No parâmetro *criterion* foram utilizadas as métricas *Gini* e *Entropy,* já o parâmetro *max_depth* foi variado entre 6 e 13 incluindo o valor None. A Tabela 6 apresenta o resultado das métricas de avaliação acurácia, precisão e *recall* na utilização da árvore de decisão variando os parâmetros citados.

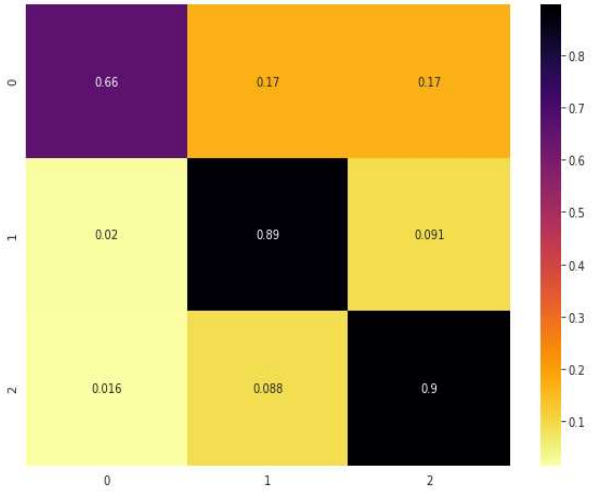
Tabela 6 - Resultado das métricas: acurácia, precisão e recall

param_criterion	*param_max_depth*	*Accuracy*	*Precision*	*Recall*
gini	6	0.7480	0.7614	0.7017
gini	7	0.7647	**0.7706**	0.7131
gini	8	0.7591	0.7665	0.7061
gini	9	0.7598	0.7647	0.7070
gini	10	0.7678	0.7670	0.7183
gini	11	0.7738	0.7586	0.7329
gini	12	0.7753	0.7577	0. 7349
gini	13	0.7717	0.7508	0.7271
gini	None	0.7570	0.7293	0.7224
entropy	6	0.7090	0.7286	0.6869
entropy	7	0.7410	0.7419	0.7091
entropy	8	0.7699	0.7586	0.7227
entropy	9	0.7707	0.7604	0.7263
entropy	10	0.7699	0.7554	0.7201
entropy	11	0.7761	0.7658	0.7242
entropy	12	0.7740	0.7582	0.7295
entropy	13	**0.7818**	0.7699	**0.7361**
entropy	None	0.7593	0.7330	0.7244

Fonte: Os Autores.

Após encontrar o modelo com os melhores resultados, foi realizada a execução no *dataset* de teste, obtendo a matriz de confusão, Figura 6. Verificou-se que processos de média e alta complexidade tiveram um alto valor de acurácia, essas são as complexidades com maior frequência no *dataset*, já a baixa complexidade teve um valor de acurácia inferior devido a menor sua frequência.

Figura 6 – Matriz de confusão da árvore de decisão



Fonte: Os Autores.

4.2 RESULTADOS - AGRUPAMENTO

Na execução do K-means os seguintes parâmetros foram variados:

- **init:** método de inicialização, utilizando k-means++ em que seleciona os centróides para o cluster de uma forma inteligente com a finalidade de acelerar a convergência e random em que a escolha é aleatória para os centróides iniciais;
- **Algorithm*:*** algoritmo k-means a ser utilizado (auto, full).

Como métricas de avaliação foram calculados a *silhouette score* e *davies bouldin score* para cada cenário definido, obtendo a seguinte tabela de resultados.

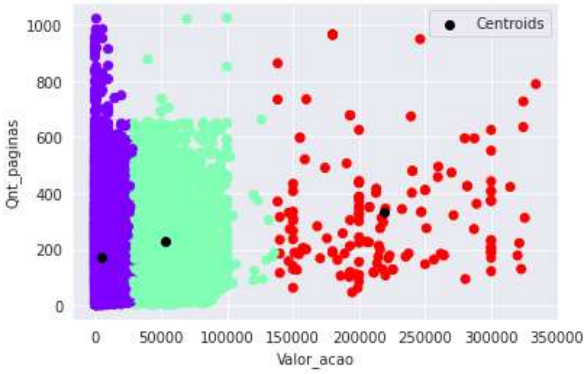
Tabela 7 - Cenários de execução do k-means.

cenario	cluster	init	algorithm	silhouette_score	davies_bouldin_score
1	3	random	auto	0.5681	**0.5167**
2	3	random	full	0.5681	0.5167
3	3	k_means	auto	**0.8078**	0.3836
4	3	k_means	full	0.80785	0.3616

Fonte: Os Autores.

O cenário 3 obteve um melhor resultado da *silhouette*. A Figura 7 mostra a distribuição dos *clusters* encontrados nesse cenário e os centróides são representados pelos pontos pretos.

Figura 7 – Resultado do k-means com 3 clusters



Fonte: Os Autores.

Também foi realizada uma análise aplicando o algoritmo de redução de dimensionalidade TSNE, antes da execução do K-means. As métricas de avaliação calculadas foram novamente a *silhouette score* e *davies bouldin*. A Tabela 8 traz os resultados das métricas e é possível observar que a métrica *davies bouldin* obteve um resultado muito superior ao resultado do cenário sem o TSNE, já a *silhouette score* obteve resultados inferiores.

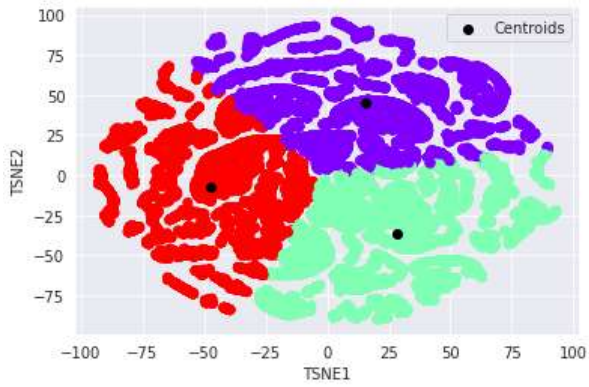
Tabela 8 - Cenários de execução do k-means + TSNE

cenario	cluster	init	algorithm	silhouette_score	davies_bouldin_score
1	3	k_means	auto	0.3645	0.8775
2	3	k_means	full	**0.3645**	**0.8759**

Fonte: Os Autores.

A Figura 8 mostra a distribuição dos clusters com a redução de dimensionalidade para duas dimensões.

Figura 8 – Resultado do k-means+TSNE com 3 *clusters*



Fonte: Os Autores.

DOI: 10.25286/repa.v6i5.1755

5 CONCLUSÕES E TRABALHOS FUTUROS

Esse artigo discute sobre o uso de técnicas de classificação e agrupamento para a resolução do problema de distribuição de processos judiciais mediante a sua complexidade, enfrentado pela Procuradoria Geral do Estado de Pernambuco.

No escopo do trabalho, foi proposto o uso de árvores de decisão para a classificação da complexidade dos processos, obtendo resultados satisfatórios de 78.18% de acurácia média e 73.61% de *recall*. Indicando, portanto, que o uso desse algoritmo no contexto de processos jurídicos pode automatizar o processo de classificação dos processos por meio de sua complexidade, diminuindo a probabilidade de erros por meio do operador do sistema e melhorando a distribuição dos processos entre os procuradores.

Ademais, também foi feito o uso do algoritmo k-means para o agrupamento dos processos, visando definir grupos de prioridades por meio das características dos processos. Com os agrupamentos realizados, é possível afirmar que o uso desta técnica se mostrou eficiente conseguindo separar bem a complexidade em três grupos distintos, obtendo uma *silhouette score* de 80% para um cenário sem redução de dimensionalidade e *davies bouldin* de 87% para um cenário com redução de dimensionalidade.

Como trabalho futuro, visamos o aprimoramento do algoritmo de classificação dos processos e o uso do classificador em conjunto com o k-means, para gerar uma distribuição dos processos baseada nas características quantitativas.

REFERÊNCIAS

[1] RAMPÃO, T. S. **Mineração de dados em bases jurídicas: um estudo de caso.** TCC (Bacharelado em Gestão da Informação) - Universidade Federal do Paraná. Curitiba, p. 159. 2016.

[3] DE CASTRO JÚNIOR, Antônio Pires; CALIXTO, Wesley Pacheco; DE CASTRO, Cláudio Henrique Araujo. **Aplicação da Inteligência Artificial na identificação de conexões pelo fato e tese jurídica nas petições iniciais e integração com o Sistema de Processo Eletrônico**. CNJ, p. 9.

[2] AMARAL, Ayrton Denner da Silva. **Predição do tempo de duração de processos e de movimentações processuais na esfera trabalhista**. 2019. 66 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Goiás, Goiânia, 2019.

[4] CUNHA, João FT; SILVA, Wellington F.; TALON, Anderson F. **Aplicação da Técnica de Mineração de Dados na Análise de Processos Jurídicos do Estado de São Paulo**. Caderno de Estudos Tecnológicos, v. 1, n. 1, 2013.

[5] KAMIŃSKI, Bogumił; JAKUBCZYK, Michał; SZUFEL, Przemysław. **A framework for sensitivity analysis of decision trees**. Central European journal of operations research, [S. l.], v. 26, p. 135–159, 24 maio 2017. DOI 10.1007/s10100-017-0479-6. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5767274/. Acesso em: 15 abr. 2021.

[6] HO, Tin Kam. Random Decision Forests. **Proceedings of the 3rd International Conference on Document Analysis and Recognition**, Montreal, QC, p. 278–282., 15 ago. 1995.

[7] Puc Rio. **Clusterização dos dados.** Disponível em: <https://www.maxwell.vrac.puc-rio.br/24787/24787_5.PDF>. Acesso em: 21, abril 2021.

[8] JÚNIOR, R. B. N., et al. **Extração de Informação e Mineração de Dados no Diário Oficial de Pernambuco,** Revista de Engenharia e Pesquisa Aplicada, v.3 n.3, p. 107, 2018.