Mineração de Dados para obtenção do grau de Complexidade de Processos Judiciais

Revista de Engenharia e Pesquisa Aplicada

Alexandre Maciel [1,2] orcid.org/0000-0001-5727-2427
Diego Teixeira [1,2] orcid.org/0000-0001-5727-2427
Marcos Pereira [1] orcid.org/0000-0001-5727-2427
Maria Gabriely [1,2] orcid.org/0000-0001-5727-2427
Matheus Henrique [1] orcid.org/0000-0001-5727-2427

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Pós-graduação em Engenharia de Sistemas, Escola Politécnica de Pernambuco, Pernambuco, Brasil,

E-mail do autor principal: Maria Gabriely mgls@ecomp.poli.br

Resumo

A mineração de dados no contexto dos processos jurídicos explora algoritmos estatísticos sobre os diferentes tipos de dados. Nesse projeto o objetivo foi analisar os dados, a fim de resolver questões de complexidade dos processos judiciais e sua distribuição. O principal objetivo é desenvolver métodos para explorar os tipos únicos de dados dos processos jurídicos, para entender melhor a complexidade de cada processo, a fim de que seja realizado um justo balanceamento de carga de trabalho entre os procuradores da justiça. O sistema atual responsável para realizar essa distribuição de processos é o Sistema de Automação da Justiça (SAJ), no qual determina o grau de complexidade para os processos judiciais, porém na maioria das vezes esta complexidade não é correta e a parte responsável pelo processo necessita alterar essa complexidade após uma revisão. A base de dados disponibilizada possui um considerável desbalanceamento em relação à complexidade do processo, em que 92% dos dados possuem complexidade média. Dessa forma, para obter um melhor balanceamento dos dados foi realizado o under-sampling. Realizado o balanceamento, foram aplicados as técnicas de classificação e agrupamento. 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;

Abstract

DOI: 10.xxxx/s11468-014-9759-3

Data mining in the context of legal cases explores statistical algorithms on different types of data. In this project the goal was to analyze the data in order to address issues of complexity of lawsuits and their distribution. The main goal is to develop methods to explore the unique data types of lawsuits in order to better understand the complexity of each case so that a fair workload balancing between the prosecutors can be performed. The current system responsible for performing this distribution of cases is the Justice Automation System (SAJ), which determines the degree of complexity for the lawsuits, but most of the time this complexity is not correct and the party responsible for the case needs to change this complexity after a review. The available database has a considerable unbalance in relation to the complexity of the process, in which 92% of the data has medium complexity. This way, to obtain a better balancing of the data, an under-sampling was performed. After balancing, the classification and clustering techniques were applied. The results were quite satisfactory for the determined scenarios, obtaining evaluation metrics with a value above 73%.

Key-words: *Data Mining; Process Complexity; Decision Tree; K-Means;*

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 (DM) sobre os diferentes tipos de dados. Nesse projeto seu principal objetivo é analisar os dados, a fim de resolver questões de complexidade dos processos judiciais e sua distribuição [1]. Essa análise está preocupada em desenvolver métodos para explorar os tipos únicos de dados dos processos jurídicos, para entender melhor a complexidade de cada processo para que seja realizado 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, o uso da mineração de dados cria um novo contexto que possibilita a utilização desses dados para um contexto de séries temporais, onde seria possível agora, prever a complexidade de processo "semelhantes" e distribuí-los 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 esta complexidade não é correta e a parte responsável pelo processo necessita alterar essa complexidade após uma revisão. Devido a este problema, 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. Isto gera desequilíbrio na carga de trabalho e insatisfação permanente daqueles que trabalham mais, sem que haja nenhuma compensação.

Dessa forma, o correto grau de complexidade dos processos torna-se 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

O objetivo deste estudo é analisar o conjunto de dados provenientes da procuradoria geral do estado

de Pernambuco (PGE-PE) usando a abordagem de mineração de dados a fim de encontrar um modelo de classificação para complexidade dos processos jurídicos, e através 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 em um desbalanceamento da carga de trabalho. Assim, a identificação da complexidade se torna um fator fundamental para a distribuição igualitária 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 devidos procuradores. A escolha de um determinado procurador não faz parte do escopo apresentado, apenas a classificação de acordo com a complexidade é o objetivo. Também não é objetivo 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, 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 Área do Negócio

Apesar dos avanços recentes no uso 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 [3] 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 de uma 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 [3].

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 demora na resolução dos casos jurídicos. Para isso, foi utilizada a técnica de mineração de dados a fim de se realizar uma análise detalhada dos 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, além disso, verificou-se que a Comarca de Marília tem os processos mais demorados, seguida por Bauru e Santos [2].

Utilizando Inteligência Artificial, a revista do 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,

DOI: 10.xxxx/s11468-014-9759-3

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. Neste 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 com finalidade de encontrar padrões consistentes entre variáveis e depois aplicá-los em novos subconjuntos de dados. Para que boas ideias sejam geradas no final da mineração, existem várias etapas que devem ser seguidas:

(i) Limpeza de dados: processo realizado para remover ruídos, incoerências, dados inconsistentes;

(ii) Integração de dados: onde várias fontes de dados podem ser combinadas;

(iii) Seleção de dados: os dados relevantes para a tarefa de análise são recuperados da base de dados;

(iv) Transformação de dados: os dados são transformados ou consolidados em formulários para mineração, gerando medidas de resumo ou agregação, por exemplo;

(v) Mineração de dados: um processo essencial onde são aplicadas técnicas inteligentes para extrair padrões dos dados;

(vi) Avaliação dos padrões: etapa para identificar os padrões relevantes que representam o conhecimento com base em medidas de interesse;

(vii) Apresentação do conhecimento: onde as técnicas de visualização e de representação do conhecimento são usadas 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, localizar correlações úteis e resgatar informações valiosas dos dados. As tecnologias de mineração de dados são separadas em estatísticas, classificação, agrupamento, regressão e associação. Redes neurais, árvores de decisão, clusterização são alguns exemplos dessas tecnologias, cada uma com sua vantagem, desvantagem e foco de problema. [1]

3 Materiais e Métodos

3.1 Descrição da Base de Dados

A base de dados utilizada neste projeto é composta pela junção das seguintes tabelas do banco de dados do sistema SAJ: Processo, Movimentação e Manifestação. Onde 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. Estas 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 com 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. Para realização desse processo foi utilizado a linguagem de programação Python, tanto na determinação de medidas de resumo como na visualização dos dados. As variáveis utilizadas para esta análise foram complexidade (os três níveis) e quantidade de páginas do processo.

As medidas de resumo calculadas para a variável numérica quantidade de páginas descrevem seu tipo, número de ocorrência, média, valores mínimos e máximos e seus quartis.

Quadro 1: Medidas de Resumo da variável quantidade de páginas

Variável: QNTPAGINASPROCESSOTOTAL	
Tipo	Float
Ocorrência	30469.000000
Média	278.999606
Desvio Padrão	534.648509
Valor Mínimo	2.000000
Quartil 25%	83.000000
Quartil 50%	165.000000
Quartil 75%	318.000000
Valor máximo	16603.000000

Para a variável nominal complexidade 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.

Quadro 2: Medidas de Resumo da variável complexidade do processo

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

Quadro 3: Frequência absoluta e relativa da variável complexidade do processo

Complexidade Processo	Frequência absoluta	Frequência relativa
Muito baixa	8	0.000263
Baixa	821	0.026945
Média	27995	0.918803
Alta	1576	0.051725
Muito alta	69	0.002265

A partir dessas medidas de resumo pode-se perceber que a maior parte da base de dados possui uma complexidade média, podendo ser visualizado melhor na figura 1.

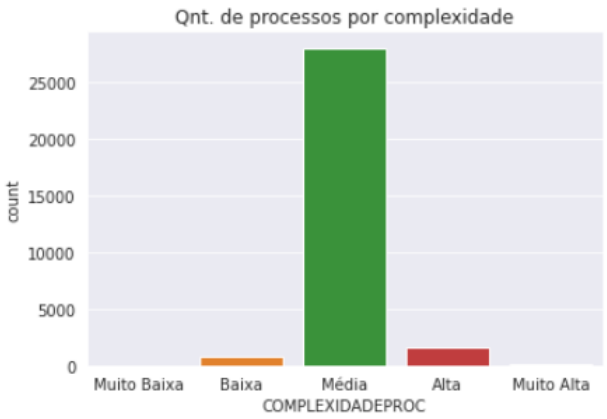

Figura 1: Quantidade de processos por grau de complexidade.

Também foram analisados os campos de complexidade da movimentação e manifestação, como mostrado nas figuras 2 e 3. Observa-se que a nível de movimentações, os processos são melhor distribuídos entre os graus de complexidade. Já a nível de manifestação só existem as complexidades baixa, média e alta.

DOI: 10.xxxx/s11468-014-9759-3

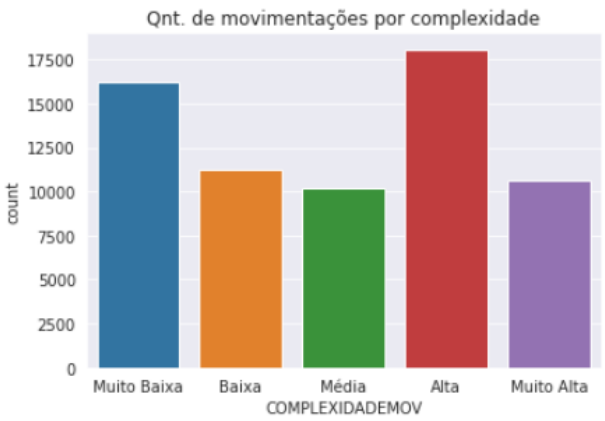

Figura 2: Quantidade de movimentações por grau de complexidade

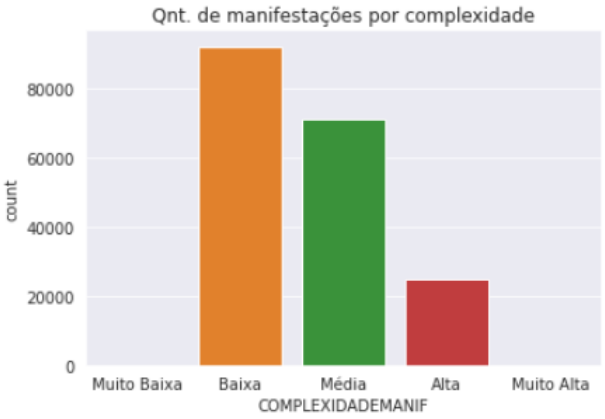

Figura 3: Quantidade de manifestações por grau de complexidade

O boxplot da 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, onde quanto maior a quantidade de páginas mais o processo era complexo (Figura 5).

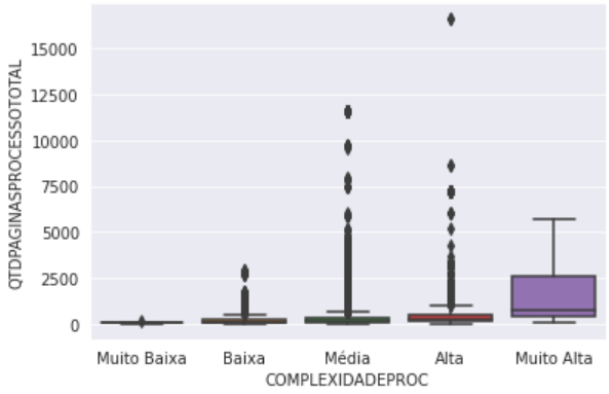

Figura 4: Boxplot do conjunto de dados

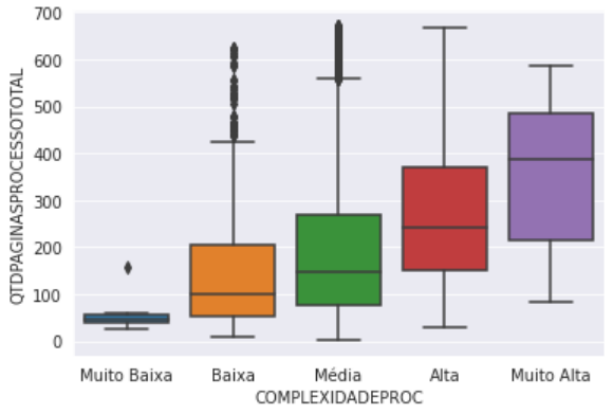

Figura 5: Boxplot do conjunto de dados desconsiderando os outliers.

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 deste projeto 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 a técnica de aprendizado supervisionado árvore de decisão. A Árvore de Decisão é um algoritmo de classificação determinado através 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].

Foi decidido 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. E em relação a Árvore de Decisão, para a escolha dos parâmetros 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.

3.4.2 Agrupamento

Para realizar o agrupamento de processos jurídicos em relação a sua complexidade, foi aplicada a técnica de clusterização K-means. A Clusterização de Dados ou Análise de Agrupamentos é uma técnica de mineração de dados multivariados que através 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].

Foi adicionado 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, onde 92% dos dados possuem complexidade média, como pode ser visto na tabela abaixo.

Quadro 4: Frequências absoluta e relativa das complexidades baixa, média e alta.

Complexidade Processo	Frequência absoluta	Frequência relativa
2.0	23165	0.925194
1.0	649	0.025921
3.0	1224	0.048886

DOI: 10.xxxx/s11468-014-9759-3

Dessa forma, para obter um melhor balanceamento dos dados foi realizado o under-sampling, a tabela a seguir mostra as frequências relativas após o balanceamento.

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

Complexidade Processo	Frequência absoluta	Frequência relativa
2.0	2000	0.516396
1.0	649	0.167570
3.0	1224	0.316034

Feito o balanceamento, foram os escolhidos os atributos de entrada da árvore de decisão, sendo eles a classificação, a subclassificação e o valor da ação. Já em relação aos parâmetros da árvore, os parâmetros *criterion* (medida de qualidade da divisão, neste parâmetro são determinadas as métricas que serão utilizadas) 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 mostra o resultado das métricas de avaliação acurácia, Precisão e Recall na utilização da árvore de decisão variando os parâmetros citados.

Quadro 6: Resultado das métricas acurácia, precisão e recall.

param_criterion	param_max_depth	Accuracy	Precision	Recall
gini	6	0.74801	0.76148	0.70171
gini	7	0.76479	**0.77069**	0.71317
gini	8	0.75911	0.76654	0.70617
gini	9	0.75989	0.76478	0.70702
gini	10	0.76789	0.76700	0.71830
gini	11	0.77382	0.75861	0.73295
gini	12	0.77537	0.75779	0.73493
gini	13	0.77175	0.75088	0.72714
gini	None	0.75704	0.72930	0.72246
entropy	6	0.70902	0.72865	0.68690
entropy	7	0.74102	0.74192	0.70910
entropy	8	0.76995	0.75868	0.72276
entropy	9	0.77072	0.76045	0.72632
entropy	10	0.76994	0.75544	0.72010
entropy	11	0.77614	0.76580	0.72428
entropy	12	0.77407	0.75827	0.72951
entropy	13	**0.78182**	0.76996	**0.73614**
entropy	None	0.75936	0.73306	0.72441

Após encontrar o modelo com os melhores, foi realizada a execução no dataset de teste, obtendo a matriz de confusão representada na figura 6. Foi verificado que processos de média e alta complexidade tiveram um alto valor de acurácia, que são as complexidades com maior frequência no dataset, já a baixa complexidade teve um valor de acurácia inferior devido a menor sua frequência.

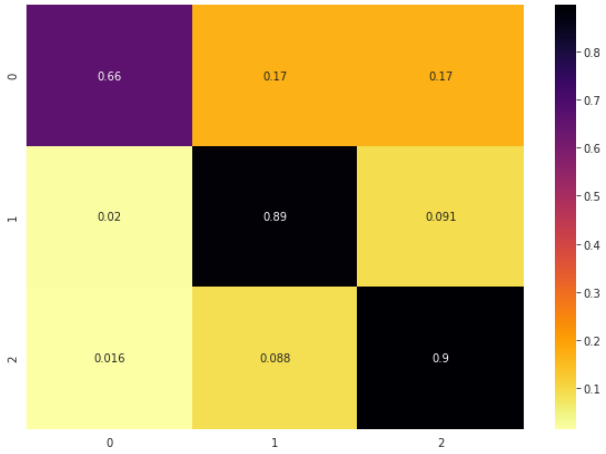

Figura 6: Matriz de confusão da árvore de decisão.

4.2 Resultados - Agrupamento

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

- init: método de inicialização, utilizando k-means++ 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.

Quadro 7: Cenários de execução do k-means.

cenario	c_clust	init	algorith	silhouett _score	davies_b uldin_sc re
1	3	random	auto	0.56819	**0.51675**
2	3	random	full	0.56819	0.51675
3	3	k_means+ +	auto	**0.80785**	0.38360
4	3	k_means+ +	full	0.807856	0.361683

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

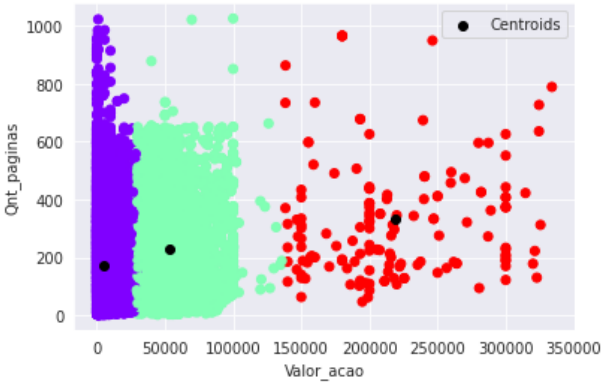

Figura 7: Resultado do k-means com 3 clusters.

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.

Quadro 8: Cenários de execução do k-means + TSNE.

cenario	c_clust	init	algorith	silhouett _score	davies_b uldin_sc re
1	3	k_means+ +	auto	0.364516	0.877514
2	3	k_means+ +	full	**0.364533**	**0.87594**

O gráfico 8 mostra a distribuição dos clusters com a redução de dimensionalidade para duas dimensões (TSNE1 e TSNE2).

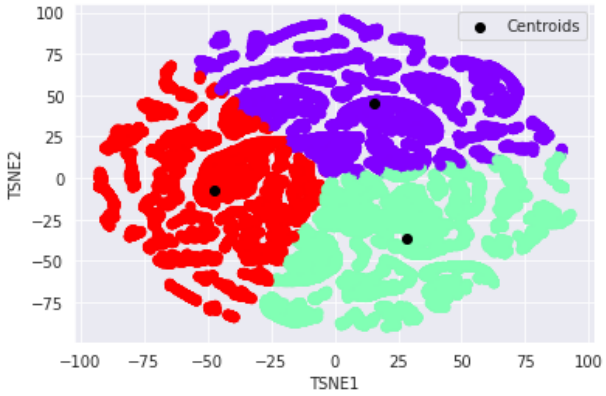

Figura 8: Resultado do k-means+TSNE com 3 clusters.

5 Conclusão e Trabalhos Futuros

Este 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, nós propomos o uso de árvores de decisão para a classificação da complexidade dos processos, assim, obtendo resultados satisfatórios de 78.18% de acurácia média e 73.61% de recall. Indicando que o uso deste 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

DOI: 10.xxxx/s11468-014-9759-3

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.

[2] 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.

[3] 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,** Information Extraction and Data Mining in the Official Gazette of Pernambuco, p. 7.