UM *FRAMEWORK* BASEADO EM APRENDIZADO DE MÁQUINA E DADOS DE PROCESSOS *RES JUDICATA* PARA ANÁLISE E PREVISÃO DE SANÇÕES PENAIS REFERENTES A CRIMES CIBERNÉTICOS

ARTIGO ORIGINAL

FONSECA, Cibele Andréa de Godoy[1], SILVEIRA, Ismar Frango[2]

FONSECA, Cibele Andréa de Godoy. SILVEIRA, Ismar Frango. **Um *framework* baseado em aprendizado de máquina e dados de processos *res judicata* para análise e previsão de sanções penais referentes a crimes cibernéticos**. Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano. 08, Ed. 02, Vol. 01, pp. 36-60. Fevereiro de 2023. ISSN: 2448-0959, Link de acesso: https://www.nucleodoconhecimento.com.br/engenharia-da-computacao/dados-de-processos, DOI: 10.32749/nucleodoconhecimento.com.br/engenharia-da-computacao/dados-de-processos

RESUMO

Este artigo apresenta uma proposta de *framework* para prever penas de multa aplicadas pelos tribunais brasileiros referentes aos crimes cibernéticos utilizando dados coletados dos processos de coisa julgada e do aprendizado de máquina. Essa previsão será feita obedecendo às fases da metodologia de descoberta de conhecimento em banco de dados (em inglês *knowledge discovery in database* – KDD) e com o uso de algoritmos de aprendizado de máquina supervisionado, especificamente os de classificação. Os resultados tendem a ajudar especialistas a descobrir padrões de aplicação de penas de multa pelos tribunais diante de um conjunto de leis por eles utilizadas e, com base nesses padrões, fazer análises e previsões.

Palavras-chave: Sanções penais, Crimes cibernéticos, Análise de dados.

1. INTRODUÇÃO

De acordo com o artigo publicado na CNN Brasil em agosto de 2022, na América Latina, o Brasil, no que se relaciona à ataques cibernéticos, é o segundo país, atrás apenas do México, que obteve 85 bilhões de tentativas. Durante o primeiro semestre de 2022, o Brasil registrou 31,5 bilhões de tentativas de ataques de crimes cibernéticos, número este 94% maior em relação ao primeiro semestre de 2021, quando houve 16,2 bilhões de registros (OLIVEIRA, 2022).

A motivação para conduzir esta pesquisa está ligada aos percentuais de crimes cibernéticos ocorridos recentemente no Brasil e ao intuito de poder apresentar o uso de um *framework* que contempla o uso dos dados dos processos de coisa julgada dos crimes cibernéticos e o aprendizado de máquina, que, ao final, apresenta, como resultados, a análise exploratória dos dados dos tribunais que utilizam áreas do direito, leis, artigos, penas restritivas de liberdade e restritivas de direito, para aplicar penas. Nesse cenário, o uso de algoritmos apresenta a previsão das decisões tomadas pelos tribunais quando visam apenar os crimes cibernéticos por meio da aplicação de multa.

Os resultados referentes às quantidades apresentadas na análise exploratória dos dados contextualizam o crime cibernético como crimes complexos ou crimes simples. Um crime é considerado complexo, quando ele ofende mais de um bem jurídico penalmente tutelado e, quanto ao simples, ele ofende apenas um bem jurídico (RANIERI, 2011). Quanto aos resultados das previsões das decisões tomadas pelos tribunais, quando visam apenar os crimes cibernéticos por meio da aplicação de multa, eles tendem a ajudar especialistas a descobrir padrões de aplicação de penas de multa pelos tribunais diante de um conjunto de leis por eles utilizadas e, com base nesses padrões, fazer análises e previsões.

Este trabalho está organizado em cinco seções. A primeira delas contempla a introdução, na qual o crime cibernético é contextualizado no atual cenário brasileiro.

Nela, também, se explanam a motivação e a justificativa para a elaboração deste trabalho e sua organização formal. A segunda está relacionada à declaração do problema; a terceira, por sua vez, mostra a solução proposta, abrangendo o *framework*, envolvendo todas as etapas relacionadas à organização e à análise de dados com base em KDD (*knowledge discovery in database*), dentre outras atividades específicas para a preparação dos dados a serem utilizados nos modelos e os resultados após sua utilização. Por fim, a quarta seção abrange a interpretação e a avaliação dos resultados do modelo proposto. Na última seção, apresenta-se, a conclusão do estudo desenvolvido e a proposta para futuros trabalhos envolvendo, inclusive, jurimetria que é definida, segundo Nunes (2019) como “a disciplina do conhecimento que utiliza a metodologia estatística para investigar o funcionamento de uma ordem jurídica”.

2. DECLARAÇÃO DO PROBLEMA

Este trabalho discute, como principal problema, a produção de resultados para ajudar especialistas a descobrirem padrões de aplicação de penas de multas pelo tribunal em face de um conjunto de leis aplicadas. Além disso, visa-se propor uma forma de fazer previsões através dos dados coletados dos processos de coisa julgada e do aprendizado de máquina.

Inicialmente, é importante contextualizar o que já é feito para se realizar previsões referentes aos crimes, sendo ou não cibernéticos, com o uso de inteligência artificial e dados heterogêneos de crimes, sobretudo para demonstrar que o trabalho aqui proposto inova na medida em que utiliza um *framework,* envolvendo as fases da metodologia KDD, o aprendizado de máquina e os dados oriundos dos processos com coisa julgada e não de relatórios criminais ou de pesquisas feitas com os cidadãos.

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

2.1 CONTEXTUALIZAÇÃO

Para este estudo, foram pesquisados trabalhos que utilizaram dados heterogêneos com aprendizado de máquina para analisar e prever crimes.

Castro (2020), contemplou em seu trabalho o uso de cinco técnicas de aprendizado de máquina: k-NN, SVM, Random Forest, XGBoost e LSTM. Os dados utilizados são provenientes de fontes heterogêneas, como o conjunto de dados dos antecedentes criminais oficiais coletados junto à Secretaria de Segurança do Estado de Minas de Gerais e um conjunto de dados não oficiais, coletados do *site* "Onde fui roubado".

Souza (2018), utiliza em seu trabalho os algoritmos: Árvore de Decisão, Classificação Gaussiana Naive Bayes e K-NN K-Nearest Neighbor, além de dados criminais fornecidos pelo *site* "Dados Abertos" do Instituto de Segurança Pública do Estado do Rio de Janeiro, pelo qual é possível acessar as bases de dados de antecedentes criminais e a atividade policial no estado do Rio de Janeiro.

Wang (2021), usou os algoritmos K-means e K-Nearest Neighbors, além de um conjunto de dados relativos às taxas de crimes de ódio praticados nos EUA em 2016, antes e depois da eleição presidencial, e em todos os estados dos EUA, de 2010 a 2015.

Safat; Asghar e Gillani (2021), demonstraram a utilização de dados criminais de Chicago, de Los Angeles (EUA) e dos algoritmos Logistic Regression, SVM, Naïve Bayes, KNN, Decision Tree, MLP, Random Forest, XGBoost e LSTM.

Stec e Klabjan (2018), por sua vez, utilizaram redes neurais profundas para prever a contagem de crimes no dia seguinte de suas ocorrências. Para isso, integraram ao *dataset* os dados de crimes de Chicago e Portland (EUA), somadas às informações referentes ao clima, ao censo e ao transporte público.

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

Rayhan e Hashem (2020), reuniram informações referentes aos crimes ocorridos em Chicago (EUA) e o *deep learning,* para conduzir estudos baseados em dados históricos.

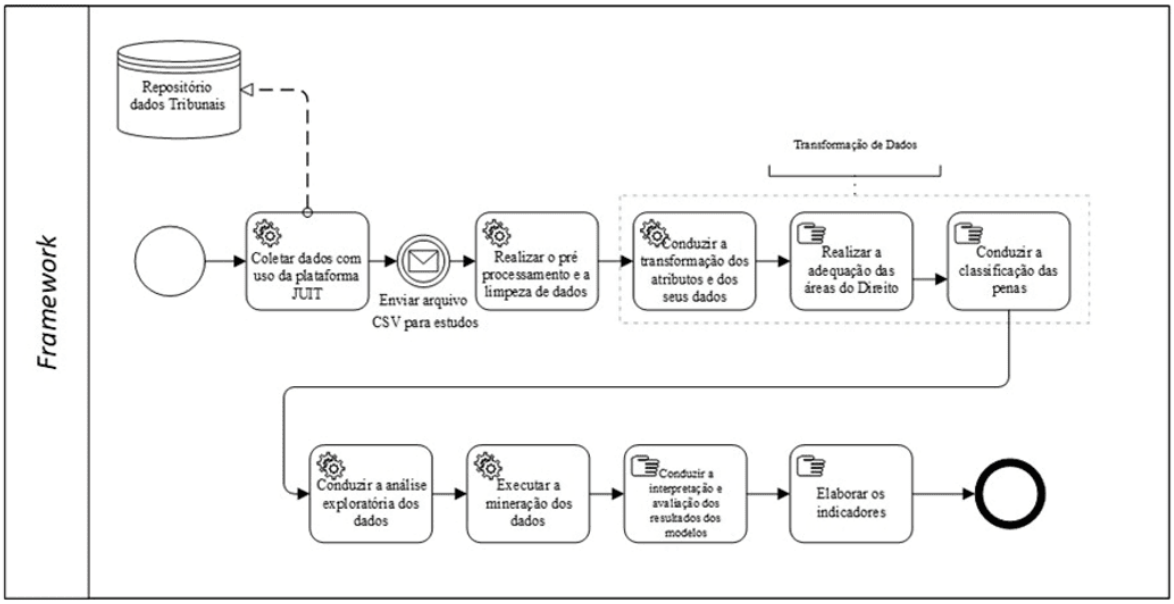
Kshatri *et al*. (2021,) por meio do aprendizado de máquina e do comitê de máquinas, pesquisaram a previsão de crimes na Índia, país desafiador no que diz respeito à identificação da natureza dinâmica dos crimes. O *dataset* foi elaborado com base nos dados criminais de 2001 a 2015, oriundos dos registros de crimes do National Crime Record Bureau (NCRB) de todos os estados da Índia referentes aos relatórios factuais sobre assassinato, estupro e roubo (crimes violentos). Cerca de 60.000 crimes ocorreram nesse período.

3. SOLUÇÃO PROPOSTA

Para descobrir padrões de aplicação de penas de multa pelos tribunais brasileiros, diante de um conjunto de leis por eles utilizadas, e, com base neles, fazer análises e previsões, é proposto um *framework* que contempla o uso dos dados coletados dos processos de coisa julgada referentes aos crimes cibernéticos, através da Metodologia KDD[3] e da utilização de algoritmos de classificação, sendo eles: XGBoost[4], Regressão Logística[5], Floresta aleatória[6], Modelos de Máquinas de Vetores de Suporte[7], e o Comitê de Máquinas[8]. As avaliações dos resultados dos modelos serão feitas com base nas métricas de acuracidade, *precision* e *recall*, da matriz de confusão e da curva ROC. A tecnologia adotada é oriunda da linguagem de programação Python com as suas bibliotecas.

O *framework* contempla as fases da Metodologia KDD e as atividades específicas para o tratamento dos dados, bem como para a análise exploratória dos dados e as análises dos resultados dos modelos, conforme ilustrado na Figura 1.

Figura 1. *Framework* proposto



Fonte: Autores.

A seguir, apresentam-se as atividades que integram o *framework* proposto.

3.1 SELEÇÃO DE DADOS

Nesta fase, foi estruturado um *dataset* intitulado *resjudicata* no formato CSV (do inglês *comma-separated values*), contendo 7.274 registros. Esses dados foram extraídos da plataforma Juit.rimor [9], em 29/06/2022, referentes aos processos de coisa julgada (*res judicata*) dos crimes cibernéticos ocorridos entre janeiro de 2006 a junho 2022. A Tabela 1 contempla o dicionário de dados desse *dataset*.

Tabela 1. Dicionário de dados do *dataset resjudicata*

Nome	Descrição	Tipo	Tamanho
process_id	número do processo	*continuous numeric*	5
Court	local do julgamento	*continuous numeric*	14
data_publish	data da publicação	*continuous numeric*	10
cnj_theme_name_list	área do direito	*nominal categorical*	207

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

cite_legislation_list	lei para aplicação da pena	*nominal categorical*	255
max_value	valor da pena	*continuous numeric*	15
max_value_currency	tipo de moeda	*nominal categorical*	10

Fonte: Autores.

3.2 PRÉ-PROCESSAMENTO E LIMPEZA DE DADOS

Nesta fase, foram conhecidos os atributos e, com o uso da linguagem Python, executadas as atividades de limpeza e padronização de *strings* por meio da remoção de informações, da limpeza de espaços vazios, da correção de caracteres e da desambiguação de informações.

3.3 TRANSFORMAÇÃO DE DADOS

Nesta etapa, também com o uso da linguagem Python, foram conduzidas as seguintes atividades:

1) Transformação dos conteúdos dos atributos *cnj_theme_name_list* e *cite_legislation_list* em colunas. Os atributos do conjunto de dados *Resjudicata* permaneceram de acordo com a Tabela 1;

2) Após a condução da atividade 1, identificou-se que algumas áreas do direito brasileiro não estavam escritas conforme publicadas pelo Conselho Nacional de Justiça (CNJ), o que levou à execução de um *script* em Python para normalizar a situação. O resultado foi um *dataset* intitulado *Resjudicata1*, com 139.991. Esse número de registros se deve ao fato de que por *process_id* os atributos *cite_legislation_list* e *cnj_theme_name_list* foram lidos e cada área do direito encontrada no atributo *cnj_theme_name_list* foi transformada em um registro; posteriormente, as leis encontradas no atributo *cite_legislation_list* também foram transformadas em registros.

Um dos resultados da transformação de dados pode ser visto na análise exploratória dos dados do *resjudicata1,* apresentada na Tabela 2. Nela, constam as quantidades de áreas do direito que integram os processos de coisa julgada utilizadas pelos tribunais em seus julgamentos.

Tabela 2. Quantidade de áreas do direito por Tribunal

Tribunal	Descrição	Quantidade
STJ	Superior Tribunal de Justiça	19
TJSP	Tribunal de Justiça de São Paulo	19
STF	Superior Tribunal Federal	18
TJMS	Tribunal de Justiça do Mato Grosso do Sul	15
TJAL	Tribunal de Justiça do Estado de Alagoas	14
TRF3	Tribunal Regional Federal da 3ª Região	13
TJCE	Tribunal de Justiça do Estado do Ceará	12
TRF1	Tribunal Regional Federal da 1ª Região	11
TRF4	Tribunal Regional Federal da 4ª Região	10
TRF5	Tribunal Regional Federal da 5ª Região	10
TJRS	Tribunal de Justiça do Estado do Rio Grande do Sul	9
TJDFT	Tribunal de Justiça do Distrito Federal e dos Territórios	7
TRF2	Tribunal Regional Federal da 2ª Região	7
TJAM	Tribunal de Justiça do Amazonas	6
TST	Tribunal Superior do Trabalho	6

Fonte: Autores.

Observa-se que os tribunais STJ, TJSP e STF são os que utilizam a maior quantidade de áreas do direito em seus julgamentos por conta da complexidade das situações fáticas que foram analisadas para apenar as condutas delituosas.

3) Para atingir o objetivo deste estudo, foi necessário classificar as penalidades. Para conduzir a classificação, foi criado, com o uso da linguagem Python, um *dataset* intitulado de *Lawarticle* composto pelo atributo *cite_legislation_list* (cujos

dados foram extraídos do *dataset Resjudicata*), bem como pelos atributos *restrictive_of_liberty, restrictive_of_right, reclusion, detention* e *pecuniary_fine* sem quaisquer dados. Seu dicionário de dados pode ser visto na Tabela 3. Criado o *dataset Lawarticle*, a próxima etapa foi analisar as leis que estão no atributo *cite_legislation_list* com base nas leis brasileiras e, manualmente, incluir suas penas nos atributos *restrictive_of_liberty, restrictive_of_right, reclusion, detention e pecuniary_fine.*

Tabela 3. Dicionário de dados do *dataset Lawarticle*

Nome	Descrição	Tipo	Tamanho
cite_legislation_list	lei para aplicação da pena	*nominal categorical*	255
restrictive_of_liberty	pena restritiva de liberdade	*continuous numeric*	1
restrictive_of_right	pena restritiva de direito	*continuous numeric*	1
reclusion	pena de reclusão	*continuous numeric*	1
detention	pena de detenção	*continuous numeric*	1
pecuniary_fine	multa	*continuous numeric*	1

Fonte: Autores.

4) Após classificar as penalidades, os *datasets Resjudicata1* e *Lawarticle* foram unidos com a execução do comando *join* da Linguagem Python, por meio do atributo *process_id*. O resultado foi o *dataset resjudicata1*, composto de 19.172 registros, cujos atributos constam na Tabela 4.

Tabela 4. Dicionário de dados do *dataset Resjudicata1*

Nome	Descrição	Tipo	Tamanho
process_id	número do processo	*continuous numeric*	5
Court	local do julgamento	*continuous numeric*	14
data_publish	data da publicação	*continuous numeric*	10
cnj_theme_name_list	área do direito	*nominal categorical*	207
cite_legislation_list	lei para aplicação da pena	*nominal categorical*	255

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

max_value	valor da pena	*continuous numeric*	15
max_value_currency	tipo de moeda	*nominal categorical*	10
restrictive_of_liberty	pena restritiva de liberdade	*continuous numeric*	1
restrictive_of_right	pena restritiva de direito	*continuous numeric*	1
reclusion	pena de reclusão	*continuous numeric*	1
detention	pena de detenção	*continuous numeric*	1
pecuniary_fine	Multa	*continuous numeric*	1

Fonte: Autores.

A partir da utilização da análise exploratória de dados, a Tabela 5 mostra o número de leis e de artigos usados pelos tribunais para apenar crimes cibernéticos.

Tabela 5. Quantidade de leis e artigos por tribunal

Tribunal	Quantidade
TJSP	1.441
STJ	1.154
TST	337
STF	209
TRF3	198
TRF5	87
TRF1	76
TJMS	48
TRF4	25
TJCE	21
TJAL	17
TRF2	12
TJDFT	8
TJAM	7
TJRS	6

Fonte: Autores.

Como resumo da análise exploratória, afirma-se que os tribunais TJSP, STJ e TST são os que utilizam maior quantidade de leis e artigos se comparados aos outros tribunais, logo, estes são os que julgam crimes cibernéticos de maior complexidade.

A Tabela 6 apresenta as quantidades por tipo de penalidade utilizada pelos tribunais em seus julgamentos. Os resultados demonstram que os tribunais do STJ e do TJSP são os que mais apenam os crimes cibernéticos com multa.

Tabela 6. Quantidade de penalidades por tribunal

Tribunal	Restritiva de liberdade	Restritiva de direito	Multa
STF	45	41	79
STJ	1.033	948	1.618
TJAL	1	1	3
TJAM	0	0	2
TJCE	1	2	7
TJDFT	3	2	4
TJMS	16	14	52
TJRS	5	6	7
TJSP	822	1.451	2.726
TRF1	35	29	46
TRF2	0	0	3
TRF3	195	177	230
TRF4	17	19	22
TRF5	77	66	83
TST	24	402	1.225

Fonte: Autores.

Na Tabela 7, abaixo, apresenta-se a quantidade de penas de reclusão e de detenção aplicadas por tribunal.

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

Tabela 7. Quantidade de penas de reclusão e de detenção por tribunal

Tribunal	Reclusão	Detenção
STF	41	38
STJ	960	960
TJAL	1	1
TJAM	0	0
TJCE	1	1
TJDFT	3	3
TJMS	15	13
TJRS	5	5
TJSP	447	752
TRF1	35	29
TRF2	0	0
TRF3	187	191
TRF4	17	15
TRF5	75	62
TST	15	21

Fonte: Autores.

Ao se observar a distribuição da reclusão e da detenção, verifica-se que o STJ, seguido do TJSP, são os tribunais que mais aplicam essas penalidades. O terceiro é o TRF3.

3.4 MINERAÇÃO DE DADOS

Nesta fase, os algoritmos foram executados. Todavia, antes disso, por meio da linguagem Python, ocorreram as seguintes atividades em sequência:

1) Criou-se um novo *dataset* intitulado *resjudicata2,* baseado no *dataset resjudicata1*. Do *dataset resjudicata2* foi removido o atributo *cnj_theme_name_list*, pois o atributo *cite_legislation_list* foi escolhido para fazer parte do modelo. Após a

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

remoção deste atributo, o *dataset resjudicata2*, com 19.172 registros, ficou composto pelos seguintes atributos: *process_id, date_publish, court, cite_legislation_list, max_value, max_value_currency, restrictive_of_liberty, reclusion, detention, restrictive_of_right*, e pecuniary_fine;

2) Foram eliminadas por *process_id* a repetição de leis e áreas de direito;

3) Os atributos *court, cite_legislation_list, max_value, restritive_of_right* e *pecuniary_fine* foram selecionados;

4) Os registros cujos *pecuniary_fine* estavam vazios foram removidos do conjunto de dados;

5) O atributo *pecuniary_fine* foi transformado em *boolean* (*true* ou *false*) porque os algoritmos trabalham com esse tipo de atributo. Após realizar as atividades 2, 3, 4 e 5, o *dataset resjudicata2* ficou com os atributos *court, cite_legislation_list, max_value, restrictive_of_right,* e *pecuniary_fine*, totalizando 19.171 registros;

6) Foram separados os registros que continham o que é *pecuniary_fine* e o que não é *pecuniary_fine,* ou seja; o que é pena de multa e o que não é pena de multa;

7) Todos os atributos do conjunto de dados foram transformados em números (categorizados).

Quanto ao modelo de classificação com o uso dos algoritmos XGBoost, regressão logística, árvore aleatória, modelos de máquinas de vetores de suporte e comitê de máquinas, que serão estimados em função do atributo *pecuniary_fine* em relação aos atributos *court, cite_legislation_list, max_value, restrictive_of_right,* a equação utilizada foi:

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

$$P(x = multa) = pecunaryfine \sim court + cite\_legislation\_list + \max\_value + restrictive\_of\_right$$

Neste modelo, classifica-se a probabilidade (*odd*) de uma determinada linha ser penalizada com uma multa em razão das entradas (*features* do modelo). Na metodologia de estimação do modelo, a etapa de validação cruzada separou a base de dados, dividindo-a da seguinte forma: 66% para a base de treinamento e 33% para a base de testes e de validação. Não foi necessário ajustar nenhum hiperparâmetro, mas, apenas, estabelecer um critério de randomicidade do gerador de números aleatórios com a *seed* de valor 42. Destaca-se o fato de que os algoritmos foram executados utilizando a linguagem Python.

3.4.1 ALGORITMO XGBOSST

Após a execução do algoritmo XGBoost, integram o resultado modelo: a acurácia, a matriz de confusão, conforme se vê na Tabela 8, as métricas *precision*, *recall*, *f1-score* e *support,* cujos resultados estão na Tabela 9, e a curva ROC, apresentada na Figura 2. Quanto à acurácia, seu resultado é de 0.8790427751695358.

Tabela 8. Matriz de confusão

	Pena de multa	Não pena de multa
Pena de multa	2.551	50
Não pena de multa	397	837

Fonte: Autores.

Tabela 9. *Precision* (%), *recall* (%), *f1-score* (%) e *support* (quantidade)

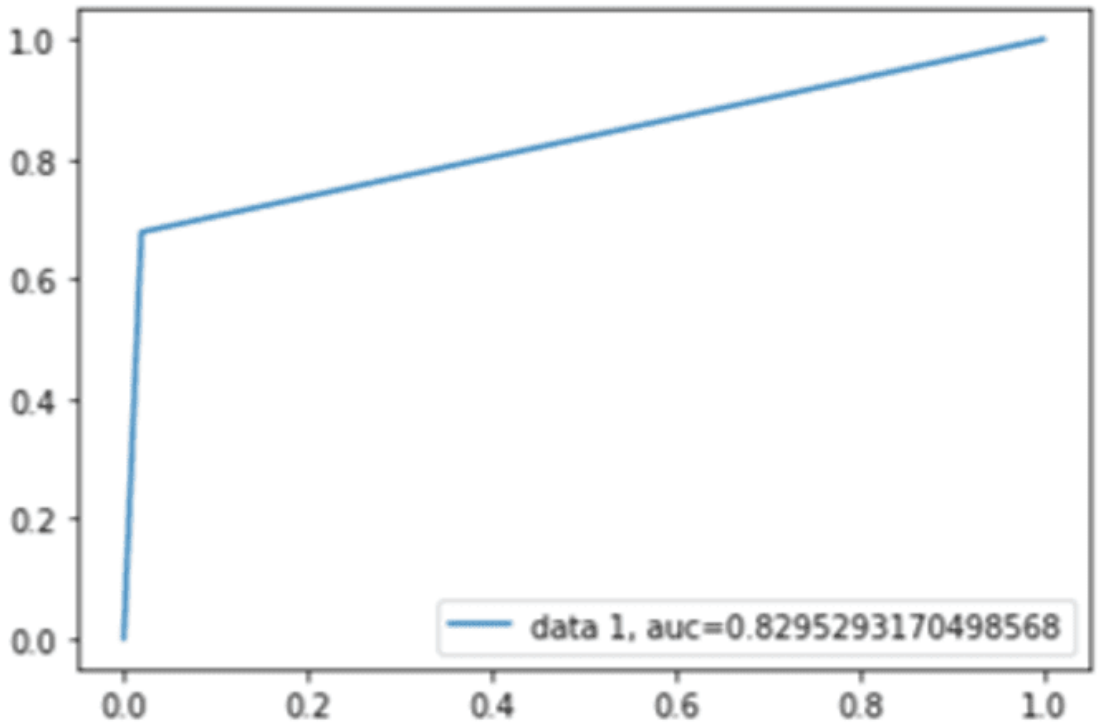
	precision	*recall*	*f1-score*	*support*
False	0.87	0.98	0.92	2.601

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

True	0.94	0.68	0.79	1.234
Accuracy			0.88	3.835
Macro avg	0.90	0.83	0.85	3.835
Weighted avg	0.89	0.88	0.88	3.835

Fonte: Autores.

Figura 2. Curva ROC



Fonte: Autores.

3.4.2 REGRESSÃO LOGÍSTICA

Após a execução do algoritmo de Regressão Logística, integram o resultado do modelo: a acurácia, a matriz de confusão (conforme se vê na Tabela 10), as métricas *precision*, *recall*, *f1-score* e *support* (Tabela 11), e a curva ROC (Figura 3). Quanto à acurácia, seu resultado é de 0.984415753781951.

Tabela 10. Matriz de confusão

	Pena de multa	Não pena de multa
Pena de multa	2.599	2
Não pena de multa	38	1.196

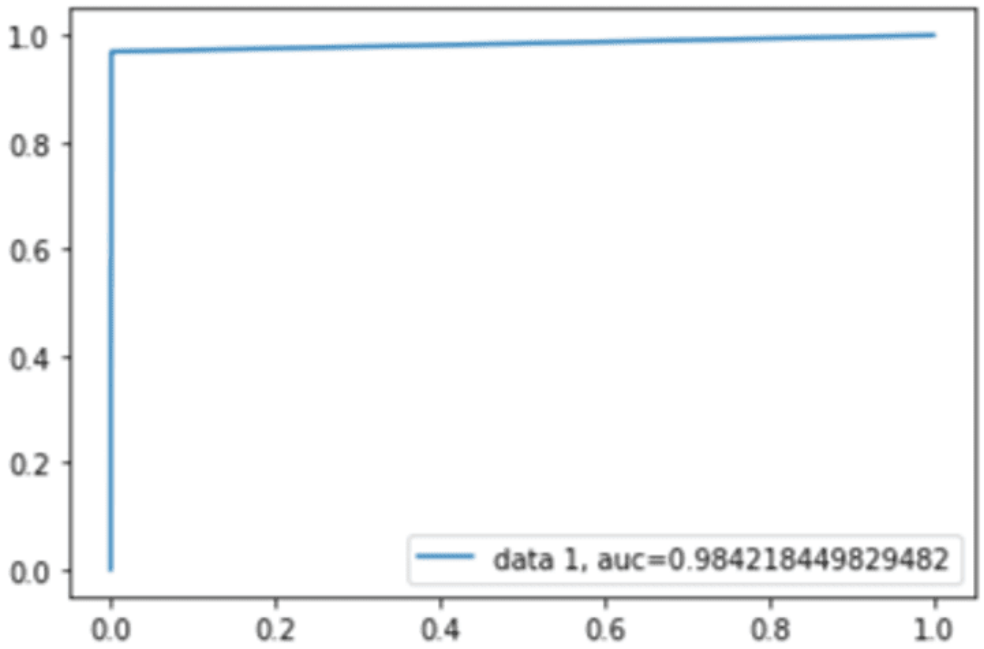
Fonte: Autores.

Tabela 11. *Precision* (%), *recall* (%), *f1-score* (%) e *support* (quantidade)

	precision	*recall*	*f1-score*	*support*
False	0.99	1.00	0.99	2.601
True	1.00	0.97	0.98	1.234
Accuracy			0.99	3.835
Macro avg	0.99	0.98	0.99	3.835
Weighted avg	0.99	0.99	0.99	3.835

Fonte: Autores.

Figura 3. Curva ROC



Fonte: Autores.

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

3.4.3 ÁRVORE ALEATÓRIA

Após a execução do algoritmo árvore aleatória, integram o resultado do modelo: a acurácia, a matriz de confusão (Tabela 12), as métricas *precision*, *recall*, *f1-score* e *support* (Tabela 13), e a curva ROC (Figura 4). Quanto à acurácia, seu resultado é de 0.8903234220135628.

Tabela 12. Matriz de confusão

	Pena de multa	Não pena de multa
Pena de multa	2.599	2
Não pena de multa	399	835

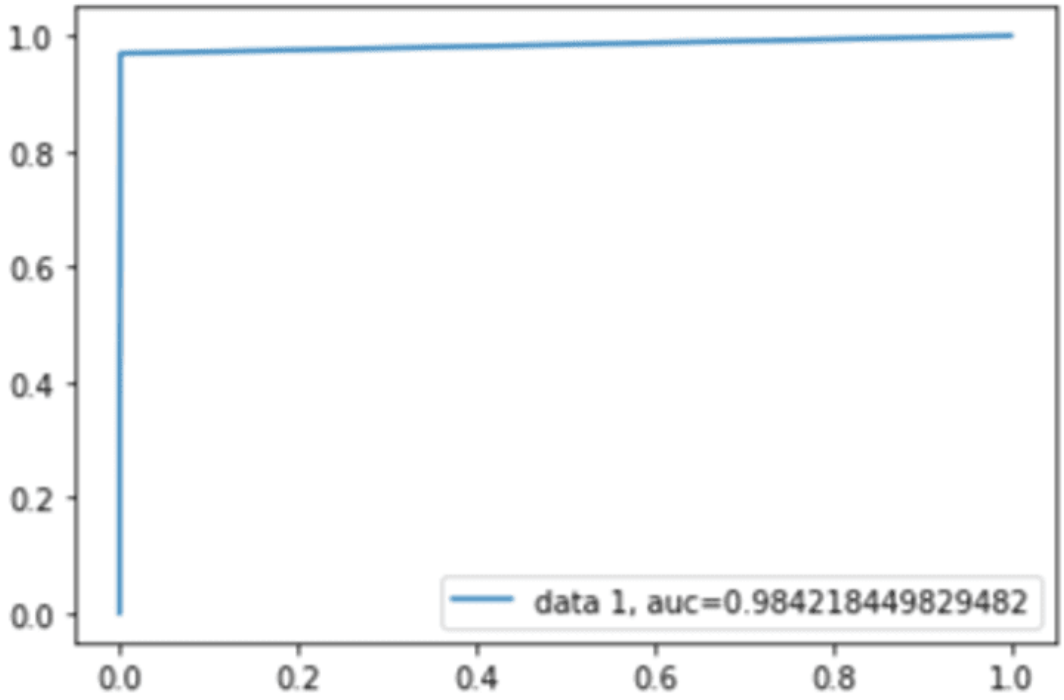
Fonte: Autores.

Tabela 13. *Precision* (%), *recall* (%), *f1-score* (%) e *support* (quantidade)

	precision	*recall*	*f1-score*	*support*
False	0.87	1.00	0.93	2.601
True	1.00	0.68	0.81	1.234
Accuracy			0.90	3.835
Macro avg	0.93	0.84	0.87	3.835
Weighted avg	0.91	0.90	0.89	3.835

Fonte: Autores.

Figura 4. Curva ROC



Fonte: Autores.

3.4.4 MÁQUINAS DE VETORES DE SUPORTE

Após a execução do algoritmo Máquinas de Vetores de Suporte, integram o resultado do modelo: a acurácia, a matriz de confusão (Tabela 14), as métricas *precision*, *recall*, *f1-score* e *support* (Tabela 15), e a curva ROC (Figura 5). Quanto à acurácia, seu resultado é de 0.9952399582681273.

Tabela 14. Matriz de confusão

	Pena de multa	Não pena de multa
Pena de multa	2.600	1
Não pena de multa	9	1.225

Fonte: Autores.

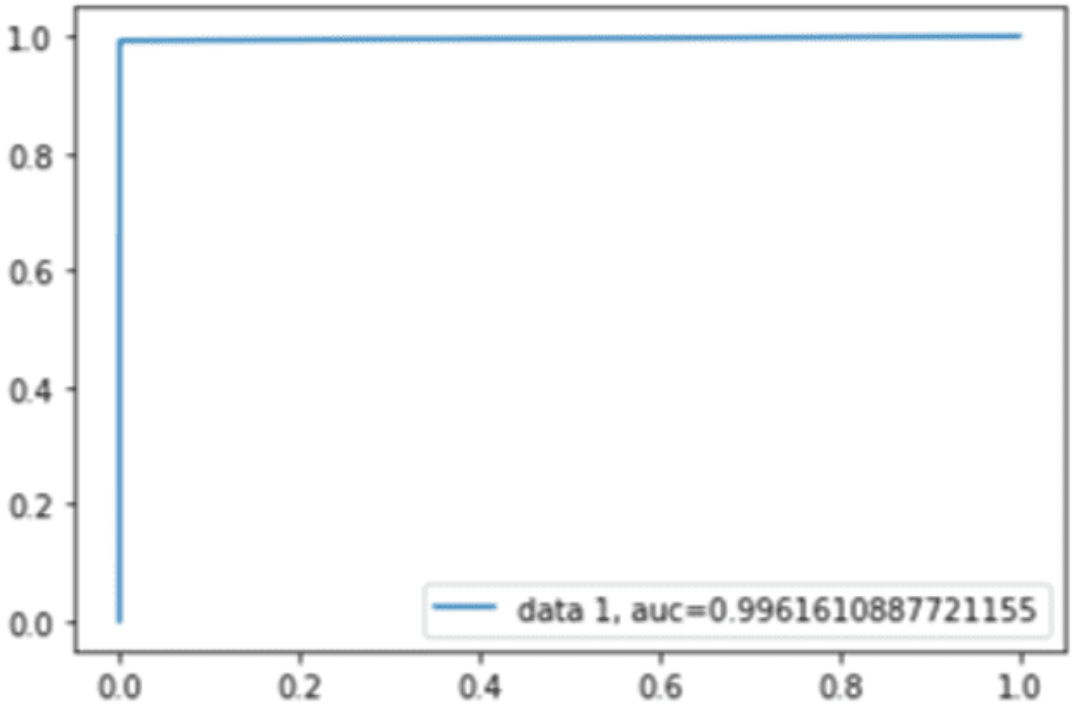
RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

Tabela 15. *Precision* (%), *recall* (%), *f1-score* (%) e *support* (quantidade)

	precision	*recall*	*f1-score*	*Support*	
False	1.00	1.00	1.00	2.601	
True	1.00	0.99	1.00	1.234	
Accuracy			1.00	3.835	
Macro avg	1.00	1.00	1.00	3.835	
Weighted avg	1.00	1.00	1.00	3.835	

Fonte: Autores.

Figura 5. Curva ROC



Fonte: Autores.

3.4.5 COMITÊ DE MÁQUINAS

Após a execução do algoritmo Máquinas de Vetores de Suporte, integram o resultado do modelo: a acurácia, a matriz de confusão (Tabela 16), as métricas

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

precision, *recall*, *f1-score* e *support* (Tabela 17), e a curva ROC (Figura 6). Quanto à acurácia, seu resultado é de 0.896452790818988.

Tabela 16. Matriz de confusão

	Pena de multa	Não pena de multa
Pena de multa	2.599	2
Não pena de multa	378	856

Fonte: Fonte: Autores.

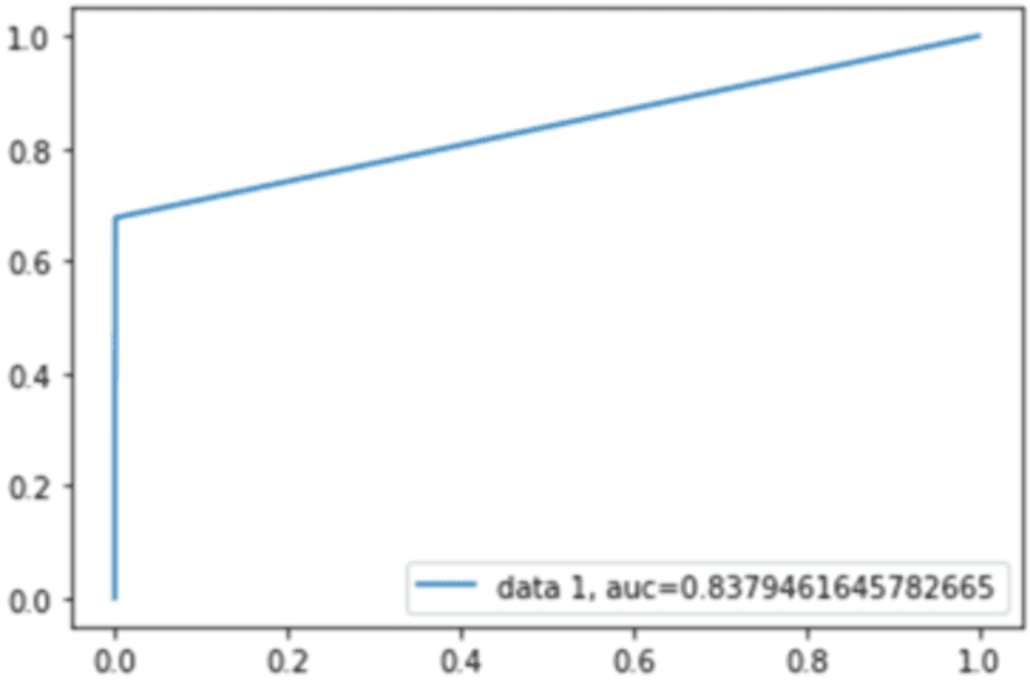
Tabela 17. *Precision* (%), *recall* (%), *f1-score* (%) e *support* (quantidade)

	precision	*recall*	*f1-score*	*Support*
False	0.87	1.00	0.93	2.601
True	1.00	0.69	0.82	1.234
Accuracy			0.90	3.835
Macro avg	0.94	0.85	0.88	3.835
Weighted avg	0.91	0.90	0.90	3.835

Fonte: Autores.

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

Figura 6. Curva ROC



Fonte: Autores.

4. INTERPRETAÇÃO E AVALIAÇÃO DOS RESULTADOS DOS MODELOS

4.1 MODELOS

É possível dizer que as estatísticas coletadas consistem em indicadores de acuracidade global que variaram entre 87,90% a 99,25%. Isso indica que os modelos atingiram bons valores sem a necessidade de se alterar parâmetros (para cada 100 casos, os modelos foram capazes de prever corretamente entre 87,90 vezes até 99,25 vezes, se a pena aplicada contempla multa pecuniária). Destaca-se o fato de que o algoritmo que apresentou a maior acurácia foi o Máquina de Vetores de Suporte.

O próximo passo consiste em analisar as estatísticas *precision* e *recall* para observar o *trade-off* entre viés e variância, o que demonstra precisamente o que está ocorrendo com os modelos no processo de tomada de decisão.

De acordo com as tabelas 9, 11, 13, 15 e 17, as médias ponderadas revelam *precision* que variam entre 0.89% e 100%, e, quanto ao *recall*, variam entre 88% e 100%, indicando um modelo bastante confiável em relação ao balanceamento entre os verdadeiros positivos e os verdadeiros negativos, mitigando as probabilidades de se cometer os erros tipo 1 e os erros tipo 2 (tipo 1: aceitar a hipótese nula quando ela é falsa; tipo 2: rejeitá-la quando é verdadeira). Ao analisar os resultados das métricas *f1-score* dos modelos, seus percentuais variaram entre 88% e 100%. Todas as métricas estão acima de 90%, o que leva à conclusão de que o modelo traz uma boa resposta com base nos dados usados para testes e treino. O algoritmo que apresentou os maiores percentuais de *precision, recall e f1-score* foi o Máquina de Vetores de Suporte.

A qualidade dos modelos também pode ser validada observando-se a matriz de confusão. Neste caso, tanto os verdadeiros positivos quanto os verdadeiros falsos estão aderentes; observam-se poucas situações de falsos positivos e falsos negativos, o que demonstra que os modelos estão generalizando bem.

Por fim, é possível validar a curva ROC dos modelos conforme apresentados nas figuras 2, 3, 4, 5 e 6, cujas estatísticas estão entre 82,95% e 99,61%, aderente com o encontrado até aqui. A curva mostra que os pontos de interseção entre o FPR e o TPR variam entre 0,8295% e 0,9961%, próximo de 1, revelando, assim, alta taxa de precisão. O algoritmo que apresentou o maior percentual próximo a 1 na curva ROC foi o Máquina de Vetores de Suporte.

RC: 135278
Disponível em: https://www.nucleodoconhecimento.com.br/saude/desinfeccao-hospitalar

5. CONCLUSÕES

Devido à indisponibilidade de ocorrências policiais relacionadas a crimes cibernéticos no Brasil, este artigo procurou demonstrar o uso de dados do processo com coisa julgada, juntamente com algoritmos de regressão e classificação com o objetivo de ajudar os especialistas a analisar padrões de aplicação de penas de multas para criminosos envolvidos em crimes cibernéticos, tendo em vista o conjunto de leis e artigos usados pelos tribunais brasileiros para essa modalidade.

Com esses dados em mãos e baseado no *framework* proposto, é possível apresentar conclusões importantes. A primeira delas refere-se à análise exploratória dos dados, pela qual é possível verificar que os tribunais TJSP, STJ e TST são os que utilizam a maior quantidade de leis e de artigos para apenar, quando comparados aos outros tribunais. Ainda, os tribunais STJ, TJSP e STF são os que utilizam a maior quantidade de áreas do direito em seus julgamentos. Baseado nas quantidades e conforme já citado, esses tribunais são os que mais julgaram os crimes cibernéticos que ofenderam mais de um bem jurídico penalmente tutelado, consequentemente, de acordo com a doutrina, foram os que mais julgaram crimes complexos. Desse modo, enfatiza-se que é essencial implementar inteligência e *expertise* no judiciário para tornar suas ações mais eficazes, reduzindo a impunidade sobre esses crimes.

Quanto à execução dos algoritmos, o modelo que utiliza o algoritmo Máquina de Vetores de Suporte foi o que apresentou métricas de valores altos, portanto, dentre todos, é o mais indicado para ser utilizado como parte de modelos para prever penas de multas com o uso dos dados de coisa julgada oriundos do Poder Judiciário. O uso desse algoritmo, ainda, permitiu demonstrar que os dados possuem grande convergência.

Assim, os resultados deste modelo se revelam promissores para a análise de dados jurídicos, em especial, quanto às *features* escolhidas, como: *proxies* das disciplinas

(as grandes áreas do direito, e estas estão relacionadas às leis, que são pedaços menores das disciplinas) que, por sua vez, determinam – de acordo com o tipo de crime cometido – o que acontece com o réu após o juiz proferir a sentença. Importante acrescentar, também, o fato de o ganho de poder gerar a predição a partir de informações públicas que são coletadas por meio de ferramentas computacionais, trazendo conhecimento e suporte para a tomada de decisão através da mineração dos dados.

Portanto, conclui-se que os resultados do *framework* podem ser utilizados para a descoberta de padrões de aplicação de penas de multa pelos tribunais brasileiros diante de um conjunto de leis por eles utilizadas e, com base nesses padrões, fazer análises e previsões.

Em uma futura ampliação do trabalho, podem ser escolhidos novos atributos, por exemplo, quem é o promotor, o juiz, parte do caso, e, assim, ampliar o seu escopo rumo à jurimetria, disciplina que tem se tornado protagonista nas áreas de intersecção entre direito e ciência de dados em geral. Outras propostas incluem o uso de um algoritmo de classificação e previsão com o conjunto de dados de coisa julgada, conforme especificado abaixo:

1. Verificar a associação das áreas do direito, das leis e dos artigos utilizados pelos tribunais para a aplicação das penas, visando entender as complexidades dos crimes cometidos;
2. Verificar se os dados sobre as associações e as ocorrências são suficientes para prever as ações dos criminosos cibernéticos e, assim, permitir a implementação de ações do Estado que garantam melhor proteção dos cidadãos;
3. Estruturar a análise exploratória dos dados dos resultados, bem como dos treinamentos com algoritmos e apresentar seus resultados por meio de gráficos, relatórios e indicadores;

4. Estruturar uma proposta contendo gráficos, relatórios e indicadores das análises realizadas após o treinamento como uma das atividades do processo de investigação criminal e, por meio dessa ação, auxiliar na redução dos riscos de crimes cibernéticos.

REFERÊNCIAS

ALBON, Chris. **Machine learning with Python cookbook**: Practical solutions from preprocessing to deep learning. Sebastopol: O'Reilly Media, 2018.

CORTES, Corinna; VAPNIK, Vladimir. Support vector networks. **Machine Leaming**, vol. 20, p. 273-297, 1995. Disponível em: https://link.springer.com/content/pdf/10.1007/BF00994018.pdf. Acesso em: 07 fev. 2023.

CASTRO, Úrsula Rosa Monteiro de. **Explorando aprendizagem supervisionada em dados heterogêneos para predição de crimes**. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica de Minas Gerais. Belo Horizonte, 2020.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURUSAMY, Ramasamy. **Advances in knowledge discovery and data mining**. California: American Association for Artificial Intelligence, 1996.

GONZALEZ, Leandro de Azevedo. **Regressão Logística e suas Aplicações**. Monografia (Bacharelado em Ciência da Computação) - Universidade Federal do Maranhão. São Luís, 2018. 46 f.

HAYKIN, Simon O. **Neural Networks**: A Comprehensive Foundation. United States: Pearson Education, 2004.

KSHATRI, Sapna Singh; SINGH, Deepak; NARAIN, Bhavana; BHATIA, Surbhi; QUASIM, Mohammad Tabrez; SINHA, G. R. An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach. **IEEE Access**, vol. 9, 2021. p. 67488–67500. Disponível em: DOI:10.1109/access.2021.3075140. Acesso em: 07 fev. 2023.

MÜLLER, Andreas C; GUIDO, Sarah. **Introduction to machine learning with python**: a guide for data scientists. 2ª Ed. Sebastopol: O'reilly Media, Inc., 2017.

NUNES, Marcelo Guedes. **Jurimetria**: Como a Estatística pode reinventar o Direito. São Paulo: Thomson Reuters, Revista dos Tribunais, 2019.

OLIVEIRA, Ingrid. Levantamento mostra que ataques cibernéticos no Brasil cresceram 94%. **CNN Brasil,** agosto de 2022. Disponível em: https://www.cnnbrasil.com.br/tecnologia/levantamento-mostra-que-ataques-ciberneticos-no-brasil-cresceram-94/. Acesso em: 13 set. 2022.

RANIERI, Silvio. **O crime complexo**. São Paulo: Quartier Latin, 2011.

RAYHAN, Yeasir; HASHEM, Tanzima. AIST: An interpretable attention-based deep learning model for crime prediction. **ArXiv**, 2020. Disponível em: https://doi.org/10.48550/arXiv.2012.08713. Acesso em: 07 fev. 2023.

SAFAT, Wajiha; ASGHAR, Sohail; GILLANI, Saira Andleeb. Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. **IEEE Access**, vol. 9, 2021. Disponível em: 10.1109/ACCESS.2021.3078117. Acesso em: 07 fev. 2023.

SOUZA, José Renato Mendes de. **Utilização de aprendizagem de máquina na predição de crime.** Trabalho de Conclusão de Curso (Tecnologia em Sistemas de Computação) - Universidade Federal Fluminense. Niterói, 2018. 54 f.

STEC, Alexander; KLABJAN, Diego. Forecasting crime with deep learning. **ArXiv,** 2018. Disponível em: https://doi.org/10.48550/arXiv.1806.01486. Acesso em: 07 fev. 2023.

WANG, Shaoxuan. Hate crime analysis based on artificial intelligence methods. **E3S Web of Conferences**, vol. 251, 2021. Disponível em: https://doi.org/10.1051/e3sconf/202125101062. Acesso em: 07 fev. 2023.

APÊNDICE - REFERÊNCIA NOTA DE RODAPÉ

3. Segundo Fayyad *et al*. (1996), KDD é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. Elas que se organizam nas etapas de: seleção, pré-processamento e limpeza, transformação e mineração de dados (*data mining*), além da interpretação e da avaliação dos resultados.

4. Para Müller e Guido (2017), os modelos de *gradient boosting*, intitulados também de *gradient boosted regression trees*, são composições de múltiplas árvores de decisão que buscam incorporá-las para formar um modelo mais eficiente. Esses modelos podem ser usados para tarefas, tanto de regressão, quanto de classificação.

5. Segundo Gonzalez (2018), a regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo

que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias.

6. Segundo Albon (2018), o método da floresta aleatória é uma composição de múltiplas árvores de decisão que usam seções do conjunto de dados para serem treinadas.

7. De acordo com Cortes e Vapnik (1995), “Máquinas de vetores de suporte” (MVS): trata-se de um modelo desenvolvido por Vapnik em 1995, fundamentado na teoria de aprendizado estatístico, utilizado para a classificação e a regressão de dados. É usado para estimar uma função que classifique dados de entrada em duas classes (normalmente, mas é multiclasses).

8. Haykin (2004), define Comitê de Máquinas (ensemble) como uma proposta baseada no princípio de “dividir-para-conquistar”, que busca superar o desempenho de uma máquina de aprendizado operando isoladamente. Trata-se de método de aprendizado (supervisionado ou não supervisionado), cujo objetivo é aumentar a capacidade de generalização de estimadores (aproximadores de função/regressores, classificadores etc.)

9. https://juit.io/

Enviado: Janeiro, 2023.

Aprovado: Fevereiro, 2023.

[1] Bacharel em Ciências com Licenciatura em Matemática, Mestre em Engenharia Elétrica e Computação, Doutoranda em Engenharia Elétrica e Computação com foco em inteligência artificial e crimes cibernéticos. Profissional com mais de 20 anos de experiência em cargos executivos responsável pela Área de Tecnologia da informação. ORCID: 0000-0001-5270-1480. CURRÍCULO LATTES: http://lattes.cnpq.br/0780179952438526.
[2] Orientador. Graduado em Matemática, Mestre em Ciências e Doutor em Engenharia Elétrica. Professor Adjunto da Universidade Presbiteriana Mackenzie e Professor Titular da Universidade Cruzeiro do Sul. Membro das comunidades científicas LA CLO (Comunidade Latino-Americana de Tecnologias de Aprendizagem), HCI-Collab (Rede Colaborativa para apoiar os processos de ensino e aprendizagem em área de Interação Humano-Computador em nível Iberoamericano), e VG-Collab (Rede Colaborativa de pesquisa e desenvolvimento de jogos na Iberoamérica). ORCID: 0000-0001-8029-072X. CURRÍCULO LATTES: http://lattes.cnpq.br/3894359521286830.