UNIVERSIDADE PAULISTA – UNESP
Faculdade de Engenharia – Campus de Guaratingetá-SP

OSIRIS DE SOUZA E SILVA NETO

O uso de modelagem estatística para análise preditiva: uma aplicação no setor de serviços bancários na área Jurídica

Guaratinguetá - SP
2022

Osiris de Souza e Silva Neto

O uso de modelagem estatística para análise preditiva: uma aplicação no setor de serviços bancários na área Jurídica

Dissertação apresentada à Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, para a etapa de qualificação como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção – Mestrado Profissional.

Orientadora: Prof.ª Dr.ª Gislaine Cristina Batistela
Coorientador: Prof. Dr. Bruno Chaves Franco

Guaratinguetá - SP

2022

FICHA CATALOGRÁFICA

Silva Neto, Osiris de Souza e
S586u O Uso de modelagem estatística para análise preditiva: uma aplicação no setor de serviços bancários na área Jurídica / Osiris de Souza Silva Neto – Guaratinguetá, 2024.
50 : il.
Bibliografia: f. 45-50

Dissertação (Mestrado) – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2026.

Orientadora: Prof.ª Dr.ª Gislaine Cristina Batistela
Coorientador: Prof. Dr. Bruno Chaves Franco

1. Direito do trabalho. 2. Desempenho - medição. 3. Processo decisório. I. Título.

CDU 65.012.4 (043)

Luciana Máximo
Bibliotecária CRB-8/3595

IMPACTO POTENCIAL DESTA PESQUISA

A pesquisa apresenta contribuição científica ao integrar métodos estatísticos e de aprendizado de máquina à área jurídica, com aplicação técnica e inovadora no setor bancário. Espera-se impacto na eficiência econômica e na qualificação da tomada de decisão em demandas trabalhistas empresariais.

POTENTIAL IMPACT OF THIS RESEARCH

The research presents a scientific contribution by integrating statistical methods and machine learning into the legal field, with technical and innovative application in the banking sector. It is expected to impact economic efficiency and enhance decision-making in corporate labor disputes.

unesp

UNIVERSIDADE ESTADUAL PAULISTA
CAMPUS DE GUARATINGUETÁ

OSIRIS DE SOUZA E SILVA NETO

ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE
"MESTRE EM ENGENHARIA DE PRODUÇÃO"

PROGRAMA: ENGENHARIA DE PRODUÇÃO CURSO: MESTRADO PROFISSIONAL

APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO

gov.br Documento assinado digitalmente
GISLAINE CRISTINA BATISTELA
Data: 10/07/2023 17:47:29-0300
Verifique em https://validar.iti.gov.br

B A N C A E X A M I N A D O R A:

gov.br Documento assinado digitalmente
GISLAINE CRISTINA BATISTELA
Data: 10/07/2023 00:15:27-0300
Verifique em https://validar.iti.gov.br

Prof.ª Dr.ª GISLAINE CRISTINA BATISTELA

Orientador - UNESP

participou por videconferência

gov.br Documento assinado digitalmente
ANDRE LUIS DEBIASO ROSSI
Data: 10/07/2023 17:00:18-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. ANDRE LUIS DEBIASO ROSSI

UNESP

participou por videconferência

gov.br Documento assinado digitalmente
TERESA CRISTINA MARTINS DIAS
Data: 10/07/2023 17:40:03-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. TERESA CRISTINA MARTINS DIAS

UFSCAR

participou por videconferência

JANEIRO de 2023

DADOS CURRICULARES

OSIRIS DE SOUZA E SILVA NETO	
NASCIMENTO	17.03.1989, em Marília, SP
FILIAÇÃO	Osíris de Souza e Silva Filho Terezinha de Souza e Silva
GRADUAÇÃO	Engenharia Mecânica, concluída em 2012
PÓS-GRADUAÇÃO	MBA em Finanças pela Universidade Federal de São Carlos (UFSCAR) de Sorocaba/SP, concluído em 2016

RESUMO

Nesta pesquisa o objetivo é comparar, por meio de indicadores de desempenho, a precisão de dois modelos na previsão dos resultados (ganho ou perda das causas) das ações trabalhistas abertas contra instituições financeiras brasileiras. Para o estudo foram consumidos dados reais das sentenças, disponíveis nos Tribunais de Justiça e Tribunais Regionais Federais, e processados, seguindo o procedimento de Knowledge Discovery in Databases (KDD). Para possibilitar a aplicação dos modelos e aumentar a confiabilidade dos resultados, foi feito o pré-processamento dos dados, realizando cinco tarefas: tratamento dos dados (agrupamentos e eliminação de variáveis); imputação de dados faltantes; normalização (para as variáveis numéricas); transformação de variáveis qualitativas em variáveis dicotômicas dummies (aplicada às variáveis categóricas); e redução de dimensionalidade e multicolinearidade. Para a construção do modelo preditivo foram aplicadas duas metodologias para fins de comparação: Regressão Logística (RL) e o Light Boosting Gradient Model (LGBM). Quanto à conclusão, foram obtidos resultados expressivos quando comparados com os mesmos tipos de estudo encontrados na literatura, evidenciando que os modelos de RL e LGBM com machine learning podem ser altamente eficientes na predição dos resultados aplicados no contexto de Direito, sendo uma ferramenta importante na estratégia de defesa de ações trabalhistas para empresas deste seguimento, principalmente quanto à tomada de decisão de entrar em um acordo ou investir na defesa.

Palavras-chave: direito trabalhista; jurimetria; mineração de dados; regressão logística.

ABSTRACT

This research aims to compare, through performance indicators, the accuracy of two models in predicting the outcomes (win or loss) of labor lawsuits filed against Brazilian financial institutions. The study was conducted using real sentencing data available from the Courts of Justice and Regional Federal Courts, processed following the Knowledge Discovery in Databases (KDD) procedure. To enable the application of the models and enhance the reliability of the results, data preprocessing was performed, comprising five tasks: data treatment (grouping and variable elimination); imputation of missing data; normalization (for numerical variables); transformation of qualitative variables into dummy dichotomous variables (applied to categorical variables); and dimensionality and multicollinearity reduction. For the development of the predictive model, two methodologies were applied for comparison: Logistic Regression (LR) and the Light Gradient Boosting Model (LGBM). The findings revealed significant results compared to similar studies in the literature, demonstrating that LR and LGBM models with machine learning can be highly effective in predicting case outcomes within the legal context. This makes them a valuable tool in the strategic defense of labor lawsuits for companies in this sector, particularly in the decision-making process regarding settlement agreements or investing in defense.

Keywords: labor law; legal analytics; data mining; logistic regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

As fortes transformações sofridas pelo segmento do setor financeiro nas últimas décadas vêm influenciando drasticamente o ambiente de competição do setor bancário. Fatores como a expansão global de tecnologias seguras e eficientes, aplicáveis aos serviços financeiros (exemplo *blockchain*), e a flexibilização de regulamentações trouxeram um aumento significativo de novos concorrentes para as tradicionais instituições financeiras. Dentro desse cenário, os objetivos estratégicos das grandes empresas financeiras exigem inovação e melhoria contínua dos seus processos (LIMA, 2014).

Outra dificuldade, quando se trata de setores de serviços, é a natureza intangível dos produtos que esse setor oferece e a alta variabilidade presente nos seus processos (Canel, 2000; Arfmann; Topolansky, 2014), a qual está associada à dependência de interação humana e se refere ao padrão de execução das atividades operacionais, como tempo de processamento e qualidade do serviço entregue. Esse contexto torna a gestão deste setor complexa, com alto índice de desperdícios e problemas com a qualidade.

Tal cenário obriga as empresas financeiras a investirem cada vez mais em uma melhor utilização de seus recursos, principalmente direcionando-os para atividades que garantam a qualidade de seus serviços e a inovação de seus produtos. Porém, quando se trata de empresas de grande porte (com mais de 5 mil funcionários), alguns custos não relacionados ao escopo de sua atuação começam a ganhar significância e a comprometer essa estratégia de investimento, caso de custos relacionados com ações trabalhistas.

Segundo Salama e Carlotti (2018), as empresas do setor bancário brasileiro apareceram no ano de 2017 como réus em números superiores a mil processos trabalhistas, sendo 89% dos casos dando procedência total ou parcial aos reclamantes trabalhadores, resultando em valores médios de condenação de 28 mil reais por processo.

1.2 QUESTÃO DE PESQUISA

Os custos relacionados aos processos trabalhistas não envolvem apenas o valor das sentenças, mas também custos administrativos com advogados, pessoal e *compliance*.

Contudo, diminuir esses gastos depende de vários fatores externos e internos às instituições, e dentre os que pertencem à organização destaca-se a efetividade de sua estratégia perante cada processo. A estratégia compreende desde a decisão de entrar com um acordo ou defesa até quais subsídios deverão ser levantados para amparar a atuação dos advogados. Entende-se como subsídios documentos, informações processuais, histórico de processos similares, cálculos de perda financeira, entre outros.

No que se refere à efetividade da estratégia, Katz (2013) afirma que o alto volume de informações reais e valiosas sobre processos ativos, assim como legado e concorrentes irão definir grande parte da inovação dentro do assunto de serviços legais. Nunes e Coelho (2010) denominam essa área de conhecimento que busca entender os fenômenos jurídicos por meio da aplicação estatística e probabilística como jurimetria.

Portanto, nesse contexto, entende-se a jurimetria como ferramenta fundamental para o gerenciamento da estratégia nas áreas responsáveis pelo setor jurídico das empresas, de forma a torná-las mais eficazes, identificando a probabilidade de cada estratégia gerar resultados negativos antes mesmo de executá-las. Isso leva à questão de pesquisa deste trabalho: como aumentar a eficácia da estratégia sobre os processos trabalhistas para empresas do setor financeiro por meio de análises preditivas baseadas nos históricos processuais disponíveis?

1.3 OBJETIVO, DELIMITAÇÃO E JUSTIFICATIVA

1.3.1 Objetivos geral

Nesta pesquisa, o objetivo é comparar, por meio de indicadores de desempenho, a precisão de dois modelos em prever os resultados (ganho ou perda das causas) das ações trabalhistas abertas contra instituições financeiras brasileiras.

1.3.2 Delimitação da pesquisa

O trabalho limita-se a comparar os resultados de dois modelos preditivos (regressão logística e LGBM) desenvolvidos, utilizando informações retiradas das petições iniciais e das sentenças, disponíveis nos Tribunais de Justiça e nos Tribunais Regionais Federais, relacionadas apenas às ações trabalhistas abertas contra instituições financeiras brasileiras. Ademais, não são abordados temas relacionados a ações penal e civil.

1.3.3 Justificativa

Segundo Pompeu (2018), o Brasil tinha, em 2018, mais de 80 milhões de processos judiciais ativos, gerando custos para empresas litigadas, não somente envolvendo o valor das sentenças, mas também custos administrativos com advogados, pessoal, e custo de oportunidade com o provisionamento dessas despesas. Logo, tornar mais efetivo os processos que tratam dessas ações torna-se uma questão importante para a sustentação da competitividade para empresas de grande porte.

Quanto ao estudo dessa volumetria de processos, o Conselho Nacional de Justiça estabeleceu a gestão das demandas repetitivas dos grandes litigantes como um dos macros desafios do Poder Judiciário no período 2015-2020 (Tribunal Regional do Trabalho, 2014), em conjunto com o Conselho Superior da Justiça do Trabalho, que definiu o tema como objetivo estratégico da Justiça do Trabalho nesse mesmo lapso temporal (Conselho de Justiça Federal, 2014). Assim, para o exercício do ano de 2018, foram identificados no Tribunal Regional do Trabalho do Rio de Janeiro os dez maiores litigantes, entre 31.12.2017 e 31.12.2018, sendo que 45%, correspondendo a 7.130 processos, estão concentrados em instituições financeiras (Itaú, Bradesco, Caixa Econômica e Santander), conforme pode ser visualizado na Figura 1.

Figura 1 – Acervo dos dez maiores litigantes 2017

Posição	Litigante	Em 31.12.2017	Em 31.12.2018	Variação
1	ITAÚ UNIBANCO	3.106	2.663	-14,3%
2	VIA VAREJO S.A.	2.913	1.823	-37,4%
3	PETRÓLEO BRASILEIRO S.A. - PETROBRAS	2.556	1.664	-34,9%
4	COMPANHIA ESTADUAL DE ÁGUAS E ESGOTOS - CEDAE	2.440	1.320	-45,9%
5	BANCO BRADESCO S.A.	2.070	1.956	-5,5%
6	PRO SAÚDE – ASSOC. BENEFICENTE DE ASSIST. SOCIAL E HOSPITALAR	1.654	1.056	-36,2%
7	SERVIÇOS DE REDE S.A. - SEREDE	1.642	1.215	-26,0%
8	CAIXA ECONÔMICA FEDERAL - CEF	1.632	1.223	-25,1%
9	BANCO SANTANDER (BRASIL) S.A.	1.580	1.288	-18,5%
10	HOSPITAL E MATERNIDADE THEREZINHA DE JESUS	1.316	1.667	26,7%
Total de processos dos dez maiores litigantes pendentes de julgamento no TRT/RJ		**20.909**	**15.875**	**-24,1%**

Fonte: Tribunal Regional do Trabalho (2018)

Quanto ao estudo dos desfechos dos processos, a subjetividade no julgamento torna ainda mais difícil a mitigação desses gastos para as empresas litigantes.

> A imprevisibilidade das decisões judiciais também acarreta custos econômicos. Primeiro, porque a imprevisibilidade do Judiciário desincentiva acordos. Isso contribui para o aumento da litigância, aumentando a necessidade de movimentação

> da máquina judicial. Segundo, porque ao dificultar a precificação de investimentos, riscos e retornos, a imprevisibilidade tende a deprimir o mercado de trabalho e a reduzir a atividade econômica de modo geral (SALAMA; CARLOTTI; YEUNG, 2019, v.1, p 4).

Por fim, a alta volumetria e a imprevisibilidade das decisões, uma vez que cada juiz tem seu próprio convencimento a respeito de uma determinada matéria, tornam os métodos estatísticos importante ferramenta para que esses litigantes tenham uma compreensão melhor do todo. De acordo com uma pesquisa da Lex Machina (2019), a estratégia jurídica para tribunais e juízes pode ser otimizada em até 98% com análises de dados judiciais.

Na prática, as predições realizadas por algoritmos e modelos estatísticos aplicados a uma grande quantidade de dados (*big data*) podem auxiliar, entre outros, na quantificação de riscos e no cálculo da probabilidade de vitória em litígio, que são informações relevantes para alcançar resultados promissores nas sentenças (Nunes, 2020). Além disso, permite o melhor gerenciamento da política estratégica de acordos, auxiliando na decisão de entrar com uma defesa ou com um acordo para cada caso. Esses insumos tornam o Direito mais previsível, descomplicando análises judiciais e tornando mais eficiente esse processo operacional.

Visto esse contexto, o presente trabalho é relevante no campo teórico, pois engloba o desenvolvimento de um modelo que possibilita a ordenação dos dados processuais das petições e das sentenças, considerando a perda de causas trabalhistas via modelo de Regressão Logística.

Nesse sentido, este trabalho contribuirá de forma prática no processo de decisão da estratégia a ser seguida pela empresa financeira contra as ações trabalhistas recebidas (escolha de entrar com um acordo ou uma defesa), trazendo uma ferramenta estatística que aumenta a eficácia dessa decisão e a eficiência da utilização dos recursos internos responsáveis pelas análises desses processos.

1.4 ETAPAS E CLASSIFICAÇÃO DA PESQUISA

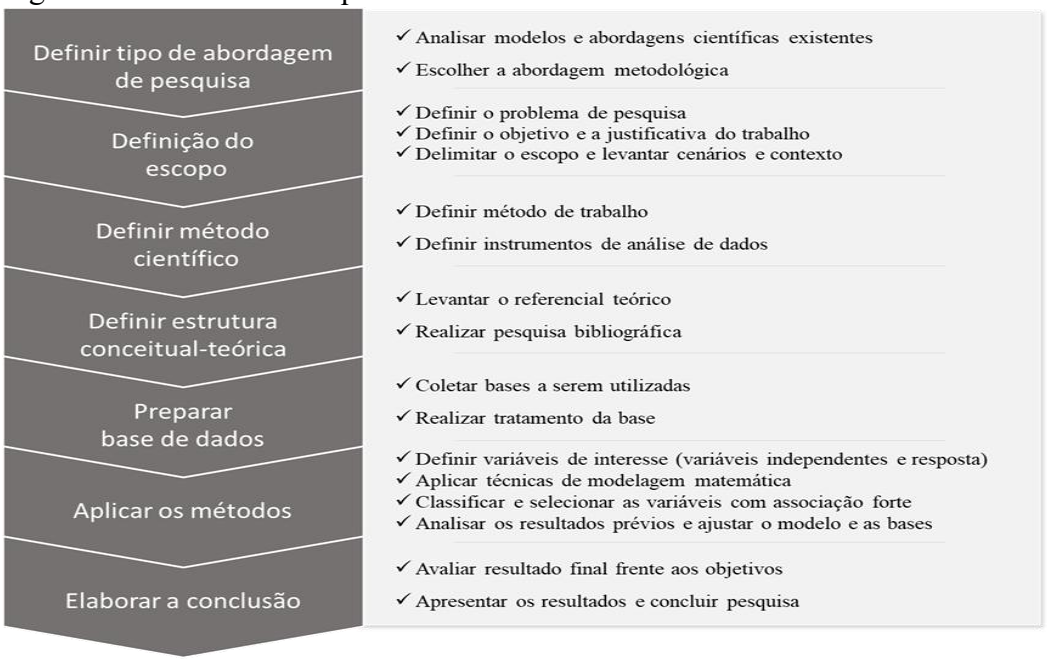
Esta pesquisa utiliza como método a modelagem que, segundo Bertrand e Fransoo (2002), é caracterizada por pesquisas cujo modelo relaciona as variáveis independentes com a variável resposta, permitindo análise e testes das relações e interações entre as variáveis.

Quanto à forma de abordagem, esta investigação pode ser classificada como quantitativa aplicada, que, segundo Furlan (2017), tem como característica principal, além de realizar o estudo dos dados, buscar realizar análises críticas, discutir e explicar os resultados.

Essa aplicação foi realizada por meio da análise de dados reais de uma instituição financeira brasileira para o estudo da associação das variáveis independentes nominais (neste caso os fatos relacionados aos conflitos judiciais trabalhistas) com a variável dependente (neste caso a sentença final dos processos trabalhistas), utilizando-se análises estatísticas descritivas e modelagem estatística por regressão logística e LGBM (*Light Gradient Boosting Machine*).

Durante o desenvolvimento da pesquisa, foram considerados artigos publicados entre 2010 e 2021, pesquisados nos repositórios *Scopus* e *Web of Science*. Na Figura 2, está representado o detalhamento do método aplicado, com suas etapas em ordem de execução.

Figura 2 – Modelo de Pesquisa



Fonte: elaborado pelo autor (2022)

2 FUNDAMENTAÇÃO TEÓRICA

2.1 INTRODUÇÃO

Segundo Soares Junior e Quintella (2005), o processo de extração de informações úteis de uma base de dados não é trivial, e o crescimento rápido do volume dos dados gera a necessidade de se extrair sistematicamente o conhecimento nelas contidas. Fayyad (1996) identificou o aumento da complexidade desse processo já na década de 90 que, segundo ele, é causado pela sobrecarga de dados trazido pela "era da informação". Propondo solucionar esse problema foi desenvolvida uma nova área de pesquisa conhecida como Knowledge Discovery in Databases (KDD), ou Descoberta de Conhecimento em Bases de Dados. Esse processo visa extrair conhecimento de uma grande base de dados que frequentemente apresentam diversos problemas como excesso de valores faltantes, inconsistências e desbalanceamento de classes (BATISTA, 2003).

O KDD é muitas vezes confundido com a mineração de dados (MD), no entanto a MD é apenas um dos passos que compõem o KDD (Silva.; Peres; Boscarioli, 2016). O KDD pode ser dividido em cinco etapas que compõem os três grandes grupos: o primeiro é o pré-processamento, que inclui as etapas de seleção, de limpeza e de transformação dos dados; já o segundo grupo, a mineração de dados, gera as informações (conhecimento) por meio de, por exemplo, modelagens estatísticas; e enfim no terceiro grupo, chamado de pós-processamento, ocorre a interpretação dos dados (SOARES JUNIOR; QUINTELLA, 2005).

Neste capítulo são apresentadas as fundamentações teóricas que deram suporte aos procedimentos realizados neste trabalho. As seções foram organizadas de acordo com os passos do KDD, trazendo os principais métodos utilizados.

2.2 PRÉ-PROCESSAMENTO

Grande parte dos métodos utilizados para a extração de conhecimento das bases, como, por exemplo, *machine learning*, induz conhecimentos estritamente a partir dos dados, sendo assim, a qualidade do conhecimento extraído é determinada pela qualidade dos dados fornecidos (Batista, 2003). Desse modo, no processo de KDD, a fase que tem como finalidade melhorar a qualidade dos dados, identificando e removendo problemas antes que os métodos de extração de conhecimento sejam aplicados, é denominada pré-processamento.

Essa fase pode assumir diversos objetivos, a depender dos dados coletados, como identificação de atributos irrelevantes, tratamento de dados corrompidos e variáveis faltantes, além da alteração da estrutura dos dados, como o grau de granularidade (Batista, 2003). Nas seções seguintes, é apresentada uma descrição geral dos principais métodos utilizados neste trabalho para esta etapa: técnica de imputação **de dados; técnica de transformação de dados; e técnica de redução de dimensionalidade.**

2.2.1 Técnica de imputação de dados

Um problema comum e recorrente em investigações científicas é a ocorrência de dados faltantes (*missing data*), especialmente na área da Saúde e das Ciências Sociais (Rubin, 1996). De acordo com Little e Rubin (2019), existem três tipos de dados faltantes: *Missing completely at random* (MCAR), *Missing at random* (MAR) e *Missing not at random* (MNAR).

O processo de preencher dados faltantes por valores prováveis e, assim, minimizar os impactos nas análises, é chamado de imputação de dados faltantes e contém diversos métodos agrupados em dois tipos de imputação: simples e múltipla; sendo que o primeiro gera um único valor para cada dado faltante escolhido, por meio de métodos como média, mediana, estimativa de máxima verossimilhança e até regressão linear (Little; Rubin, 2019), e, no segundo, são gerados ‘n’ valores para cada dado faltante por meio da técnica proposta por Rubin (1978), que leva em conta a incerteza associada à imputação de dados faltantes por meio dos erros-padrão das estimativas, considerando a variabilidade dos resultados (CANIZARES *et al*., 2004; STERNE *et al.*, 2009).

2.2.2 Técnicas de transformação dos dados

Muitos modelos e algoritmos trabalham apenas com valores categóricos, enquanto outros apenas com dados numéricos, e dentro do mesmo tipo de variável ainda podem ser aplicados métodos diferentes para a transformação dos dados, dependendo dos objetivos pretendidos. Alguns exemplos são: a normalização, que coloca as variáveis em uma mesma escala; a generalização, para a conversão de valores muito específicos para valores mais genéricos; o agrupamento, para agrupar valores em faixas sumarizadas; e a criação de novos atributos gerados a partir de outros já existentes, entre outros (CAMILO; SILVA, 2009).

Nesta seção são apresentadas duas formas de transformação: a transformação de *dummies* variáveis qualitativas; e a normalização, aplicada em variáveis quantitativas.

2.2.2.1 Transformação de variáveis qualitativas (dummies)

As variáveis qualitativas ou categóricas são aquelas que representam características de um indivíduo, objeto ou elemento, e que não podem ser medidas ou quantificadas. Segundo Fávero e Belfiore (2017), essas variáveis podem ser classificadas em função do número de categorias da seguinte forma:

a) dicotômicas ou binárias (*dummies*), quando assumem apenas duas categorias;

b) politômicas, quando assumem mais de duas categorias.

Para as variáveis que assumem apenas duas categorias qualitativas (variáveis dicotômicas), atribuímos o valor 1 quando a característica de interesse está presente e o valor 0, caso contrário. Em algumas técnicas, como a regressão logística binária durante o processo de pré-processamento, faz-se necessária a transformação das variáveis qualitativas com *n* categorias em variáveis *dummies* de (*n-1*) categorias (FÁVERO, BELFIORE, 2017).

2.2.2.2 Normalização de variáveis quantitativas

As variáveis quantitativas são variáveis cujas características podem ser calculadas em uma escala de valores numéricos, podendo ser contínuas ou discretas (Bussab; Morettin, 2017). Para evitar que o modelo fique enviesado na direção das variáveis com maior ordem de grandeza, pode-se utilizar a padronização, que emprega variáveis em uma mesma ordem de grandeza. subtraindo cada valor pela média da população e, em seguida, dividindo a diferença pelo desvio padrão da população, c**onforme a Equação 1:**

$$Z = \frac{X - \mu}{\sigma} (1)$$

em que:

X representa o valor de cada variável;
Z representa o valor padronizado da variável X;
μ é a média populacional;
σ é o desvio padrão populacional.

2.2.3 Redução de dimensionalidade

Devido à possibilidade de muitas variáveis independentes utilizadas em uma regressão apresentarem um elevado grau de correlação por medirem as mesmas características, ou seja, de serem redundantes, a redução de dimensionalidade apresenta-se como uma das etapas mais importantes para a análise multivariada (Maitra; Yan, 2008). Como descrito por Miloca e Conejo (2008), uma variável adicional pode influenciar na previsão da variável dependente não somente com a correlação existente com a variável dependente, mas também com a correlação dessa variável com as demais variáveis independentes existentes no modelo. Em situações em que essas dependências forem fortes, caracteriza-se a multicolinearidade.

Conforme Hair Junior (2009), além dos efeitos na explicação, a multicolinearidade pode ter sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado e, nesses casos, é aconselhado selecionar variáveis que sejam independentes, ou seja, com baixa multicolinearidade com as outras variáveis independentes, mas que também apresentem correlações elevadas com a variável dependente.

Logo, o problema de multicolinearidade pode ser contornado utilizando-se a Análise de Componentes Principais (Aranha; Zambaldi, 2008), sendo uma técnica estatística de análise multivariada que, dentre muitas aplicações, pode ser empregada na redução da dimensionalidade.

2.2.3.1 Análise de Componentes Principais

É possível que o modelo seja bom para tomar a decisão sobre dados históricos, mas que seja ruim para tomar decisões sobre dados nunca vistos. Nesse caso, significa que ocorreu *overfitting* e que o modelo não generaliza para novos dados. Segundo Monard e Baranauskas (2003), o *overfitting* é caracterizado pelo ajuste em excesso, resultando na impossibilidade de aplicar o modelo em novos dados, pois seu aprendizado tornou-se específico demais para os casos do treinamento. Uma das causas relacionadas ao *overfitting* está na quantidade elevada de parâmetros presentes na base (alta dimensionalidade).

Dentre as diversas técnicas de redução de dimensionalidade, destaca-se o *Principal Component Analisys* (PCA) (Smith, 2002), que foi inicialmente publicada por Pearson (1901) e descrita em métodos computacionais práticos (HONGYU; SANDANIELO; OLIVEIRA JUNIOR, 2016).

De acordo com Mingoti (2005), PCA é uma técnica estatística multivariada que pode ser empregada sob dois aspectos: o primeiro com o objetivo da redução do espaço paramétrico, ou seja, de obter combinações lineares das variáveis originais que descrevam características do fenômeno estudado sem perda prejudicial de informação; o segundo é sobre a utilização dessas combinações para a análise classificatória ou discriminante dos dados.

Hongyu, Sandanielo e Oliveira Junior (2016) complementam que a PCA é uma técnica multivariada de modelagem da estrutura de covariância, que consiste na transformação linear de um conjunto de variáveis originais, inicialmente correlacionadas entre si, em outro conjunto menor de variáveis não correlacionadas, denominadas componentes principais, que contêm a maior parte da informação (em termos de variação) do conjunto original.

Assim, uma das principais vantagens na aplicação da PCA é a eficácia em retirar a multicolinearidade das variáveis (Hongyu; Sandanielo; Oliveira Junior, 2016), avaliando as interrelações entre um grande número de variáveis correlacionadas, com o objetivo de condensar as informações contidas em um número de variáveis originais em um número menor de variáveis não correlacionadas, sem perda de informações (MINGOTI, 2005; HAIR JUNIOR et al., 2009).

Contudo, Aranha e Zambaldi (2008) relatam que um dos problemas relacionados à utilização da PCA está na dificuldade de interpretar o significado dos componentes. Outra desvantagem está na sua aplicação em casos de variáveis originais pouco correlacionadas, gerando pouca redução na dimensionalidade, portanto baixa eficácia na eliminação da multicolinearidade (HONGYU, 2015; REGAZZI, 2000).

2.2.4 Mineração de Dados

A mineração de dados surgiu na década de 1990 para se referir à etapa de geração de informação (conhecimento) do KDD (*Knowledge Discovery in Database*), e é um campo de pesquisa que engloba diversas áreas, tais como: inteligência artificial, redes neurais, bancos de dados, *machine learning* e estatística (FAYYAD; PIATETSKY; SMYTH, 1996).

Segundo Larose (2005), os objetivos mais comuns da mineração de dados são: classificação, utilizada na identificação de classes de novos dados por meio do aprendizado realizado em dados já pré-classificados; regressão ou estimação, que busca estimar o valor de determinada variável analisando-se os valores das demais; agrupamentos, que visa identificar

e aproximar os registros similares; e associação, em que a tarefa busca identificar quais atributos estão relacionados entre outros.

Os métodos de aprendizado podem ser divididos em aprendizado supervisionado (preditivo) e não supervisionado (descritivo) (Han; Kamber, 2006), existindo ainda uma proposta de abordagens semissupervisionadas (Seliya, 2007). A principal diferença está na presença ou não de uma base rotulada para aprendizado do modelo.

Quando é necessário um conjunto de dados com a variável alvo rotulada e os novos registros são categorizados em relação a esta, temos uma abordagem supervisionada. Exemplos dessa categoria são a classificação (que também pode ser não supervisionado) e a regressão. Já em tarefas que não têm pré-categorização geralmente se utilizam medidas de similaridades entres os atributos; essas tarefas são denominadas não supervisionadas. São as mais comuns entre as não supervisionadas as tarefas de agrupamento e de associação (Mccue, 2007). Neste estudo, são apresentados os métodos de regressão logística e árvore de decisão.

2.2.4.1 Regressão logística

Uma das técnicas bastante utilizada na mineração de dados é a regressão logística, que tem como objetivo estudar a relação entre o comportamento de uma ou mais variáveis explicativas (preditoras ou independentes) qualitativas e um determinado fenômeno, representado por uma variável dependente, por meio de um modelo (equação matemática) (Fávero; Belfiore, 2017).

A regressão logística segue o modelo de aprendizagem supervisionada. A divisão dessa tarefa em duas etapas (treinamento e teste) possibilita saber se a capacidade do modelo de responder corretamente aos dados é eficiente. Na primeira etapa é realizado o treinamento, no qual o modelo é gerado por meio de um conjunto de dados já classificados e, em seguida, um novo conjunto de dados, que não foram usados na etapa anterior, é utilizado para realizar a etapa de testes (CASTRO E FERRARI, 2016).

A classificação da regressão logística ocorre de acordo com a quantidade de possibilidades de respostas (categorias), assumindo o valor 1 para descrever a ocorrência do evento de interesse e 0 para descrever a não ocorrência do evento de interesse, ou multinominal quando oferece três possibilidades ou mais de resposta (Fávero *et al*, 2009). No caso de variáveis multinominais, é necessário definir a categoria de referência, a fim de compará-la com as demais e para que essta escolha não seja arbitrária, conforme desejo e

orientação do pesquisador (Fávero *et al.*, 2009). Ademais, para uma variável qualitativa com (*n*) categorias são necessárias (*n* - 1) variáveis *dummies* (FÁVERO, BELFIORE, 2017).

Outro fator de diferenciação das técnicas de regressão logística binária e multinomial em relação às técnicas tradicionais é a base pela qual são elaboradas. As técnicas tradicionais são elaboradas com base na estimação do modelo por meio do método de mínimos quadrados ordinários, ao passo que as técnicas de regressão binária e multinomial são elaboradas com base na estimação do modelo por máxima verossimilhança (FÁVERO; BELFIORE, 2017).

2.2.4.1.1 Modelo de regressão logística binária

O modelo de regressão logística binária quantifica a probabilidade de ocorrência de um evento de interesse com base no comportamento das variáveis explicativas. Dessa forma, o modelo logístico pode ser expresso de acordo com a Equação 2 (FÁVERO; BELFIORE, 2017):

$$L_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} \quad (2)$$

em que:

$L_i = \ln\left(\frac{p_i}{1-p_i}\right)$ é denominado por logito, sendo que i representa cada observação da amostra e p_i é a probabilidade de ocorrência do evento de interesse (para $i = 1,2,\ldots,n$, e n é o tamanho da amostra);

$X_1, X_2, \ldots, X_k$ é o conjunto de variáveis independentes (ou explicativas);

α é o parâmetro a ser estimado;

β_j são os parâmetros a serem estimados ($j = 1,2,\ldots,k$).

Além disso, o termo $\frac{p_i}{1-p_i}$ representa a chance de ocorrência do evento de interesse e é denominada razão de chance (*odds ratio*) (Fávero; Belfiore, 2017), ou seja, é a probabilidade de sucesso comparada com a de fracasso (DIAS FILHO; CORRAR, 2007).

Em termos práticos, por meio da regressão logística binária, é possível estimar a probabilidade de ocorrência do evento para cada observação e não os valores previstos da variável dependente (FÁVERO; BELFIORE, 2017).

Conforme Dias Filho e Corrar (2007), a probabilidade associada à ocorrência do evento de interesse é obtida por meio da Equações 3 e 4.

$$p_i = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n}}{1 + e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}} . (3)$$

Simplificando, obtém-se:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}, (4)$$

isto é, p_i pode ser entendida como a probabilidade de a variável dependente ser igual a 1, dado o comportamento das variáveis independentes.

2.2.5 Árvores de decisão

Árvores de decisão são algoritmos classificadores que funcionam como um fluxograma em forma de árvore, em que cada nó indica um teste feito sobre um valor, as ligações entre os nós representam os valores possíveis do teste do nó superior e as folhas indicam a classe (categoria) à qual o registro pertence. Por fim, após o modelo pronto para classificar um novo registro, basta seguir o fluxo na árvore, começando no nó raiz até chegar a uma folha. Devido à sua estrutura, as árvores de decisão também podem ser convertidas em Regras de Classificação (CAMILO; SILVA, 2009).

2.2.5.1 Light boosting gradient model

O *Light Boosting Gradient Model* (LGBM) é um algoritmo baseado em árvores de decisão que aplica um *boosting* gradiente, com alta velocidade e eficiência no treinamento do modelo, baixo uso de memória, melhor acurácia, suporte para aprendizado paralelo; e capaz de lidar com grandes escalas de dados (MICROSOFT, 2019a).

O *boosting* presente no LGBM consiste em utilizar várias cópias do classificador para realizar o aprendizado de forma interativa e, para classificar um resultado, é dado um peso para cada predição desses classificadores e, então, é calculada a predição final.

Para o aprendizado, cada dupla de treinamento, formada por um atributo e sua classe, tem um peso. Assim que o primeiro classificador executa o processo de aprendizado, os pesos são alterados de modo a aumentar a atenção para as duplas classificadas erroneamente por esse classificador. Assim, o próximo terá maiores chances de acertar que o primeiro classificador, e assim sucessivamente. O último classificador impulsionado (*boosted*) combina os votos de cada classificador individualmente, sendo que os pesos de cada classificador é

uma função de sua acurácia (HAN; KAMBER, 2012).

Quando voltado para classificar múltiplas classes, o LGBM pode utilizar algumas métricas, e uma destas é por meio da função objetivo *softmax* (MICROSOFT, 2019b). Essa função retorna um vetor que representa a distribuição de probabilidades de uma lista de potenciais resultados.

2.3 PÓS-PROCESSAMENTO

A etapa seguinte à mineração de dados é o pós-processamento e o objetivo, nessa etapa, é avaliar a utilidade do conhecimento descoberto. Segundo Goldschmidt e Passos (2005), a complexidade do processo de KDD está na dificuldade de perceber e interpretar corretamente os inúmeros fatos observados. Nesta seção são abordados os métodos utilizados na análise dos dados minerados.

2.3.1 Indicadores de Desempenho

Diversos indicadores de desempenho podem ser utilizados, a fim de se estimar e comparar os resultados de diferentes modelos, comparando-se a saída predita com a saída esperada. Um dos mais utilizados para problemas de classificação é a matriz de confusão, detalhada na próxima seção. Todas as métricas apresentadas comportam-se de maneira diferente quando utilizadas em conjuntos de dados balanceados ou desbalanceados (SAITO; REHMSMEIER, 2015).

2.3.1.1 Matriz de confusão

No campo do *machine learning,* a matriz de confusão é uma tabela que permite a visualização do desempenho de um modelo de classificação para problemas binários. Ela é formada por duas linhas e duas colunas que mostram o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

A divisão das previsões em uma matriz de confusão binária pode ser observada na Figura 3.

Figura 3 – Matriz de Confusão para Classificação Binária

		Classe Verdadeira: Positivo	Classe Verdadeira: Negativo
Classe Prevista	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: Elaborado pelo autor (2022)

2.3.1.2 Curva ROC

A curva ROC (do inglês, significa *Receiver Operator Characteristic*) é uma das formas de avaliar o diagnóstico dos modelos estatísticos obtidos por meio de classificadores binários, sendo utilizada amplamente em campos como *machine learning* (Spackman, 1989) e, por isso, foi escolhida como forma de comparação dos modelos aplicados neste trabalho.

Trata-se de uma representação gráfica, que resume as informações das funções de distribuição cumulativas das pontuações das duas classes consideradas, constituída no eixo das coordenadas pela fração de verdadeiros positivos (sensibilidade) e pela fração de falsos positivos (equivalente a 1 menos a especificidade), no eixo das abcissas (Mourão, 2015). Assim, sensibilidade é a proporção de positivos que foram identificados corretamente, enquanto a especificidade é dada pela proporção de negativos que foram identificados corretamente (CARTER et al., 2016).

2.3.1.3 Acurácia

De acordo com Paulino (2005), a acurácia apresentada na Equação (5), é a proporção dos verdadeiros positivos e negativos com relação ao total de previsões realizadas. Entende-se como verdadeiros positivos e verdadeiros negativos a classificação feita corretamente pelo modelo da classe Positivo e Negativo, respectivamente.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (5)$$

em que a A é a acurácia, VP é o total de verdadeiros positivos, VN é o total de verdadeiros negativos, FP é o total de falsos positivos e FN é o total de falsos negativos (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006).

3 PROCEDIMENTOS METODOLÓGICOS

3.1 FERRAMENTAS E BASE DE DADOS

3.1.1 Software

O *software* utilizado tanto no tratamento da base, quanto na análise descritiva, como na construção do modelo de regressão logística, foi a plataforma Anaconda®. Esta é uma ferramenta para ciência de dados em *Python,* utilizada para gerenciar pacotes e ambiente, com uma coleção de mais de 7.500 pacotes de código aberto (Bandyopadhyay, 2019). Os principais motivos para a escolha do programa, além de ser gratuito, estão na quantidade de pacotes disponíveis, úteis para esta pesquisa, e nas diversas ferramentas que facilitam a interação com a linguagem de programação.

3.1.2 Bibliotheca Scikit-Learn

Para a execução das atividades do KDD, tanto na etapa de pré-processamento, quanto para as etapas de mineração e pós-processamento, foram utilizadas as técnicas e métodos da biblioteca do *Scikit-learn*.

Conforme publicado no site oficial, o *Scikit-learn* é uma biblioteca da linguagem *Python,* desenvolvida originalmente por David Cournapeau, no *Google Summer of Code*, em 2007, especificamente para aplicação de *machine learning*. Ainda em 2007, Matthieu Brucher começou a trabalhar nesse projeto como parte de sua tese e, em 2010, Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort e Vincent Michel do INRIA assumiram a liderança do projeto e fizeram o primeiro lançamento público, no dia 1 de fevereiro de 2010 (SCKITLEARN, 2011).

Essa biblioteca tem código aberto e é projetada para interagir com as bibliotecas do *Python,* dispondo de ferramentas simples e eficientes para análise preditiva de dados, sendo reutilizável em diferentes situações. Além disso, inclui vários algoritmos de classificação, regressão e agrupamento, máquinas de vetores de suporte, florestas aleatórias, *k-means*, entre outros (SCKITLEARN, 2011).

O *Scikit-learn* está organizado em vários módulos, cada um desenvolvido para uma finalidade específica, descritos a seguir (SCKITLEARN, 2011):

i. Pré-processamento: é o processo que busca converter dados brutos em dados preparados, normalmente esta é a etapa mais trabalhosa no desenvolvimento de um modelo de *machine learning;*

ii. classificação: refere-se ao desenvolvimento de modelos capazes de detectar a qual categoria pré-determinada um elemento pertence;

iii. regressão: tem como objetivo o desenvolvimento de modelos que podem atribuir um valor contínuo a um elemento;

iv. clusterização: desenvolvimento de modelos para detecção automática de grupos com características similares em seus integrantes;

v. redução de dimensionalidade: busca reduzir o número de variáveis em um problema;com essa redução, diminui-se a quantidade de cálculos em um modelo, aumentando a eficiência, com uma perde mínima de assertividade;

vi. ajuste de parâmetros: esse módulo compara, valida e escolhe parâmetros e modelos, de maneira automatizada, com o objetivo de comparar parâmetros para ajuste de um modelo, encontrando, assim, a melhor configuração para a aplicação em questão.

3.1.3 Construção da base de dados

Para este estudo, foram utilizados dados históricos dos processos trabalhistas brasileiros, publicados nos acervos de jurisprudência (Tribunais da Justiça Trabalhista e Diários da Justiça do Brasil), coletados e processados pela empresa de tecnologia jurídica Data Lawyer®.

A Data Lawyer® foi fundada pelo CEO Caio Santos, em 5 de dezembro de 2018, sendo considerada a primeira *Legal Research* do Brasil em nível de documentos, com produtos de jurimetria e gestão de escritórios que têm como premissa uma advocacia orientada a dados (SANTOS, 2019). Seu acervo de *big data* jurídico é formado pelo cruzamento de dados extraídos de cerca de 11 milhões de processos trabalhistas cadastrados, de petições e de contratos, centralizados por meio de processamentos de dados e de inteligência artificial (SANTOS, 2019).

Para a obtenção da base de dados desta pesquisa, foi utilizado um recorte da base disponibilizada pela Data Lawyer® e a Tabela Processual Unificada de Assuntos da Justiça do Trabalho (TRIBUNAL SUPERIOR DO TRABALHO, 2021).

Tal recorte foi referente aos litigantes do setor financeiro brasileiro envolvidos em processos trabalhistas entre os anos de 2017 e 2018, resultando em 2622 processos, contendo

resultados das sentenças (variável dependente neste estudo), juízes envolvidos, Tribunal Regional, assuntos dos processos, entre outros (variáveis independentes).

A Tabela Processual Unificada do Poder Judiciário foi utilizada na formação da base de trabalho, com o objetivo de padronizar e de uniformizar a taxonomia e terminologia de classes, ou seja, assuntos processuais, trazendo agrupamentos oficiais utilizados pela Justiça Estadual, Federal, do Trabalho e do Superior Tribunal de Justiça, empregados nos seus respectivos sistemas processuais.

3.2 PRÉ-PROCESSAMENTO

Para possibilitar a aplicação dos modelos e aumentar a confiabilidade dos resultados, foi realizado o pré-processamento dos dados (Hair Junior et al., 2009), em cinco tarefas:

i. tratamento dos dados (agrupamentos e eliminação de variáveis);
ii. imputação de dados faltantes;
iii. normalização (para as variáveis numéricas);
iv. transformação de variáveis qualitativas em variáveis dicotômicas *dummies* (aplicada às variáveis categóricas);
v. redução de dimensionalidade e multicolinearidade.

3.2.1Tratamento dos dados

Na Tabela 1 é apresentada a volumetria original dos desfechos (variável dependente) antes de serem realizados os agrupamentos e a eliminação de variáveis.

Tabela 1 – Frequência absoluta por desfecho da base original de dados

Desfecho do processo	Frequência absoluta
Parcialmente Procedente	950
Acordo	656
Improcedente	183
Arquivamento Ausência do Reclamante	137
Desistência	96
Extinção da execução ou do cumprimento da sentença	78
Ausência de Pressupostos Processuais	40
Indeferimento da Petição Inicial	23
Procedente	19
Arquivamento Sumaríssimo (art. 852-B, § 1º/CLT)	14
Pronúncia de Decadência ou Prescrição	11
Abandono da Causa	10
Perempção, litispendência ou coisa julgada	8
Ausência das Condições da Ação	7
Outros	6
Total	2.238

Fonte: Produção do próprio autor(2022)

Primeiramente, foram retirados das bases de dados todos os casos que ainda não tinham desfecho, ou seja, não havia sido proferida a sentença final ou acordo, uma vez que esta é a variável resposta do modelo preditivo.

Em seguida, foi realizado o ajuste na variável dependente, classificando como ”Procedentes” todos os resultados apontados na base de dados como: ”Parcialmente Procedente”, ”Acordo” e ”Procedente”, resultando em 1.625 (73% da base de dados) casos ”Procedentes”, e o restante como ”Improcedentes”, gerando um total de 613 (27% da base de dados) casos ”Improcedentes”.

3.2.1.1 Criando uma linha para cada assunto da petição

O reclamante pode ingressar, perante o Judiciário, com mais de um pedido, ou melhor, pode requerer mais de um direito dentro de um mesmo processo. Cada direito requerido é chamado na jurisprudência de objeto de ação. A base de dados original, vinda da Data Lawyer®, separa cada processo em uma linha e todas as informações contidas nele estão nas respectivas colunas.

Os objetos da ação, podendo ter mais de um por processo, são nomeadas na base como “assuntos” e inicialmente encontram-se dentro de uma mesma célula. Para a realização deste estudo, foi necessário separar cada objeto em linhas diferentes replicando apenas as informações comuns do processo em cada nova linha.

3.2.1.2 Agrupando classes da petição

A classe processual refere-se ao procedimento adotado na esfera judicial ou administrativa para atender ao pedido e está relacionado diretamente à matéria ou aos temas discutidos nos processos. Os tipos de classes podem ser encontrados, de forma detalhada, na Tabela de Classes Processuais, criada a partir da Resolução CNJ nº. 46/2007, de uso obrigatório no âmbito da Justiça Estadual, Federal, do Trabalho, Eleitoral, Militar da União, Militar dos Estados e Superior Tribunal de Justiça (TRIBUNAL SUPERIOR DO TRABALHO, 2021).

Com o objetivo de melhorar o desempenho e simplificar os resultados do processo de modelagem por meio da redução do espaço de busca de hipóteses (Wang; Xiuju, 2005), foram criados três grupos para as classes de petição, sendo dois já existentes: “Ação Trabalhista - Rito Ordinário” e “Ação Trabalhista - Rito Sumaríssimo” – e o restante foi agrupado em “Outros”, como demonstrado nas tabelas 2 e 3.

Tabela 2 – Frequência de classes processuais antes do processamento

Classes Processuais	Frequência absoluta
Ação Trabalhista - Rito Ordinário	1.855
Ação Trabalhista - Rito Sumaríssimo	284
Execução Provisória em Autos Suplementares	67
Ação Civil Pública	6
Ação Civil Coletiva	6
Ação Trabalhista - Rito Sumário (Alçada)	4
Protesto	3
Petição Cível	2
Ação Rescisória	2
Alvará Judicial - Lei 6858/80	1
Oposição	1
Tutela Cautelar Antecedente	1
Embargos de Terceiro	1
Exibição	1
Ação de Cumprimento	1
Notificação	1
Alvará Judicial	1

Fonte: Elaborado pelo autor (2022)

Tabela 3 – Frequência de classes processuais após o tratamento

Classes Processuais	Frequência absoluta
Ação Trabalhista - Rito Ordinário	1.855
Ação Trabalhista - Rito Sumaríssimo	284
Outros	99

Fonte: Elaborado pelo autor (2022)

3.2.1.3 Agrupamento de comarcas

Comarca é o termo originalmente empregado para definir um território delimitado e, dentro do contexto jurídico, corresponde ao território em que o juiz de primeiro grau irá exercer sua jurisdição, podendo abranger um ou mais municípios (TRIBUNAL E JUSTIÇA DO TRABALHO DO AMAPÁ, 2016).

Devido à quantidade de comarcas *versus* o número de processos utilizados neste estudo, optou-se por criar uma variável chamada “is_capital” a partir do agrupamento das comarcas, classificando como 1 as comarcas que são capitais brasileiras e como 0 as comarcas que não são capitais brasileiras. Todavia, para garantir nomes de categoria consistentes, antes de criar a classificação, foi necessário remover variações ortográficas, retirando-se as letras

maiúsculas e os acentos, pois esses elementos não estavam padronizados nos nomes das comarcas.

3.2.1.4 Retirando variáveis

As variáveis “Número do processo”, “CNPJs”, “Partes pólo passivo”, “Partes pólo ativo”, “Advogados pólo ativo”, “Advogados pólo passivo” foram eliminadas da base de dados por não terem correlação com a variável dependente, uma vez que tinham um número de valores distintos na mesma ordem de grandeza dos números de processos analisados.

As variáveis que são obtidas após o encerramento do caso, tais como “Instância”, “Fase”, “Tipos de recursos”, “Valor de depósitos (R$)”, “Valor de condenação (R$)”, “Valor de liquidação (R$)”, “Valor de custas (R$)”, “Valor de acordo (R$)”, “Juízes”, ou seja, que não se sabe o resultado antes do desfecho da ação, também foram eliminadas da base.

Ademais, foram retirados valores que são constantes na base, por não influenciarem nos resultados.

3.2.2 Imputação dos dados

Com o objetivo de completar os dados faltantes por estimativas de valores coerentes e, assim, possibilitar a análise desses casos, foi necessário realizar a técnica de imputação de dados faltantes (RUBIN, 1996; SCHAFER; GRAHAM, 2002).

Para isso, dentro do modelo, foi incluída a funcionalidade *SimpleImputer,* presente na biblioteca do *scikit-learn 1.1.1,* que realiza a chamada “imputação única”. Assim, os dados ausentes de cada variável foram preenchidos por um único valor, usando-se métodos separados em dois grupos, dependendo do tipo da variável:

i. substituição dos dados faltantes pela média dos valores (para variáveis numéricas);

ii. substituição pelo valor mais frequente (para variáveis categóricas).

3.2.3 Normalização

O segundo tratamento realizado no modelo foi a normalização das variáveis numéricas, ao trazê-las para uma mesma ordem de grandeza, dimensionando e traduzindo cada

característica individualmente.

Para essa funcionalidade, foi utilizado o módulo *MinMaxScaler* da *biblioteca Sckit-Learn* em *Phyton*. Essa transformação foi usada como uma alternativa para a escala de variação unitária de média zero, e que permite escolher um *range* que não tenha números negativos.

3.2.4 Transformação de variáveis qualitativas em variáveis dicotômicas

A formatação das variáveis independentes (variáveis categóricas) foi realizada com o intuito de representar as características e incluir essas informações não numéricas na regressão. O módulo utilizado nessa transformação foi o OneHotEncorder que, em resumo, transforma as variáveis qualitativas em *dummies*, criando uma coluna binária para cada categoria (SCKITLEARN, 2011).

3.2.5 Redução de dimensionalidade

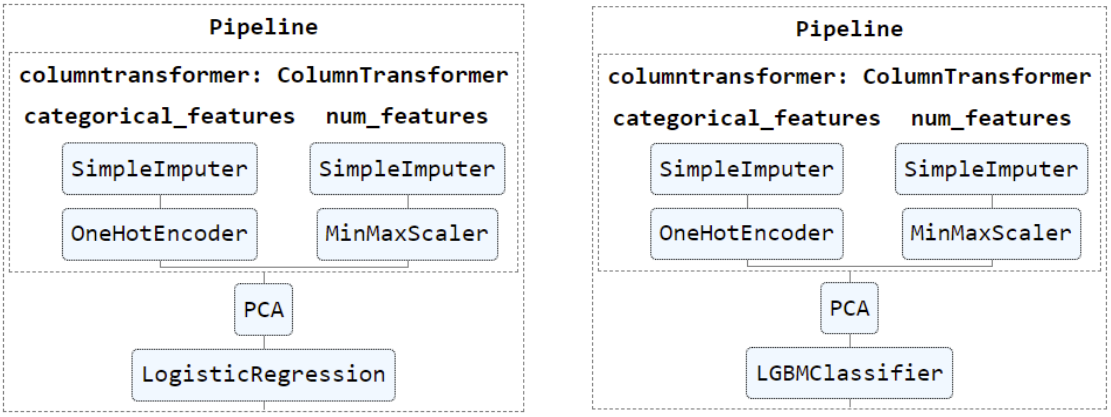
Um ajuste antes da aplicação das modelagens (Regressão Logística e LBGM) é a redução de dimensionalidade por meio da PCA, utilizada nas metodologias. Essa transformação busca eliminar o problema de multicolinearidade, causado por variáveis redundantes ou altamente correlacionadas, por meio da criação de componentes principais não correlacionadas entre si, formadas pela combinação linear de todas as variáveis originais (Hongyu, 2015; Regazzi, 2000), diminuindo, também, o número de parâmetros no modelo, mitigando um possível *overfitting*.

Para este fim, foi utilizado a *feature sklearn decomposition* PCA por meio do pacote *Sckit-Learn* em *Phyton* (SCKITLEARN, 2011).

3.3 MINERAÇÃO DOS DADOS

Na etapa de modelagem, foram utilizados dois tipos de modelos para fins de comparação: a Regressão Logística e o LGBM. As Figuras 4 (a) e (b) representam os passos utilizados no *Phyton* para aplicação dos modelos de Regressão Logística e do LBGM.

Figura 4 – Representação dos passos para aplicação dos algoritmos de regressão logística e do LBBM respectivamente em Phyton



(a) (b)

Fonte: Elaborado pelo autor (2022)

3.3.1 Amostras de Treino e Teste

Para a construção do modelo e mitigação do *Overfitting*, ou seja, para garantir a eficácia da previsão ao utilizar novos dados, realizou-se a divisão da base em duas partes denominadas conjunto de treino e conjunto de teste, caracterizando o processo de treino e teste.

Para esta pesquisa, foi escolhido separar 85% da base para "treino", isto é, para a modelagem, e 15% para o teste dos modelos criados, utilizando a data de abertura dos processos judiciais como critério de corte, ou seja, foi encontrada a data cujas quantidades de processos abertos anteriores a ela correspondem 85% da base. Essa seleção resultou em 1862 processos destinados para o treino do modelo e 327 para o teste dos dois modelos criados.

3.3.2 Ajuste dos modelos

Para fins de comparação, foram utilizados em ambos os modelos (LGBM e Regressão Logística), o mesmo pré-processamento.

Para o modelo de Regressão Logística, foi aplicada a função *LogisticRegression* por meio do pacote *Sckit-Learn* (Sckitlearn, 2011), enquanto para o modelo LGBM foi utilizada a biblioteca *lightgbm* da Microsoft para importar a *feature LGBMClassifier* (MICROSOFT, 2019a).

Foram utilizados os parâmetros padrões das bibliotecas da *Sckit-Lean* e *lightgbm,* nos dois modelos, o quais estão descritos na Tabela 4.

Tabela 4 – Parâmetros padrões das feature Logística Regression e featura LGBM classifier

feature LogisticRegression		feature LGBMClassifier	
Parâmetro	Valor padrão	Parâmetro	Valor padrão
penalty	'l2'	boosting_type	'gbdt'
dual	False	num_leaves	31
tol	0.0001	max_depth	-1
C	1.0	learning_rate	0.1
fit_intercept	True	n_estimators	100
intercept_scaling	1	subsample_for_bin	200000
class_weight	None	objective	None
random_state	None	class_weight	None
solver	'lbfgs'	min_split_gain	0.0
max_iter	100	min_child_weight	0.001
multi_class	'auto'	min_child_samples	20
verbose	0	subsample	1.0
warm_start	False	subsample_freq	0
n_jobs	None	colsample_bytree	1.0
l1_ratio	None	reg_alpha	0.0
		reg_lambda	0.0
		random_state	None
		n_jobs	None
		importance_type	'split'

Fonte: Elaborado pelo autor (2022)

4 RESULTADOS

4.1 SELECÃO DE VARIÁVEIS

Foram obtidos como resultados do PCA uma redução de 83% das variáveis do modelo, sendo gerados 27 componentes principais, reduzindo os efeitos da multicolinearidade nas estimativas dos coeficientes de regressão.

4.2 MODELOS AJUSTADOS

4.2.1 Modelo de Regressão Logística

No modelo de Regressão Logística, os valores dos parâmetros estimados para as variáveis independentes representam medidas das variações na proporção das probabilidades, que é denominada de logaritmo da razão de desigualdades (*logit*), ou seja, a razão entre as probabilidades de dois resultados ou eventos (HAIR JUNIOR et al., 2009).

Fávero et al. (2009) reforçam que o objetivo de estimar os parâmetros é encontrar uma função de tal modo que os coeficientes das variáveis independentes permitam estabelecer a importância de cada variável para a ocorrência do evento de interesse, bem como calcular a probabilidade de ocorrência desse evento.

A Tabela 5 mostra os parâmetros estimados pela modelagem por meio da Regressão Logística, considerando a probabilidade de ocorrência do evento de referência ganho da causa trabalhista (assume valor 1). Portanto, os coeficientes positivos indicam que o evento torna-se mais provável e os coeficientes negativos indicam que o evento torna-se menos provável.

Tabela 5 – Parâmetros estimados pela regressão logística

Parâmetro estimado	Coeficiente
$\hat{\alpha}$	0,888114
$\hat{\beta}_1$	0.057608
$\hat{\beta}_2$	-0.787865
$\hat{\beta}_3$	-0.291918
$\hat{\beta}_4$	0.907388
$\hat{\beta}_5$	-0.575483
$\hat{\beta}_6$	-1.682.911
$\hat{\beta}_7$	1.939.488
$\hat{\beta}_8$	1.097.315
$\hat{\beta}_9$	0.268827
$\hat{\beta}_{10}$	0.519076
$\hat{\beta}_{11}$	0.901719
$\hat{\beta}_{12}$	-0.007804
$\hat{\beta}_{13}$	1.110.903
$\hat{\beta}_{14}$	-0.181131
$\hat{\beta}_{15}$	-0.491000
$\hat{\beta}_{16}$	0.830696
$\hat{\beta}_{17}$	-0.934598
$\hat{\beta}_{18}$	-0.381790
$\hat{\beta}_{19}$	-0.198124
$\hat{\beta}_{20}$	0.219499
$\hat{\beta}_{21}$	0.744209
$\hat{\beta}_{22}$	-0.629275
$\hat{\beta}_{23}$	0.086818
$\hat{\beta}_{24}$	0.535105
$\hat{\beta}_{25}$	0.489071
$\hat{\beta}_{26}$	-1.138.076

Fonte: Elaborado pelo autor (2022)

4.2.2 Modelo de LGBM

Para a modelagem por LGBM, por se tratar de um modelo formado por um conjunto de árvores de decisão, o resultado da sua aplicação não pode ser expresso em formato de equação e seus respectivos coeficientes. Foram avaliados os resultados da *performance* do modelo visto na Seção 4.3.

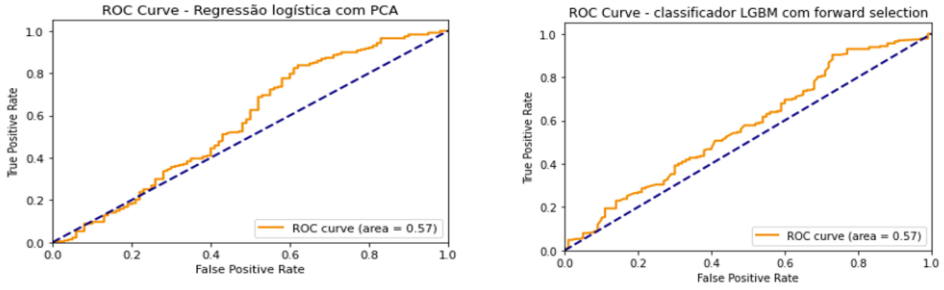
4.3 ANÁLISE DO DESEMPENHO DOS MODELOS

4.3.1 Curva ROC

Para avaliar os resultados dos modelos e compará-los, foi empregada a curva ROC, apresentada na figura 5, para os modelos de Regressão Logística e LGBM..

Fávero et al. (2009) postulam que quanto maior a área abaixo da curva ROC, maior é a capacidade do modelo discriminar os eventos de interesse, quer dizer, quanto mais próxima a curva ROC estiver da reta diagonal, pior é o poder discriminatório do modelo. Além disso, de acordo com Hosmer Junior, Lemeshow e Sturdivant (2013), uma referência para interpretação em relação à área abaixo da curva ROC é: inferior a 0,7 o modelo é incapaz de classificar variáveis; entre 0,7 e 0,8 é aceitável; entre 0,8 e 0,9 é excelente; e superior a 0,9 é fora de série (mas, extremamente rara).

Figura 5 – Curvas ROC para modelos de Regressão Logística e LGBM



Fonte: Elaborado pelo autor (2022)

4.3.2 Matriz de Confusão, Acurácia, Precisão e Sensibilidade

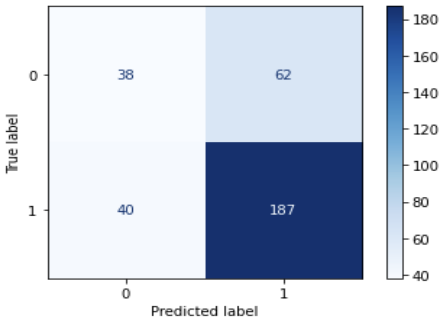
As informações apresentadas nas linhas da Matriz de Confusão, definida na página 22, representam as instâncias de uma classe predita, enquanto cada coluna representa a instância da classe observada, resultando em quatro combinações diferentes de valores reais e previstos.

Nesse contexto, o verdadeiro positivo (VP) ocorre quando, no conjunto real, a classe que se está buscando foi prevista corretamente, ou seja, o resultado do desfecho do caso é verdadeiro (valores iguais a 1) e o modelo classificou como verdadeiro (desfechos tidos como

procedentes). Já o verdadeiro negativo (VN) ocorre quando, no conjunto observado, a classe que não se está buscando prever (valores iguais a 0) foi prevista corretamente pelo modelo, isto é, o resultado do desfecho do caso é tido como improcedente.

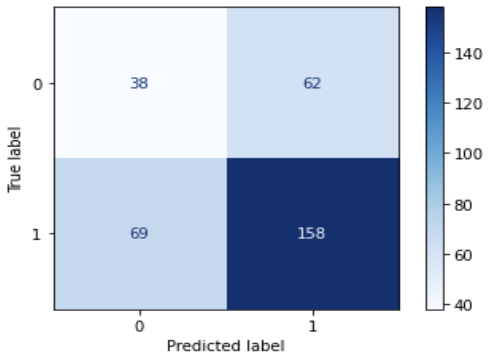
Assim, para apurar a acurácia e o poder preditivo do modelo, foi utilizada a matriz de confusão para os dois ajustes de modelos na base de testes (Figuras 6 e 7). Os valores de VP foram de 187 (82,4%) e 158 (69,6%) para os modelos de Regressão Logística e LGBM, respectivamente, e os valores de VN foram iguais a 38 (38%) para os dois modelos.

Figura 6 – Matriz Confusão para o modelo de Regressão Logística com PCA



Fonte: Elaborado pelo autor (2022)

Figura 7 – Matriz Confusão para o modelo de LGBM com PCA



Fonte: Elaborado pelo autor (2022)

A Tabela 6 mostra os valores de acurácia dos modelos e suas respectivas precisão e sensibilidade calculadas a partir das matrizes de confusão.

Tabela 6 – Comparação da performance dos modelos

Modelo	Acurácia	Precisão	Sensibilidade
Regressão Logística com PCA	68,8%	75,1%	82,4%
LGBM com PCA	59,9%	71,8%	69,6%

Fonte: Elaborado pelo autor (2022)

5 CONCLUSÃO

5.1 VERIFICAÇÃO DOS OBJETIVOS QUANTO AOS DADOS

Uma das dificuldades da utilização de modelo dinâmico e em tempo adequado, dentro da área da jurimetria, é a demora no desfecho dos casos, podendo delongar em média entre 2 e 3 anos, de acordo com informações fornecidas pelo Conselho Nacional de Justiça (CNJ, 2019), na versão de 2019 do relatório "Justiça em Números". Neste cenário, para dispor de um número relevante de casos com desfecho, optou-se por utilizar bases dos anos de 2017 e 2018. Após o trabalho da base, como a retirada de variáveis redundantes, a criação de agrupamentos para diminuição da dimensionalidade e preenchimento de dados faltantes, obteve-se uma base com 1862 linhas para a construção do modelo e 327 para o teste.

Outro ponto relevante quanto ao ambiente de informações disponíveis dentro da jurimetria é o espalhamento das informações sobre o desfecho dos casos. Não foi encontrada uma base oficial que centralizasse todas as informações dos casos e, por isso, fez-se necessário recorrer à base disponibilizada pela empresa *Data Lawyer,* que faz esse trabalho de cruzamento de dados de diversas bases no contexto judicial. O ponto de atenção é que para dados mais recentes (últimos dois anos) e um período mais extenso (mais de 3 anos) as bases não são gratuitas.

5.2 VERIFICAÇÃO DOS OBJETIVOS QUANTO À *PERFORMANCE* DO MODELO

Quanto à *performance* geral dos modelos, a Regressão Logística (RL) apresentou desempenho melhor quando comparado com resultados obtidos pelo modelo de LGBM, atingindo 68,8% no indicador de acurácia contra 59,9% do modelo LGBM. O mesmo ganho foi visto no indicador de sensibilidade, alcançando 82,4% para RL contra 69,6% do modelo LGBM.

Dentro do cenário desta pesquisa, a boa *performance* do indicador de sensibilidade é muito relevante na decisão estratégica, pois os falsos negativos, em que o modelo não acerta ao prever que a empresa irá ganhar a causa, são mais prejudiciais que os falsos positivos, em que o modelo erra ao prever que a empresa irá perder a causa. A primeira situação pode ser mais prejudicial financeiramente, pois ao entender que existe um forte potencial de ganho de

causa, a empresa investe na defesa, contratando advogados e mobilizando as equipes responsáveis, sem sucesso, perdendo a causa e ainda tendo valores financeiros a cumprir. No segundo cenário, a empresa, ao ver a probabilidade de perder, entra com um pedido de acordo, em que os valores são menores. Isso se soma com o grande número de procedência (perda da causa) visto na base de estudo, cerca de 72,6%.

Comparando com a literatura, Barnabé (2022) implementou diferentes modelos de previsão de ocorrência de desembolso financeiro em processos judiciais de instituições financeiras brasileiras, destacando os métodos de Regressão Logística (RL) e *Gradient Boosting* (GB). Para a RL, o modelo proposto atingiu uma acurácia de 70,7%, precisão de 68,6% e sensibilidade de 73,23%, enquanto para o método GB, acurácia de 75,94%, precisão de 76,6% e sensibilidade de 72,7%, demonstrando desempenho melhor quando comparado com RL, concluindo-se que métodos estatísticos e de *machine learning* são eficientes para predizer a ocorrência ou não de desembolso financeiro em processos judiciais. Os resultados para a RL são próximos aos de Barnabé (2022), enquanto o método LGBM apresentou resultados menores (Tabela 6). Essa diferença pode ter uma explicação na base de dados utilizada por Barnabé (2022).

Soares e Cunha (2020) propuseram modelos de predição para o risco de inadimplência de contribuintes do tributo Imposto Sobre Serviço (ISS) por meio de RL e LGBM. Os resultados demonstraram que o LGBM apresentou acurácia de 81,86% e área sob a curva ROC de 84,36%, com melhores resultados quando comparados com a acurácia de 79,74% e área sob a curva ROC de 76,64%) do RL, ou seja, o LGBM apresentou desempenho melhor quando comparado com o RL. Destaca-se que os resultados obtidos para o RL em Soares e Cunha (2020) estão próximos ao presente estudo, embora neste o LGBM tenha tido desempenho pior.

Nesse cenário, evidencia-se que os modelos de RL e LGBM com *machine learning* podem ser altamente eficientes na predição dos resultados aplicados no contexto de Direito, tais como levantamento do risco de inadimplência de ISS (Soares; Cunha, 2020) e previsão de ocorrências de desembolsos financeiros (Barnabé, 2022) e desfecho de processos judiciais abertos contra empresas do setor financeiro brasileiro.

Portanto, esse tipo de modelagem pode ser uma ferramenta importante na estratégia de defesa de ações trabalhistas para empresas desse seguimento, evitando acordos desnecessários, nos quais havia um grande potencial de ganho da causa, ou que invistam na

defesa de casos com grande probabilidade de perda, diminuindo, assim, os prejuízos financeiros, gastos administrativos e com advogados (POMPEU, 2018).

Apesar da perda de *performance* de 10% em acurácia e 15% na sensibilidade, os resultados obtidos com a modelagem em LGBM apresentam resultados satisfatórios e próximos dos encontrados na literatura. Isso demonstra que o LGBM pode ser uma alternativa para problemas relacionados ao Direito, criando oportunidades para trabalhar com técnicas de *machine learn*.

5.3 CONSIDERAÇÕES FINAIS

Um ponto de atenção relevante aos estudos de modelagens para a jurimetria é o dinamismo do cenário e das partes envolvidas. A utilização de modelos pode orientar as empresas a demonstrarem para os escritórios de advocacia alguma tendência nas estratégias de defesas e acordos, influenciando também na estratégia desses escritórios que se adaptam constantemente, de acordo com as mudanças dos resultados e padrões das empresas. Isso mostra a importância de bases atualizadas com dados recentes e que os modelos não resolvem todos os problemas, mas são uma ferramenta muito importante para a tomada de decisão e diminuição dos prejuízos, sendo ainda imprescindível o papel analítico dos advogados e seus conhecimentos empíricos e abordagens de defesa.

REFERÊNCIAS

ARANHA, F.; ZAMBALDI, F. **Análise fatorial em administração**. São Paulo: Cengage Learning, 2008.

ARFMANM, D.; TOPOLANSKY, F. G. The value of lean in the service sector: a critique of theory & practice. **International Journal of Business and Social Science**, Brooklin, v. 5, n. 2, fev, 2014.

BANDYOPADHYAY, A. **Hands-on GPU computing with Python**: Explore the capabilities of GPUs for solving high performance computational problems. Birmingham: Packt Publishing, maio,2019.

BARNABÉ, M. A. P. **Predição de risco jurídico de instituições financeiras**. 2022. 117f. Dissertação (Mestrado em Economia) - Faculdade de Economia, Universidade de Brasília, Brasília, 2022.

BATISTA, G. E. A. P. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. 204f. Tese (Doutorado em Ciências) - Universidade de São Paulo, 2003.

BUSSAB, W. O.; MORETTIN, P. A. Estatística básica. 9. ed. São Paulo: Saraiva, 2017.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados**: conceitos, tarefas, métodos e ferramentas. Goiânia: Universidade Federal de Goiás, 2009. Relatório Técnico.v.1

CAÑIZARES, M.; BARROSO, I.; ALFONSO, K. Methods for handling incomplete data in health research: a critical look. **Gaceta Sanitaria**, Milan, v. 18, n. 1, p. 58-63, 2004.

CANEL, C.; ROSEN, D.; ANDERSON, E. A. Just-in-time is not just for manufacturing: a service perspective. **Industrial Management and Data Systems**, Wembley, v. 100, n. 2, p. 51-60, 2000.

CARTER, J. V. *et al.* ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. **Surgery**, Bethesda, v. 159, n. 6, p. 1638-1645, 2016.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CONSELHO DE JUSTIÇA FEDERAL. **Estratégia Judiciário 2020**: Poder Judiciário – 2015/2020. Brasília,2014. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2015/05/7694a9118fdabdc1d16782c145bf4785.pdf. Acesso em: 05 out 2019.

CONSELHO DE JUSTIÇA FEDERAL. **Justiça em números 2019**. Brasília: CNJ, 2019. Disponível em:https://www.cnj.jus.br/wp-content/uploads/conteudo/arquivo/2019/08/justica_em_numeros20190919.pdf. Acesso em: 05 out 2019.

DIAS FILHO, J. M.; CORRAR, L. J. Regressão logística. *In*: CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. (org.). **Análise multivariada**: para os cursos de Administração, Ciências Contábeis e Economia. São Paulo: Atlas, 2007.

FÁVERO, L. P. *et al.* **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados:** estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Elsevier, 2017.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v.17, n.3, 1996.

GOLDSCHMIDT, R.R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Campus, 2005.

HAIR JUNIOR, J. F. *et al.* **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

HARRELL, F. E. *et al*. **Regression modeling strategies:** with applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer, 2015.

HONGYU, K. **Comparação do GGE-biplot ponderado e AMMI-ponderado com outros modelos de interação genótipo x ambiente**. 2015.155f. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba/SP, 2015.

HONGYU, K.; SANDANIELO, V. L. M.; OLIVEIRA JUNIOR, G. J. Análise de componentes principais: resumo teórico, aplicação e interpretação. **Engineering And Science**, Cuiabá, v. 5, n. 1, p. 83-90, jun. 2016. Disponível em:https://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398. Acesso em: 02 fev. 2021.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis.** London, UK: Pearson, 2014.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, Washington.v. 349, n. 6245, p. 255-260, 2015.

KATZ, D. M. Quantitative legal prediction – or – how i learned to stop worrying and start preparing for the data driven future of the legal services industry. **Emory Law Journal**, Atlanta, v. 62, n. 2, 2013.

LAROSE, D. T. **Discovering knowledge in data:** an introduction to data mining. John Wiley

& Sons, Inc., New Jersey, 2005. Disponível em:https://doc.lagout.org/science/0_Computer%20Science/4_Theory%20of%20Computation/Data%20Mining/Discovering%20Knowledge%20In%20Data%20An%20Introduction%20to%20Data%20Mining.pdf. Acesso em 16 ago.2021.

LEVINE, D. M.; SZANAT K. A.; STEPHAN, D. F. **Business statistic**s: a first course. London: Pearson, 2019.

LIMA, A. C. **Uma análise do ambiente competitivo dos bancos comerciais no Brasil**. 2014. Tese (Doutorado em Ciências) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2014.

LITTLE, R. J. A. Regression with missing X's: a review. **Journal of the American statistical association**, New York, v. 87, n. 420, p. 1227-1237, 1992.

LITTLE, R. J. A; RUBIN, D. B. **Statistical analysis with missing data**. New Jersey: John Wiley and Sons, 2019.

LEXISNEXIS **2020 Legal analytics study**. LexisNexis, 2019. Disponível em: https://www.lexisnexis.com/en-us/products/lexis-analytics/2020-LegalAnalytics-Study.page. Acesso em: 02 fev. 2021.

MAITRA, S.; YAN, J. Principle component analysis and partial least squares: two dimension reduction techniques for regression.**Applying Multivariate Statistical Models[***S.l.*], v. 79, p. 79-90, 2008.

MANLY, B.F.J.; ALBERTO, J.A. N. **Multivariate statistical methods:** a primer. 4th ed. New York**:** Chapman and Hall/CRC, 2016.

MCCUE, C. **Data mining and predictive analysis**: intelligence gathering and crime analysis. 2 nd ed. Amesterdã: Elsevier Inc., 2007.

METZ, C. E Statistical analysis of ROC data in evaluating diagnostic performance. . **Multiple regression analysis**: applications in the health sciences, v. 13. p. 365-384, 1986.

MICROSOFT CORPORATION. **LGBM Documentation**. Microsoft, 2019**.**Disponível em: https://lightgbm.readthedocs.io/en/latest/index.html . Acesso em: 10 jun. 2021.

MICROSOFT CORPORATION. **LGBM Parameters.** 2019**.** Disponível em: https://lightgbm.readthedocs.io/en/latest/Parameters.html . Acesso em: 10 jun. 2021.

MICROSOFT CORPORATION. **LGBM Parameters Tunning.** Microsoft, 2019. Disponível em: https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html . Acesso em: 10 jun. 2021.

MILOCA, S. A.; CONEJO, P. D. Multicolinearidade em modelos de regressão. *In:* XXII SEMANA ACADÊMICA DE MATEMÁTICA, p. 25-34, 2008, Cascavel. Disponível em: https://projetos.unioeste.br/cursos/cascavel/matematica/xxiisam/artigos/todos.pdf.

Acesso em 13 jun.2021

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005. 297 p.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, Barueri: Manole,, n. 1, 2003. v. 1.

MOURÃO, M. F. *et al*. Adjusting covariates in CRIB score index using ROC regression analysis.*In*p. 157-171. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-21407-8_12. Acesso em: 20 jun. 2021.

NILSSON, N. J. **Introduction to Machine Learning**. Stanford: Stanford University, 1998.

NUNES, M. G; COELHO, F. U. Pesquisa a serviço da advocacia. **Jornal Valor**, SP. Caderno Legislação & Tributos, sexta-feira e fim de semana, 20,21 e 22 de agosto de 2010, p. E2.

NUNES, D.; GUEDES, M. Jurimetria: como a estatística pode reinventar o direito. **Revista dos Tribunais**, São Paulo, p. 115-116, 2016.

NUNES, D.; DUARTE, F. A. Jurimetria e tecnologia: diálogos essenciais com o direito processual. **Revista de Processo**, v. 299, p. 407-450, 2020.

PAULINO, C. D. A Evolução da estatística bayesiana em Portugal.. *In*: ROSADO, F. **Memorial da Sociedade Portuguesa de Estatística**. Lisboa: Sociedade Portuguesa de Estatística, 2005. p. 215-217.

PEDREGOSA, F. *et al*. SCIKIT-LEARN: machine learning in Python. 2011. **Journal of Machine Learning Research 12.**Disponível em: https://www.jmlr.org/papers.v12/pedregosa11.html . Acesso em 10 jun,2021.

POMPEU, A. Judiciário brasileiro tem 80,1 milhões de processos em tramitação. **Consultor Jurídico**, São Paulo, 2018. Disponível em: https://www.conjur.com.br/2018-ago-27/judiciario-brasileiro-801milhoes-processos-tramitacao. Acesso em: 2 fev. 2021.

REGAZZI, A. J.; CRUZ, C. D. **Análise multivariada aplicada**. Viçosa: Editora UFV, 2020.

RUBIN, D. B. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *In*: SURVEY RESEARCH METHODS SECTION, p. 20-34,.1978, Alexandria. **Proceedings[…]** Alexandria: American Statistical Association, 1978.

RUBIN D.B. **Multiple Imputation for Nonresponse in Surveys.** New York: Wiley. 1987

RUBIN, D. B. Multiple imputation after 18+ years. **Journal of the American Statistical Association**,Washington, v. 91, n, 434, p. 473-489, 1996.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **Plos One**, Bergen, v. 10, n. 3,

p. 1-21, 4 mar. 2015. . Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432. Acesso em: 20 jun. 2021.

SALAMA, B.; CARLOTTI, D.; YEUNG, L. **As decisões da justiça trabalhista são imprevisíveis?**. São Paulo: Insper, 2018.

SALAMA, B. M.; CARLOTTI, D.; YEUNG, L. **Quando litigar vale mais a pena do que fazer acordo**: os grandes litigantes na justiça trabalhista. São Paulo: FGV, 2019. 11 p. (Relatório 2: O Judiciário destrinchado pelo 'Big Data'). Disponível em: https://works.bepress.com/bruno_meyerhof_salama/143/ . Acesso em: 26 mar. 2020.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. **Psychological methods**, Washington, v. 7, n. 2, p. 147, 2002.

SELIYA, N; KHOSHGOFTAAR, T. M. Software quality modeling with limited apriori defect data. *In*: ZHU, X.; Davidson, I. **Knowledge discovery and data mining**: challenges and realities. Hershey: Information Science Reference, 2007.p. 1–16.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier, 2016.

SMITH, L. I. **A tutorial on principal components analysis**. Otago: University of Otago, 2002.

SOARES, V. G.; CUNHA, R. Predição de irregularidade fiscal dos contribuintes do tributo ISS. *In*: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 35.,p.223-228, 2020.**Anais**

SOARES JUNIOR, J. S.; QUINTELLA, R. H. Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica. **Revista de Administração Pública**, Rio de Janeiro, v. 39, n. 5, p. 1077-1107, 2005.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Heidelberg:Springer. , 2006. p. 1015-1021(Série)Disponível em: https://link.springer.com/chapter/10.1007/11941439_114. Acesso em: 20 jun. 2021.

SPACKMAN, K. A. Signal detection theory: valuable tools for evaluating inductive learning. *In*: international workshop on Machine learningachusetts/EUA: Morgan Kaufmann, 1989.p. 160-163.

STERNE, J. A. C. *et al.* **Multiple imputation for missing data in epidemiological and clinical research**: potential and pitfalls. London: BMJ, , 2009, v.338.

TRIBUNAL REGIONAL DO TRABALHO DA 1ª REGIÃO. **Relatório estatístico do 1º grau**: 2018. Rio de Janeiro, 2018. Disponível em: https://www.trt1.jus.br/documents/21792. Acesso em: 16 de jun. 2021.

TRIBUNAL DA JUSTIÇA DO PARÁ. **Os conceitos e diferenças entre comarca, vara, entrância e instância.** Belém: TJAP, 2016. Disponível em: https://www.tjap.jus.br/portal/ . Acesso em 20 jun. 2021.

TRIBUNAL SUPERIOR DO TRABALHO.**Tabelas processuais unificadas da justiça do trabalho**. Brasília, 2021. Disponível em: https://www.tst.jus.br/web/corregedoria/tabelas-processuais . Acesso em 16 de jun. 2021.

VAN D. H.; GEERT, J. M. G *et al.* Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. **Journal of clinical epidemiology**, Oxford, v. 59, n. 10, p. 1102-1109, 2006.

WANG, L.; FU,X.. **Data mining with computational intelligence**. New York: Springer Science and Business Media, 2006.

ZHANG, P. Multiple imputation: theory and method. **International Statistical Revi ew/Revue Internationale de Statistique**, Edinburgh, v. 71, n.3, dec.2003.