DOI: 10.5748/19CONTECSI/PSE/DSC/7117

JURIMETRIA E VISUALIZAÇÃO DE DADOS: ANÁLISE DE DECISÕES DE PROCESSOS DO TJSP COM BASE EM DATA PIPELINES.

Marcos Antonio Speca Junior
Universidade Presbiteriana Mackenzie

Fabio Silva Lopes
Universidade Presbiteriana Mackenzie - Faculdade de Computação e Informática - Programa de pós-graduação em Computação Aplicada

JURIMETRY AND DATA VISUALIZATION: ANALYSIS OF JUDGMENTS IN SÃO PAULO COURT OF JUSTICE BASED ON DATA PIPELINES.

ABSTRACT

The objective of this work was to structure a data pipeline to perform collection, storage, processing and visualization of data regarding a specific theme. The theme chosen was the comparison of lawsuits from the court of justice of the state of São Paulo (TJSP) of companies that operate in the same segment. The purpose of the designed pipeline is to collect lawsuits and their first decisions, store the raw data, apply some data treatments to synthesize them and later make them available in an analytical layer so that it would be possible to compare companies of the same segment and their respective success rates in the aforementioned court. The generated datasets allowed the construction of visualizations in Power BI and made it possible to observe discrepant situations between similar lawsuits, which lack further studies of these lawsuits in the São Paulo courts.

Keywords: jurimetrics, data pipelines, data visualization

JURIMETRIA E VISUALIZAÇÃO DE DADOS: ANÁLISE DE DECISÕES DE PROCESSOS DO TJSP COM BASE EM DATA PIPELINES.

RESUMO

O objetivo deste trabalho foi estruturar um pipeline de dados para realizar a coleta, armazenamento, processamento e visualização de dados a respeito de um tema específico. O tema escolhido foi a comparação de processos judiciais do Tribunal de Justiça do estado de São Paulo (TJSP) de empresas que atuam no mesmo segmento. O pipeline desenhado tem como objetivo realizar a coleta de processos e suas decisões de 1ª instância, armazenar os dados brutos, aplicar algumas tratativas de dados para sintetizá-los e posteriormente disponibilizar em uma camada analítica para que fosse possível realizar a comparação entre empresas do mesmo segmento e suas respectivas taxas de êxito no tribunal citado. Os datasets gerados permitiram a construção de visualizações em Power BI e possibilitaram observar situações discrepantes entre processos similares, que carecem de novos estudos sobre a tramitação dos processos nos foros de São Paulo.

Palavras-chave: jurimetria, data pipelines, visualização de dados

1. Introdução

O Direito, como objeto de pesquisa, é uma área que apresenta muitos desafios para a Ciência de Dados. Entre eles, está o alcance do equilíbrio entre o uso de indicadores qualitativos e quantitativos de modo a reduzir a imprecisão e erros de estimação (Andrade, 2018).

Não obstante, a eficiência judicial tem gerado discussões interessantes no âmbito acadêmico e comercial pois a morosidade dos processos, aliada a outras questões como burocracia, má-

gestão, legislação processual inadequada, falta de transparência, judicialização excessiva, estrutura inadequada e ausência de democratização do acesso à justiça (Ponciano, 2015).

Embora o tempo do processo judicial seja tido como diferido, encarado como sinônimo de segurança e concebido como uma relação de ordem e autoridade, as questões indicadas por Ponciano (2015) geram efeitos indesejados para a sociedade, entre eles, a insatisfação com o serviço público oferecido (Faria, 2004).

Segundo o Anuário da Justiça Brasil 2022, o Brasil lida com 80 milhões de processos em tramitação nos tribunais do judiciário, sendo que, neste ano, foram protocolados aproximadamente 27 milhões de novos processos contra 26 milhões julgados. Estes números refletem processos de diversas naturezas e assuntos, em diversas áreas da justiça do Brasil (Consultor Jurídico, 2022). O primeiro grau de jurisdição é o mais sobrecarregado e, por conseguinte é aquele que presta serviços mais aquém da qualidade esperada (CNJ, 2018). O mesmo relatório apresenta uma série histórica com dados coletados desde 2009 onde há um crescimento acumulado de 18,3% dos processos em tramitação.

Além dos números elevados de processos, outro fato relevante é que a grande maioria destes processos já tramita em meios eletrônicos, em tribunais estaduais, e compreendem processos da esfera cível, direito do consumidor e execuções fiscais, ou seja, os dados destes processos estão armazenados em sistemas dos próprios tribunais e são disponibilizados para consulta. Um exemplo é o site do Tribunal de Justiça do Estado de São Paulo (TJSP, 2022).

Conforme legislação atual do Brasil, os dados dos processos judiciais são dados públicos e são passíveis de serem coletados e analisados de forma livre segundo descrito no Artigo 93, inciso IX da Constituição Federal, excetuando os casos que tramitam em segredo de justiça.

Se os dados são públicos, podem ser coletados para posterior análise. Logo, a pergunta desta pesquisa se pautou na seguinte interrogação: Do ponto de vista da Ciência de Dados, é possível gerar insights a partir da coleta de dados públicos de processos de modo a prover melhorias na gestão de demandas judiciais por parte das empresas?

Considerando os pressupostos já descritos, o objetivo deste estudo foi desenvolver uma arquitetura de dados capaz de coletar, processar e entregar produtos analíticos e contribuir para a geração de novos insights no âmbito dos processos jurídicos.

Entender fenômenos a partir dos dados disponíveis é a contribuição do cientista de dados nos diversos contextos em que ele atua. No cenário apresentado, é importante conhecer o papel da jurimetria aliada à ciência de dados para promover contribuições significativas ao contexto jurídico brasileiro. O estudo de Maia e Bezerra (2020) apontou carência de estudos desta natureza. Foram publicados somente 84 artigos sobre o tema entre 2002 e 2019. A ciência de dados vem evoluindo sobremaneira nos últimos anos, em diversos contextos, logo, o gap é grande e não há justificativas pois, os dados, a fonte primeira dos estudos desta natureza, está disponível.

Este artigo, além da introdução, foi estruturado em cinco seções: A Seção 2 apresenta o referencial teórico sobre jurimetria e data pipelines, e foi complementada com trabalhos correlatos sobre o tema. A seção 3 traz o descritivo da metodologia utilizada. A seção 4 apresenta os resultados da pesquisa, bem como a discussão pertinente. Por fim, a seção 5 traz as considerações finais e trabalhos futuros para continuidade da pesquisa nesta área.

2. Referencial Teórico

2.1.Jurimetria

A Jurimetria é um termo cunhado por Loevinger em 1949 para definir o conjunto de investigações, tanto lógica-matemática como estatística voltada para tipos distintos de análise de informação jurídica (Ramirez, et al, 2016). A proposta de Loevinger estava atrelada a um contexto no qual a aplicação de métodos analíticos pode trazer progresso e segurança jurídica (Maia e Bezerra, 2020).

Contudo a utilização de análises estatísticas relacionadas ao Direito remonta ao trabalho de Bernoulli em 1709. Uma definição que adotou-se para este estudo foi a de Machado (2003), "*A convergência entre o privilégio de exploração da criação intelectual e a elaboração de um direito do espaço virtual com suas consequências sobre o domínio público*". Esta definição toca em pontos relevantes para este estudo como o domínio público dos dados e o espaço virtual em que estamos atuando.

A visão de Loevinger apontava para a importância científica dos métodos estatísticos no contexto do direito, pois achava que o conhecimento da lei poderia ser melhor compreendido por meio da observação mais do que da especulação.

É fato que o método estatístico consiste em etapas subsequentes de coleta de dados, organização, e aplicação de modelos oriundos de teorias descritivas e probabilísticas, usados na expectativa de explicar fenômenos como a frequência com que determinados eventos ocorrem no mundo. No campo jurídico, não é diferente.

Segundo Salzano (2020), quando se leva o processo para o ambiente eletrônico, ganha-se ferramentas que podem colaborar com as demandas massificadas que sobrecarregam o judiciário. A inteligência artificial como ferramenta de auxílio ao processo judicial é uma realidade já posta na jurisdição.

Tais análises são interessantes pois para as partes envolvidas nestes processos é relevante entender tendências do judiciário. Um limitador, porém, é que tais dados não são disponíveis via APIs, mas sim através de consultas diretamente nos sites dos tribunais. Adicionalmente, conforme a estrutura do judiciário, os tribunais de diferentes justiças (Estadual, do Trabalho, Militar e Eleitoral) possuem diferentes sistemas, e diferentes métodos de armazenamento e disponibilização destes dados processuais. Além disso, um processo judicial segue diferentes procedimentos a depender de sua tramitação, instâncias e tipo de justiça.

2.2.Data Pipelines e a Jurimetria

Pipelines de dados são cadeias de atividades interconectadas que visam coletar os dados de sua origem, processá-los, enriquecê-los e por fim disponibilizá-los para análise e geração de insights (Munappy et al, 2020).

Considerando o contexto de grande volume de processos judiciais no Brasil, o uso de data pipelines é indicado pois agregam práticas consolidadas de mercado em termos de coleta e armazenamento e processamento destes dados.

O processo inicia-se como a definição dos mecanismos de coleta automatizada, seja no formato batch, streaming ou eventual. O armazenamento dos dados está presente em diversas etapas, desde o armazenamento temporário de dados brutos até o armazenamento dos dados enriquecidos, bem como, aqueles já prontos para consumo, em aplicações ou disponíveis para usuários como DAAS (*data-as-a-service*). Em cada etapa do pipeline, a depender de seu objetivo, os dados e metadados podem ser armazenados em diferentes formatos, com diferentes tecnologias e abordagens conceituais.

A etapa seguinte do pipeline é a aplicação de modelos ao conjunto de dados, ou a implementação de visualizações, gerando assim, produtos analíticos de interesse aos stakeholders do projeto.

A última etapa de um data pipeline configura a disponibilização e o consumo dos produtos analíticos. Isso se dá por meio de representações numéricas, tabulares e gráficas, ou por meio de APIs (*Application Programming Interface*) que disponibilizam os resultados para um determinado processo. Segundo Chen e Zhang (2014), a visualização de dados tem como objetivo a representação do conhecimento, de modo mais intuitivo e efetivo, por meio de diferentes formas gráficas.

No caso do poder judiciário, os dados são disponibilizados por meio de aplicações de consulta em Websites, em páginas HTML (HyperText Markup Language). Os Tribunais não possuem APIs para consumo de dados destes repositórios.

Muitas informações relevantes estão disponíveis em tags HTML que dão contexto para a informação. O Direito é fonte geradora de processos em diferentes áreas de atuação com informações distintas, podendo ser alteradas de acordo com seu contexto. Isso faz com que os atributos dos processos variem, tornando a modelagem de dados uma tarefa complexa. Tais particularidades deste contexto são consideradas na coleta de dados, processada por meio de técnicas de Web Scraping, como ocorreu neste estudo.

Considerando as características apontadas e a constante evolução e inclusão de atributos adicionais através do tempo, permitem classificar o conjunto de dados dos processos disponíveis nos sites dos Tribunais, como dados semiestruturados uma vez que não possuem estrutura definida. Isso requer uma etapa adicional no pipeline para estruturação dos dados. De modo complementar, é importante que o desenvolvimento de uma solução neste contexto, apresente aptidão para escalabilidade devido ao grande volume de dados a serem coletados e armazenados.

Após a transformação e enriquecimento dos dados, os dados analíticos estruturados são carregados em um banco de dados relacional para facilitar o cruzamento e a visualização dos dados por meio de ferramentas de visualização. Este armazenamento permite uma modelagem dimensional, utilizando o modelo "star-schema" com tabela Fato e Dimensões. Esta modelagem favorece a evolução para data marts segmentados para estudos específicos como decisões de 2ª instância e outras análises referentes ao mesmo conceito.

2.3.Trabalhos Correlatos

O relacionamento entre o Direito e a Estatística é próximo e importante. A palavra Probabilidade é um termo usual em ambos os contextos. Contudo, o estatístico deve entender que um litígio é um processo entre adversários. Um considera a estratégia do outro para preparar a sua defesa (Gray, 2014).

O estudo de Luvizotto e Garcia (2020) aplica jurimetria em dados públicos do TCU (Tribunal de Contas da União) para identificar casos de jurisprudência. Os autores partem da hipótese de que o uso da jurimetria pode melhorar o desempenho deste órgão além de contribuir para a segurança jurídica e accontability.

Já o trabalho de Garcia (2022) pautou-se na construção de indicadores de corrupção para o estado do Rio de Janeiro para melhorar ações de planejamento e monitoramento de contas públicas. Utilizou-se os dados do cadastro de contas consideradas irregulares fornecido pelo TCU. O estudo definiu dois indicadores: o CIPDK (quantidade de contas irregulares para cada

dez mil habitantes do município) e o VDPK (valor do débito das contas irregulares para cada mil reais do produto interno bruto do município).

O projeto Victor é uma parceria entre o Supremo Tribunal Federal (STF) e a Universidade de Brasília para executar pesquisa textual em processos e implementar recursos de Inteligência Artificial para análise de admissibilidade recursal. Um classificador identifica peças de acordo com a nomenclatura utilizada pelo STF (STF, 2021).

3. Materiais e Métodos

Este estudo foi estruturado a partir do modelo de referência Crisp-DM (Cross-Industry Standard Process for Data Mining) posposto por Shearer (2000), onde o processo de data mining é fracionado em seis etapas.

Na etapa 1 (Business Understanding), desenvolveu-se um estudo exploratório sobre os dados de processos jurídicos, disponíveis, bem como ferramentas de apoio para coletar e estruturar o dataset que serviu de base para as etapas seguintes.

Na etapa 2 (Data Understanding) , os dados foram coletados, armazenados e analisados por meio de técnicas de análise exploratória de dados. Um metadados foi organizado nesta etapa.

A etapa 3 (Data Preparation) consistiu em organizar e transformar os dados para estruturas mais adequadas aos modelos analíticos que seriam expostos.

Optou-se para esta etapa o armazenamento relacional com o banco de dados PostgreSQL por ser um banco de dados relacional open source que favorece a análise de dados, pois diversas ferramentas de análise de dados já possuem conectores disponíveis para este banco de dados, além de permitir a utilização da linguagem SQL o que acelera a análise de dados por parte dos analistas de dados. Além disso, existem diversas opções de utilização deste banco em nuvens dos principais players de mercado.

Os modelos de visualização de dados foram planejados e executados na etapa 4 (Modeling). Os resultados foram analisados na etapa 5 (Evaluation) e os resultados foram compilados na etapa 6 (Deployment).

Optou-se por utilizar scripts Python executados em Jupyter, rodando localmente em computadores desktops para executar as etapas supracitadas. O armazenamento dos dados brutos foi feito em um banco de dados Mongodb e posteriormente processados e enviados a um banco de dados relacional PostgreSQL e as visualizações foram desenvolvidas em Microsoft PowerBI.

4. Resultados e Discussão

O processo descrito permitiu estruturar um pipeline de dados a partir de coletas batch que foram realizadas no período de agosto a outubro de 2022. Optou-se por desenvolver uma aplicação de web scraping, por meio da biblioteca BeautifulSoup v.4, que foi escrita na forma de script Python v. 3.10.7 e executada em uma instância local Jupyter Notebook. A Figura 1 apresenta um diagrama esquemático da arquitetura proposta para este estudo.

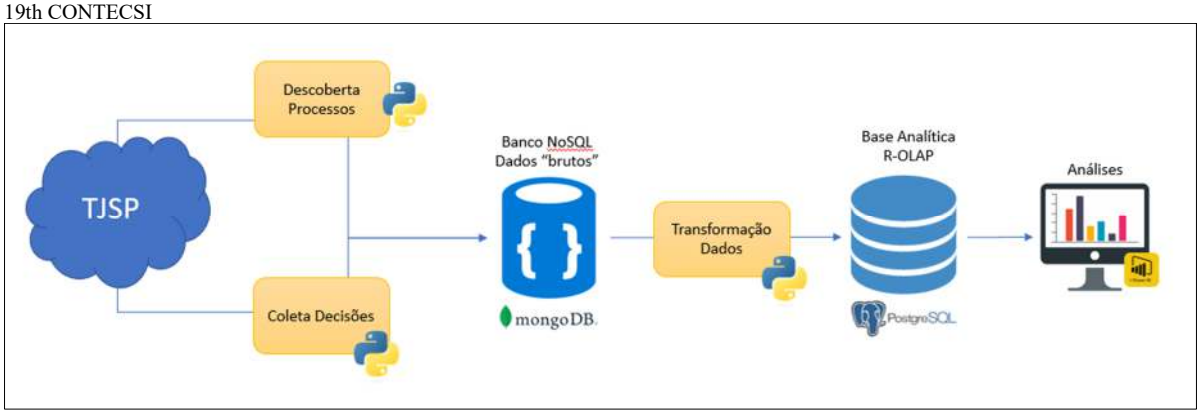

Figura 1: Diagrama detalhado do Pipeline de Dados. Elaborado pelos autores.

4.1.Coleta de Dados

A coleta de dados foi realizada em duas etapas distintas: A descoberta dos processos de determinadas empresas, utilizando a busca por CNPJ do site do tribunal; posteriormente foi realizada a busca da decisão de 1ª Instância do processo coletado através de seu número.

Todos os dados já tratados pelo webscrapper foram armazenados em duas coleções no banco de dados NoSQL, coleção "Processos" e coleção "Decisões".

As etapas de coleta de Processos e Decisões estão detalhadas na Figura 2.

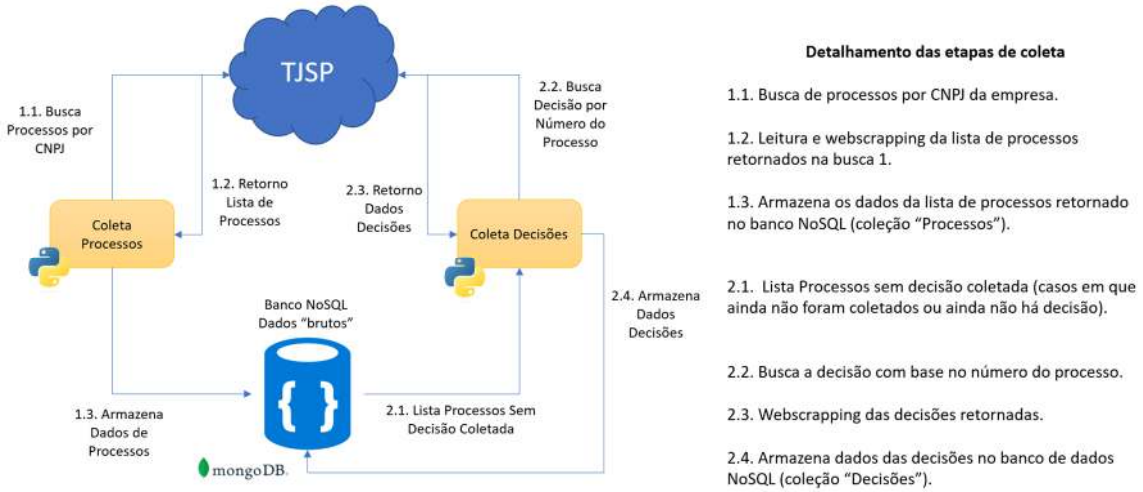

Figura 2: Diagrama detalhado da etapa de coleta de dados. Elaborado pelos autores.

Durante a fase de coleta, coletamos dados de 55.954 processos da justiça estadual, e destes coletamos 28.064 decisões de 1ª instância. Vale destacar que já era esperado que nem todos os processos coletados teriam decisões, pois existem muitos casos ainda em fase inicial de tramitação.

Considerando o cenário descrito, elaborou-se o metadados do dataset Processo. O Quadro 1 apresenta este metadado e algumas características que vão direcionar as análises futuras.

Atributo Processo	Descritivo
nro_processo	Número do Processo (chave), numeração única no CNJ, ligação entre processo e decisão.
link_processo	Hiperlink para o acesso ao processo completo.
nome_parte	Nome da Parte, no caso da empresa cuja o CNPJ foi utilizado na pesquisa.
tipo_participacao	De uma forma resumida, indica se a empresa é Ré ou Autora no processo.
classe_processo	Indicação de classe processual (Procedimento Comum, Juizado Especial etc.).
assunto_processo	Assunto ou tema em discussão no processo.
data_local_distribuicao	Informação sobre data da distribuição (início) do processo e o local de tramitação (os tribunais apresentam a informação consolidada, sendo necessário o desmembramento futuro).
cod_empresa	Código criado manualmente para facilitar a unificação de processos do mesmo cnpj porém com nomes de partes diferentes.

Quadro 1: de metadados da coleção "Processo".

Além dos atributos dos processos, no contexto de decisões, outros atributos relevantes foram coletados e armazenados, conforme apresenta o Quadro 2, contendo o metadados do dataset Decisões.

Atributo Decisões	Descritivo
nro_processo	Número do Processo (chave), numeração única no CNJ, ligação entre processo e decisão.
classe_processo	Indicação de classe processual (Procedimento Comum, Juizado Especial etc.).
assunto_processo	Assunto ou tema em discussão no processo.
magistrado	Nome do magistrado/juiz que realizou o julgamento em 1ª instância
comarca	Comarca relacionada a decisão
foro	Foro relacionado a decisão
vara	Vara relacionada a decisão
data_disp	Data de disponibilização da decisão
texto_decisao	Texto da decisão (texto completo do julgamento).

Quadro 2: Metadados da coleção "Decisões". Elaborado pelos autores.

4.3.Preparação e Transformação de Dados

Na etapa de preparação e transformação de dados realizamos duas tarefas distintas, a primeira foi identificar e sintetizar o resultado da decisão e posteriormente carregamos os dados em uma base de dados relacional para que esta possa servir de consulta analítica para a análise dos dados.

4.3.1. Identificação dos resultados nas decisões.

Embora haja um padrão textual nas decisões proferidas pelos magistrados, não há um atributo sintético onde seja apontado de maneira estruturada o resultado da decisão, portanto foi necessário realizar um processo de identificação destes resultados com base no texto da decisão.

Cada decisão possui detalhes e aspectos únicos relacionados ao seu julgamento, porém podemos sintetizar os resultados conforme apresentado no Quadro 3.

Resultado	Observações
Procedente	Decisão indica que o pedido do autor é procedente totalmente e condena o Réu a alguma ação.
Parcialmente Procedente	Decisão indica que o pedido é procedente apenas em parte e condena o Réu a alguma ação apenas onde foi considerado procedente.
Improcedente	Decisão indica que o pedido do autor é improcedente.
Acordo	Indica que foi realizado um acordo entre as partes e a Decisão homologa este acordo.
Extinto	O caso foi extinto sem julgamento do mérito por algum motivo.

Quadro 3: Resultados identificados nas Decisões. Elaborado pelos autores.

Para identificar cada resultado com base no texto das decisões, utilizamos análises textuais baseadas em expressões regulares (ReGex), visto que os textos das decisões seguem um padrão.

Elaborou-se um script em Python que lê o texto da decisão e de acordo com uma combinação do texto classifica o resultado em uma das opções citadas anteriormente.

Com apenas algumas regras de expressão regular permitiu classificar 22.981 decisões (84,67% das decisões). Este montante compreende todas as decisões do TJSP disponíveis para consulta em agosto/2022. A Figura 3 apresenta a proporção de decisões classificadas por segmento.

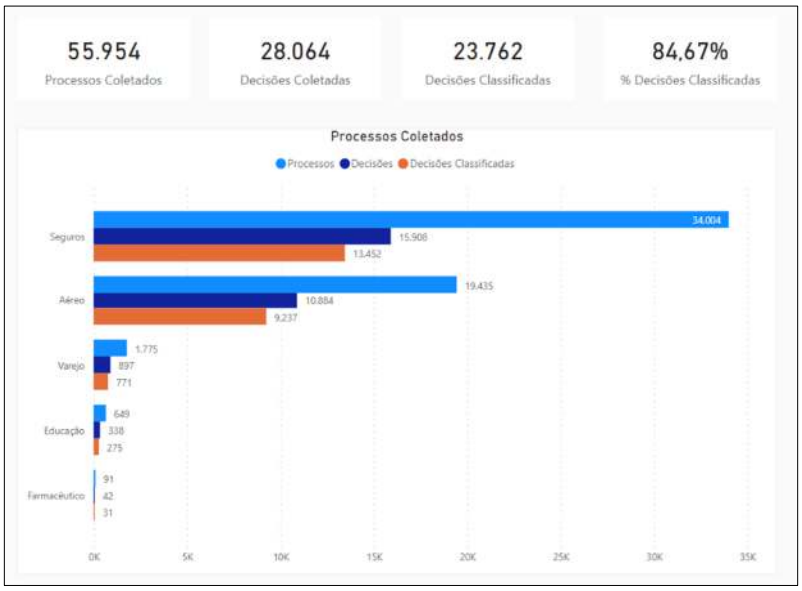

Figura 3: Proporção de Processos, Decisões e Decisões Classificadas por Segmento. Elaborado pelos autores.

Além desta classificação dos resultados foram realizadas algumas tratativas básicas, como a identificação da data de distribuição dos processos a partir do atributo "data_local_distribuicao" e a consolidação das empresas em segmentos de negócio conforme seu ramo de atuação.

Por fim foi realizada uma carga em um banco de dados relacional (PostgreSQL) em duas tabelas para que ficassem disponíveis para a conexão com a ferramenta de análise.

4.4. Análise e Visualização de Dados

Para a análise e visualização de dados optamos por uma ferramenta "self-service BI" que facilita a análise exploratória bem como a criação de dashboards para a apresentação de tais informações. Como em um primeiro momento a análise de dados será de cunho exploratório, ou seja, não temos certeza de que indicadores e informações podemos extrair, o uso de uma ferramenta que acelera a visualização dos dados sem codificação pesou na decisão da escolha.

Escolhemos a plataforma Power BI por ser uma plataforma gratuita (versão Desktop), ser uma ferramenta líder de mercado (Microsoft, 2022) e permitir a visualização dos dados de maneira ágil pois já possui conector nativo com a base de dados PostgreSQL. Além disso, a plataforma Power BI é amplamente utilizada no mercado, o que favorece a curva de aprendizado de seus usuários.

Posteriormente gráficos e análises poderão ser gerados em outras tecnologias, visto que os dados analíticos serão tratados e persistidos no banco de dados PostgreSQL e não em etapas de transformação de dados existentes na ferramenta Power BI.

Para realizar as análises optou-se por desenvolver algumas métricas chaves que nortearam o processo de identificação de insights, conforme apresenta o Quadro 4.

Métrica	**Descrição**
Quantidade de Processos	Quantidade de processos coletados.
Quantidade de Decisões	Quantidade de decisões coletadas.
Quantidade de Decisões Classificadas	Quantidade de decisões onde foi possível classificar seu resultado.
% de Decisões Classificadas	Percentual de decisões classificadas em relação ao total de decisões
Quantidade de Improcedências	Quantidade de decisões cuja o resultado foi "Improcedente" (Resultado mais relevante do ponto de vista da empresa que é processada).
% de Improcedências	Percentual de decisões improcedentes em relação ao total de decisões.

Quadro 4: Métricas criadas com base nos dados coletados. Elaborado pelos autores.

Para realizar algumas análises foi criado um painel com as principais métricas e diversas visões por Foro, Assuntos, Magistrado, Segmento Empresarial e outras. Foram incluídos alguns filtros neste painel para que fosse possível segmentar algumas análises por empresa, assunto e foro. A Figura 4 apresenta uma visão parcial deste painel.

Figura 4: Painel de Análise de Decisões. Elaborado pelos autores.

4.5.Insights identificados

Como o principal objetivo deste trabalho é identificar possíveis insights para empresas com base na coleta de dados públicos, fizemos algumas análises com o segmento de “Seguros”.

Uma primeira análise é a identificação de taxa de improcedência por Foro. Considerando que os processos do segmento de seguros são similares, pois tratam do mesmo assunto, identificamos algumas discrepâncias em relação aos percentuais de improcedências e procedências dos processos.

Pode-se observar que entre foros analisados, Campinas se destaca pelo alto percentual de resultados improcedentes, como é possível observar na Figura 5 que apresenta os 10 fóruns com maior número de processos.

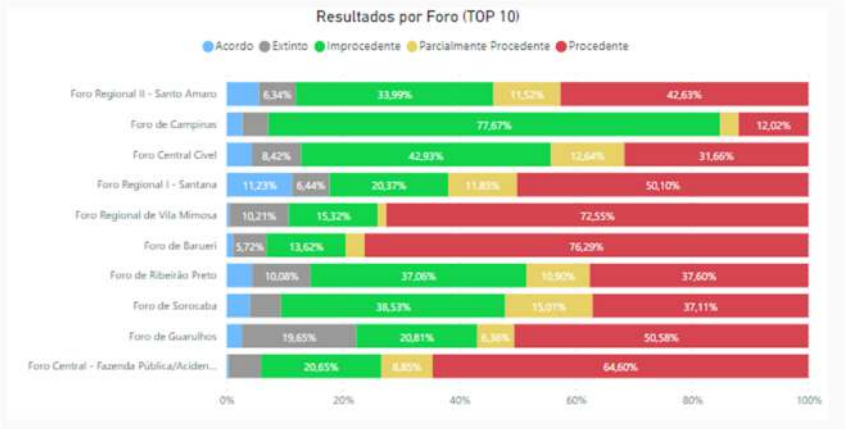

Figura 5: Comparativo parcial de Resultados por Foro. Elaborado pelos autores.

Esta diferença percentual de improcedência significativa entre os foros dos tribunais podem ser oriundas de diversos fatores como: concentração de um determinado assunto em um determinado foro, processos com diferentes tipos de participação das empresas (participam como Réus ou Autores), entre outros. De qualquer maneira podemos avaliar que este tipo de visualização (comparação percentual de improcedência entre os foros) pode trazer insights relevantes para as empresas, mas que demandam análises futuras com maior profundidade.

Considerando a data da decisão, outra informação relevante foi a identificação de um “pico” de acordos realizados em agosto de 2022, em relação a outros períodos analisados. Esta análise pode indicar alguns fatores como mutirões de acordo, ou ações dos tribunais para diminuir os casos em tramitação, conforme observa-se na Figura 6.

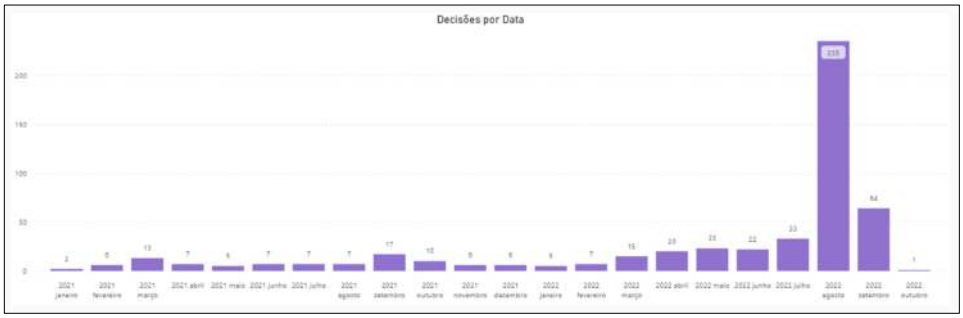

Figura 6: Quantitativo de Decisões por data de julgamento. Elaborado pelos autores.

Uma das análises mais relevantes é a comparação entre empresas do mesmo segmento e sua taxa de improcedência. Consideramos apenas o assunto “Indenização por Dano Moral”, o

assunto com maior número de processos, As Figuras 7 e 8 apresentam Decisões e Taxas de improcedência, respectivamente, para empresas do segmento de seguros.

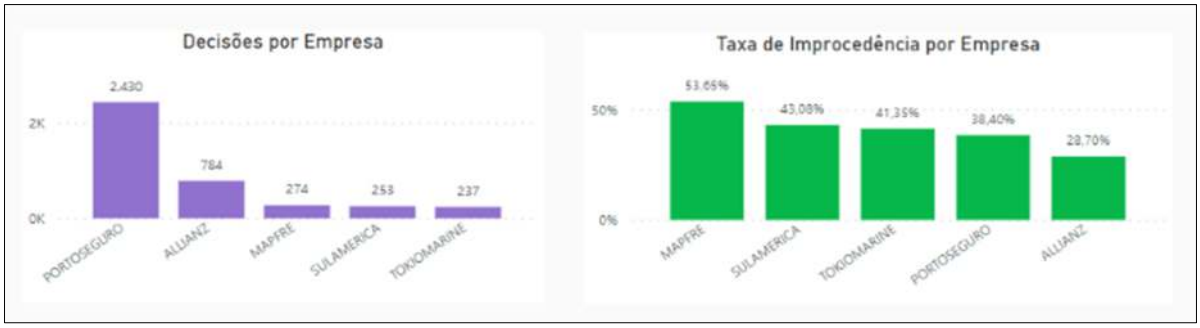

Figura 7 - Decisões por Empresas de Seguros.

Figura 8- Taxa de Improcedência por empresa de Seguros.

Elaborado pelos autores.

É possível observar que uma das empresas possui taxa de improcedência de 53,65% e outra, a com menor taxa, tem um valor de 28.70% de casos improcedentes. Esta diferença também é significativa e pode apontar para uma melhoria de processos internos das empresas após uma análise mais aprofundada dos casos. A diferença nas taxas pode suscitar questionamentos. Porque há diferença significativa entre as taxas?

As análises supracitadas são apenas alguns exemplos de possibilidades de geração de insights através do pipeline de coleta, transformação e visualização de dados construídos neste trabalho. Outras análises podem ser realizadas a partir dos dados coletados, como: Similaridade de Decisões nos mesmos temas, Tempo médio entre a distribuição do processo e decisão, identificação de valores de danos morais e sua variabilidade entre outros.

4.6.Limitações do Trabalho

Durante a execução deste estudo, observaram-se algumas limitações que cabe destacar.

- A coleta considerou apenas o tribunal de justiça de São Paulo, e para as empresas que possuem um volume grande de processos faz sentido expandir a análise para os demais tribunais de justiça do Brasil.

- No pipeline construído não foi implantado um processo de automação e coleta contínua, porém entende-se que, com a utilização de ferramentas de ingestão de dados, como o Apache Airflow, este fluxo pode ser 100% automatizado.

- No processo de transformação de dados apenas algumas transformações foram aplicadas. Outras poderiam ser aplicadas a título de comparação de resultados.

- Além dos resultados das decisões, os valores de indenização e condenação podem ser extraídos dos textos das decisões através de processos de data mining mais sofisticados, isto pode trazer uma visão mais aprofundada do impacto das condenações para as empresas.

- Coletaram-se dados apenas de algumas empresas do segmento de Seguros, portanto, as comparações realizadas podem gerar resultados não generalizáveis. A generalização carece de estudos mais amplos.

5. Considerações Finais

O objetivo deste trabalho foi estruturar um pipeline de dados para a coleta de informações dos tribunais de justiça, especificamente do tribunal de justiça do estado de São Paulo e criar mecanismos para analisar processos de empresas do mesmo segmento.

Considerando a implementação do pipeline de dados de ponta a ponta e a realização de análises após estruturação dos dados, é lógico concluir que este tipo de arquitetura permite produzir insights relevantes para empresas que precisam gerir e atuar em demandas judiciais, principalmente aquelas em que as empresas figuram o polo passivo (são processadas).

No entanto, vale destacar o desafio da coleta de dados, pois ainda que os dados sejam de acesso público, conforme a constituição federal, os tribunais não fornecem uma forma de consumo mais simples, como APIs dedicadas para esta finalidade, fazendo com que a coleta de dados seja feita por meio de técnicas baseadas em webscrappers, o que pode impactar na qualidade dos dados coletados. De modo complementar, existem problemas de disponibilidade do tribunal para fornecer os dados durante períodos de maior demanda.

Do ponto de vista tecnológico e os aspectos do desafio do pipeline de dados, em termos de coleta, tratamento e armazenamento de dados semiestruturados de processos judiciais, depende diretamente da ação do cientista de dados em conjunto com os especialistas da área.

Foi possível também observar a aderência e flexibilidade para incluir documentos com diferentes atributos, realizar pesquisas, ajustes e consolidação de informações para análises. Outrossim, os datatasets e visualizações criadas, ainda podem contribuir para a geração de novos insights, a depender da criatividade dos especialistas do contexto.

A título de trabalhos futuros, a automação do processo e a ampliação do escopo geográfico são questões a serem implementadas neste contexto.

Referências

ANDRADE, M.D. “A utilização do sistema R-Studio e da jurimetria como ferramentas complementares à pesquisa jurídica. **Quaestio Iuris** , vol. 11, no 02, p. 680–692, 2018, doi: 10.12957/rqi.2018.29221.

CHEN, C.L.P; ZHANG, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. **Information Sciences**. 277 (2014) 314-347.

CNJ, Conselho Nacional de Justiça. **Justiça em números 2018: ano base 2017.** Disponível em <https://www.cnj.jus.br/wp-content/uploads/2011/02/8d9faee7812d35a58cee3d92d2df2f25.pdf>. Acesso em <08/11/2022>.

CONSULTOR JURÍDICO. Anuário da Justiça, 16ª. Ed. **Consultor Jurídico**. 2022. Disponível em <https://anuario.conjur.com.br/pt-BR/profiles/78592e4622f1-conjur/editions/anuario-da-justica-brasil-2022>. Acessado em <08/11/2022>.

GARCIA, Gilson Piqueras. Corrupção e Jurimetria. **Cadernos**, [S.l.], v. 1, n. 8, p. 7-34, jan. 2022. ISSN 2595-2412. Disponível em: <https://www.tce.sp.gov.br/epcp/cadernos/index.php/CM/article/view/175>. Acesso em: 10 nov. 2022.

GRAY, M. W. Statistics and the Law. **International Encyclopedia of Statistical Science**. Springer. 2014,

FARIA, J.E. O sistema brasileiro de Justiça: experiência recente e futuros desafios. **Estud. av.**, May/Aug. 2004, vol.18, n.51, p.103-125. ISSN 0103-4014.

LUVIZOTTO, J. C.,; GARCIA, G. P. A jurimetria e sua aplicação nos tribunais de contas: análise de estudo sobre o Tribunal de Contas da União (TCU). **Revista Controle - Doutrina E Artigos**, 2020. *18*(1), 46-73. https://doi.org/10.32586/rcda.v18i1.585

MAIA, M.; BEZERRA, C.A. Análise bibliométrica dos artigos científicos de jurimetria publicados no Brasil. **RDBCI** v. 18 (2020). DOI: https://doi.org/10.20396/rdbci.v18i0.8658889

Microsoft. 2022 Gartner Magic Quadrant for Analytics and Business Intelligence Platforms. **Microsoft Azure**. Disponível em <https://info.microsoft.com/ww-landing-2022- gartner-mq-report-on-bi-and-analytics-platforms.html?LCID=EN-US > Acessado em <12/11/2022>.

MUNAPPY, A.R.; BOSCH, J. e OLSSON, H.H. "Data Pipeline Management in Practice: Challenges and Opportunities", nov. 2020, Acessado: 7 de novembro de 2022. [Online]. Disponível em: https://research.chalmers.se/publication/523476/file/523476_Fulltext.pdf

PONCIANO, V.L.F. "O CONTROLE DA MOROSIDADE DO JUDICIÁRIO: EFICIÊNCIA SÓ NÃO BASTA". **Tribunal Regional Eleitoral-PR**. Disponível em < https://www.conjur.com.br/2015-ago-05/vera-ponciano-eficiencia-nao-basta-combater-morosidade >. Acesso em <08/11/2022>.

SALZANO, J.G.F. Virtualização do Processo: Jurimetria, Inteligência Artificial e Processo Eletrônico no Ordenamento jurídico. **Conhecimento Interativo**, SJP/PR ISSN 1809-3442 v.14, N.1, p. 163-187. jan/jun 2020.

SHAERER, C. The CRISP-DM Model: The New Blueprint for Data Mining. **Journal of Data Warehousing**. Vol 5 Nro. 4, 2000.

STF. Projeto Victor avança em pesquisa e desenvolvimento para identificação dos temas de repercussão geral. **Portal STF**. Disponível em <https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=471331&ori=1>. Acesso em <17/11/2022>.

TJESP. Site do Tribunal de Justiça do Estado de São Paulo para Consulta de Processos. Disponível em < https://www.tjsp.jus.br/Processos>. Acesso em <08/11/2022>.