UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO UNIVERSITÁRIO NORTE DO ESPÍRITO SANTO

DEPARTAMENTO DE COMPUTAÇÃO E ELETRÔNICA

BACHARELADO EM ENGENHARIA DA COMPUTAÇÃO

Thiago Oliveira Lima

Análise do Perfil de Gastos na Legislatura 55 do Senado Federal:Ênfase na Relação Fornecedor-Senador

São Mateus, ES

2020

Thiago Oliveira Lima

Análise do Perfil de Gastos na Legislatura 55 do Senado Federal:Ênfase na Relação Fornecedor-Senador

Monografia apresentada ao Colegiado do Curso de Engenharia de Computação do Departamento de Computação e Eletrônica da Universidade Federal do Espírito Santo, campus São Mateus, como requisito parcial para obtenção do Grau de Bacharel em Engenharia de Computação.

Universidade Federal do Espírito Santo

Departamento de Computação e Eletrônica

Colegiado do Curso de Engenharia de Computação

Orientador: Prof. Silvia das Dores Rissino

São Mateus, ES

2020

Thiago Oliveira Lima
Análise do Perfil de Gastos na Legislatura 55 do Senado Federal:Ênfase na Relação Fornecedor-Senador/ Thiago Oliveira Lima. – São Mateus, ES, 2020-
28 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Silvia das Dores Rissino

Monografia (PG) – Universidade Federal do Espírito Santo
Departamento de Computação e Eletrônica
Colegiado do Curso de Engenharia de Computação, 2020.

1. Apriori. 2. KDD. 3. Mineraçãoa de Dados. 4. Senado Federal. 5. Dados abertos 6. Python. 7. PowerBI I. Universidade Federal do Espírito Santo. II. Análise do Perfil de Gastos na Legislatura 55 do Senado Federal:Ênfase na Relação Fornecedor-Senador

CDU 02:141:005.7

Thiago Oliveira Lima

Análise do Perfil de Gastos na Legislatura 55 do Senado Federal:Ênfase na Relação Fornecedor-Senador

Monografia apresentada ao Colegiado do Curso de Engenharia de Computação do Departamento de Computação e Eletrônica da Universidade Federal do Espírito Santo, campus São Mateus, como requisito parcial para obtenção do Grau de Bacharel em Engenharia de Computação.

Trabalho aprovado. São Mateus, ES, 15 de dezembro de 2020:

Prof. Silvia das Dores Rissino
Orientador

Prof. Luciana Lee
Avaliador 1

Prof. Wilian Hiroshi Hisatugu
Avaliador 2

São Mateus, ES
2020

Dedico esse trabalho a todas pessoas que se envolveram e que foram importantes, de alguma forma, na minha jornada dentro da UFES.

Agradecimentos

A Deus, por me conceder o dom da vida e me dar sabedoria para superar todo os obstáculos encontrados durante o curso. A minha família, por me auxiliar e dar todo o suporte necessário para vencer essa etapa da vida. Aos meus amigos, que tornaram esse processo muito mais divertido e menos árduo. A todos os professores que com muita competência me ensinaram e que de alguma forma me marcaram.

"Não tenha medo de tentar, tenha medo de não tentar e ver que a vida passou e você não se arriscou como deveria.
(Chorão, Charlie Brown Jr.)

Resumo

Com a popularização da internet a quantidade de dados estão em constante crescimento, viabilizando o armazenamento de grandes conjunto de dados. A Mineração de Dados (DM) tem grande importância quando se trata de analisar, interpretar e relacionar esses dados. O processo de descoberta de conhecimento em base de dados, ou KDD, desenvolve métodos e técnicas com intuito de dar sentido aos dados além de auxiliar em estratégias e tomadas de decisão, esse processo tem a DM como a etapa mais importante. Este trabalho apresenta o estudo e análise dos dados abertos do Senado Federal relacionado as CEAPS, especificamente da legislatura 55. Usa-se o processo de KDD, onde na etapa de pré-processamento foi utilizado a linguagem de programação Python com o auxílio da biblioteca pandas, no processo de Mineração de Dados foi utilizado o algoritmo Apriori para gerar regras de associações com índice de confiança 1. E no pós-processamento foi utilizado a ferramenta Power BI para desenvolvimento de um *dashboard*. Permitindo a visualização de informações como a quantidade total de registros, o valor total reembolsado, o senador que mais recebeu reembolso, o fornecedor que mais recebeu pela cota e principalmente constatar a coerência das regras geradas na etapa de Mineração de Dados. O algoritmo Apriori, inicialmente, não se mostrou adequado para o conjunto de dados utilizado, porém, após algumas adaptações, foi possível identificar atividades atípicas.

Palavras-chaves: Apriori. KDD. Mineração de Dados. Senado Federal. Dados abertos. Python. PowerBI.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Considerações Gerais

Com os avanços constantes em Tecnologia da Informação e com a popularização da Internet, a quantidade de dados trocado e armazenado tem aumentado cada vez mais, viabilizando o armazenamento de grandes e múltiplas bases de dados. Fazer a análise desses dados acaba se tornando um trabalho impraticável para o ser humano sem o auxílio de ferramentas computacionais adequadas. Com isso, tem se tornado cada vez mais importante o desenvolvimento de ferramentas automáticas e inteligentes para auxiliar o ser humano nas tarefas de analisar, interpretar e relacionar esses dados, com o intuito de criar e escolher algumas estratégias de ação em cada contexto de aplicação [1].

Com isso, o conceito de Mineração de Dados, do inglês, Data Mining (DM) está se tornando cada vez mais popular como uma ferramenta de descoberta de informação, que pode revelar estruturas de conhecimento e que possam guiar decisões em condições de incertezas [2]. O processo de Descoberta de Conhecimento em Bases de Dados, do inglês, Knowledge Discovery in Databases (KDD) é considerado uma análise automática de dados exploratórios de grandes bancos de dados, que tem como a sua etapa mais importante a de DM [3]. Existem autores que considera DM e KDD como termos sinônimos.

O Senado Federal é uma casa legislativa e uma assembleia deliberativa, suas funções típicas é legislar e fiscalizar. É chamado de câmara alta, assim como a dos Deputados Federais é chamado de câmara baixa. Essa designação surgiu a partir do primeiro parlamento bicameral do mundo, o do Reino Unido e hoje em dia é frequentemente utilizada para distinguir casas legislativas dentro de um sistema bicameral, que é o caso do sistema Brasileiro [4].

Os Senadores da República são os responsáveis por exercer as funções do Senado Federal, eles são eleitos segundo o princípio majoritário como representantes dos estados e do Distrito Federal. Cada estado e Distrito Federal elegem três senadores, totalizando 81, sendo que o mandato tem duração de oito anos, ou seja, duas legislaturas. A renovação da representação é feita, alternadamente por um e dois terços, a cada quatro anos. Além disso, cada senador é eleito com dois suplentes [5].

A atuação de um parlamentar envolve muitas atividades dentro e fora do Congresso, como reuniões, viagens, diligências e contato com a sociedade por diferentes meios. Para auxiliar no mandato os senadores possuem à disposição uma série de serviços e auxílios [6]. Um desses auxílios é a Cotas para Exercício da Atividade Parlamentar dos Senadores (CEAPS) que corresponde a um valor destinado a cobrir despesas relacionadas ao exercício

do mandato. Tais despesas podem ser: passagens, serviços postais, manutenção de escritórios de apoio à atividade parlamentar, hospedagem, combustível, entre outros.

Com o advento da Lei da Transparência que foi sancionada em 2009, a divulgação na internet dos dados em relação a gastos da União, dos estados e dos municípios se tornou obrigatória. Logo, tornou-se possível que qualquer pessoa interessada tenha acesso a esses dados e com isso o processamento e análise de tais dados, seja para fiscalizar as gestões políticas, para fins acadêmicos ou qualquer outra finalidade se tornou mais fácil [7].

Nesse sentido, é possível utilizar técnicas de DM para analisar os dados, que devido ao surgimento da Lei da Transparência, passou a ser divulgado no site do Senado Federal a fim de esclarecer conhecimentos incertos em relação às CEAPS.

1.2 Descrição do Problema

Como identificar atividades atípicas, dentro do conjunto de dados abertos do Senado Federal, que estão relacionadas às Cotas para Exercício da Atividade Parlamentar dos Senadores(CEAPS) na legislatura 55 (período entre 01/02/2015 e 31/01/2019). Como exibir de forma compreensível os dados, relacionados as atividades atípicas, utilizando ferramentas de exibição dos dados.

1.3 Objetivo Geral

Esse trabalho tem como objetivo geral aplicar as etapas do KDD no conjunto dados abertos do Senado Federal relativo a Cota para o Exercício da Atividade Parlamentar dos Senadores(CEAPS) durante à legislatura 55 com o intuito de analisar perfil de gastos, principalmente entre a relação Fornecedor x Senador.

1.4 Objetivos Específicos

Os objetivos específicos deste trabalho estão descritos a seguir:

- Realizar levantamento bibliográfico;
- Estudo das etapas do KDD;
- Pesquisa e definição do conjunto de dados;
- Estudo sobre o Senado Federal e seu conjunto de dados;
- Estudo do Python e da sua biblioteca pandas, assim como do Power BI;
- Aplicação do KDD no conjunto de dados selecionado;

- Estudo e aplicação do algoritmo Apriori para a análise de perfil;
- Apresentação dos resultados no formato de *dashboard*.

1.5 Organização do Trabalho

O trabalho foi organizado em 5 seções. Na Seção 1, a atual, é feita a introdução do trabalho, apresentado o problema, os objetivos gerais e os objetivos específicos. Na Seção 2, é feito um levantamento bibliográfico do método utilizado e também apresentado alguns trabalhos relacionados. A Seção 3 descreve toda a metodologia utilizada no trabalho. Na Seção 4, é mostrado os resultados obtidos. E na Seção 5 são feitas algumas considerações finais e apresentado possíveis abordagens futuras.

2 Levantamento Bibliográfico

2.1 Considerações Iniciais

O termo KDD foi formalizado em 1989 em alusão ao conceito de buscar conhecimento a partir de bases de dados [8]. Uma das definições mais populares é: “KDD é um processo não trivial para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados” [1]. Em outras palavras, o KDD se preocupa com o desenvolvimento de métodos e técnicas para dar sentido aos dados e através disso auxiliar em estratégias e tomadas de decisão.

Um aspecto fundamental que o caracteriza é a maneira como ele é divido em etapas. Existem algumas formas de se dividir essas etapas, cada um com suas vantagens e desvantagens [1]. Uma dessas abordagens divide o processo em 5 etapas conforme a Figura 1, seus nomes e atribuições são [9]:

- Seleção: A primeira etapa do processo tem como objetivo a escolha dos conjuntos de dados referente ao domínio do problema de forma que os resultados contenham informações úteis.
- Pré-processamento: Consiste na etapa onde é feita a limpeza dos dados, ou seja, corrige problemas como a ausência de alguns valores, erros e inconsistências dos dados. Essa etapa requer um cuidado para que os dados não sejam comprometidos.
- Transformação: Essa etapa consiste na adequação e reorganização dos dados fornecidos pelas etapas anteriores de forma que os algoritmos de Mineração de Dados consiga interpretá-los.
- Mineração de Dados: Apesar da importância de todas etapas para o sucesso do processo, essa etapa carrega a maior importância e é considerada o núcleo de todo o processo. É nessa etapa que os dados serão propriamente transformados em informações úteis através de algoritmos apropriados.
- Interpretação/Avaliação: Etapa onde o que foi indicado pela etapa anterior será interpretada e avaliada. A partir dessa análise pode-se descobrir novos fatos, padrões e relacionamentos que podem ser utilizados para tomadas de decisão ou apenas para exibição de resultados.

As 5 etapas do KDD são agrupadas em outras 3 grandes fases que são chamadas de etapas operacionais, apresentadas na Figura 1, são elas: Pré-Processamento, Mineração de Dados e Pós-Processamento [1]:

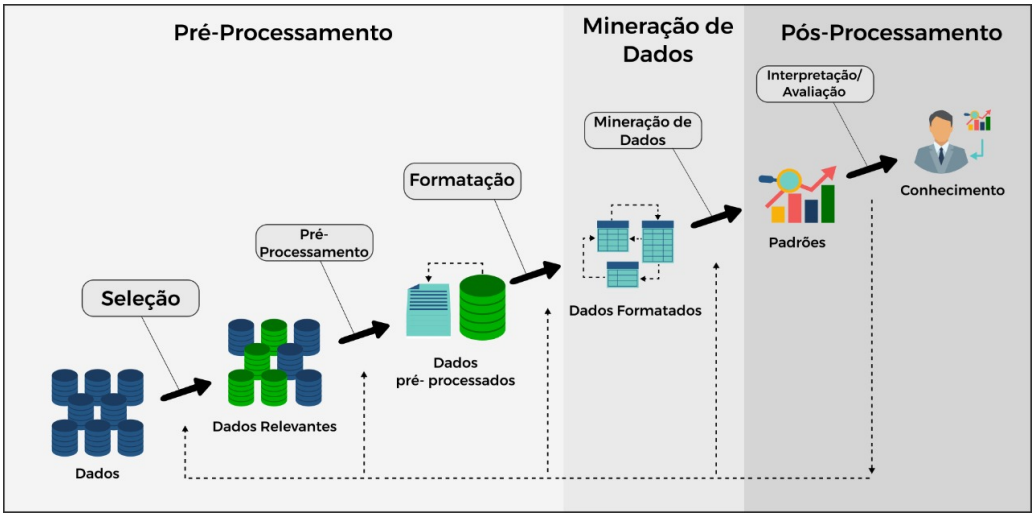

Figura 1 – Etapas do KDD.

- Pré-processamento: Essa etapa abrange as funções relacionadas a adquirir, organizar, tratar e preparar os dados para a etapa da Mineração de Dados. Possui uma relevância imprescindível para o processo de KDD e é determinante para o seu resultado final. Ela abrange a retirada de dados ruidosos (que contenham valores discrepantes do esperado), inconsistentes e incompletos.

- Mineração de dados: Compreende ao processo de busca efetiva de novos conhecimento e úteis a partir dos dados pré-processados. Nessa etapa, através da aplicação de algoritmos, os dados são transformados em informações que, após submetidas a análise e interpretação, são transformadas em conhecimentos para tomadas de decisões. Dentre as atividades que podem ser implementadas na Mineração de Dados, destacam-se a classificação, clusterização, agrupamentos, sumarização.

- Pós-processamento: Essa etapa compreende a visualização, análise e a interpretação do modelo gerado pela Mineração de Dados. Nessa etapa que os dados obtidos são analisados para saber se podem auxiliar em novas alternativas e tomadas de decisão sejam elas automatizadas ou não.

2.2 Trabalhos Relacionados

Os trabalhos apresentados a seguir demonstram conceitos, técnicas utilizadas, aplicações e ferramentas para o processo e Mineração de Dados em conjunto de dados relacionados com política.

Em “Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do Rio Grande do Sul” [10], os autores propõem um estudo de descoberta de conhecimento em uma base de dados das eleições de 2012, para qual analisam os candidatos a vereador de uma região do Rio Grande do Sul, tais dados foram

disponibilizados pelo Tribunal Superior de Justiça. O objetivo do trabalho foi descobrir os motivos que levaram um candidato a vereador dessa região a ser eleito. Para isso, os autores utilizaram técnicas de mineração de dados de árvores de decisão, com o algoritmo J48 e o auxílio da ferramenta WEKA. A árvore de decisão foi capaz de identificar corretamente 88,2% das instâncias (total de 544). Além disso, os autores testaram outros algoritmos disponíveis no WEKA, porém, segundo eles, nenhum mostrou um desempenho melhor que o J48.

Um exemplo de padrão descoberto durante o desenvolvimento do trabalho é que o fator mais importante para o candidato ser eleito é ser político. Por outro lado, se o candidato não for político e possuir somente o ensino fundamental, em grande parte do casos não é eleito. Outro fator relevante é candidatos com idade entre 31 e 56 anos juntamente com um grau de instrução mais elevado (ensino superior completo) na maioria dos casos obtém sucesso na eleição.

Em "Mineração de Dados na Base de Dados Aberta da Câmara Legislativa Federal Brasileira: Ênfase na Análise dos Dados da Legislatura 54 (2011-2013)" [11], é proposto a aplicação do KDD no conjunto de dados abertos da Câmara Legislativa Federal Brasileira para encontrar padrões em gastos feitos pelos partidos políticos durante a legislatura 54 que compreende os anos de 2011 a 2013. Para isso, utilizaram dois algoritmo de Mineração de Dados, um supervisionado (J48) e um não supervisionado (Apriori). Além disso, foi utilizado a ferramenta WEKA.

Alguns padrões observados pelos autores a partir da árvore de decisão gerada pelo algoritmo j48 foram:

- As cotas referentes a manutenção de escritório de apoio à atividade parlamentar, combustíveis e lubrificantes, telefonia e serviço de segurança prestado por empresa especializada, em 2011 e 2013, foram usadas predominantemente pelo PT, independente da época e valor.
- A cota referente a serviço de segurança prestado por empresa especializada em 2011 e 2013, foi gasta predominantemente pelo PSD, independente da época e valor.

Com a utilização dos dois algoritmos os autores concluíram que a utilização do algoritmo Apriori forneceu melhor desempenho quando comparado com os j48 e que, provavelmente, os algoritmos não supervisionados se enquadram melhor para esse tipo de análise.

3 Metodologia

Para a execução desse trabalho, primeiramente foi feito um levantamento bibliográfico em sites onlines, livros, portal Google Acadêmico e Periódicos Capes. O objetivo era encontrar artigos e informações sobre Mineração de Dados e o processo de descoberta de conhecimento em banco de dados ou KDD. Em paralelo ao levantamento, foi feita uma análise de diversos banco de dados do portal de Dados Abertos do Governo e de outras instituições, para definir o conjunto de dados para a aplicação do método desejado.

Após busca e análise, foi decidido trabalhar com os dados referentes aos auxílios pagos para os Senadores Federais do Brasil, em específico as CEAPS. Esse conjunto de dados foi encontrado na seção de Dados Abertos no site do Senado Federal. Como o intuito deste trabalho é aplicar a Mineração de Dados para analisar o pagamento dessas cotas em uma legislatura, foram selecionados os dados dos anos 2015 a 2019, respeitando o período vigente da legislatura 55.

3.1 Ambiente de Dados

Dados são abertos quando qualquer pessoa pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, exigências que visem preservar sua proveniência e sua abertura [12].

O Governo Brasileiro criou em 2012 um portal de dados abertos, onde é centralizado a busca e o acesso aos dados e às informações públicas. Para o Senado Federal, a divulgação de informações aos cidadãos é fundamental e contribui para o enrijecimento das políticas de participação social, de inovação tecnológica e para a integridade pública [13].

No site do Senado Federal é possível acessar conjuntos de dados abertos relacionados às diversas esferas de atuação dessa câmara, o conjunto de dados selecionado para este trabalho foram os dados relativos a um benefício chamado de Cotas para Exercício da Atividade Parlamentar (CEAP) . Para esta análise foram selecionados apenas os dados correspondentes à legislatura 55. Os dados são disponibilizados em formato CSV com os nomes *2015.csv* (utilizado dados a partir 01/02/2015), *2016.csv* (dados do ano todo), *2017.csv* (dados do ano todo), *2018.csv* (dados do ano todo) e *2019.csv* (dados apenas até 31/01/2019).

Cada arquivo é separado nas seguintes colunas: *ANO*, *MES*, *SENADOR*, *TIPO_DESPESA*, *CNPJ_CPF*, *FORNECEDOR*, *DOCUMENTO*, *DATA*, *DETALHAMENTO*, *VALOR_REEMBOLSADO* e *COD_DOCUMENTO*. Com uma particularidade para o ano 2018 que não possui a coluna *COD_DOCUMENTO*. A Figura 2 apresenta um fragmento

ANO	MES	SENADOR	TIPO_DESPESA	CNPJ_CPF	FORNECEDOR	DOCUMENTO	DATA	DETALHAMENTO	VALOR_REEMBOLSADO	COD_DOCUMENTO
2017	1	ACIR GURGACZ	Aluguel de imóve	05.914.650/0001-66	ENERGISA	34079	18/01/2017	Despesa com paga	97	2060286
2017	1	ACIR GURGACZ	Aluguel de imóve	004.948.028-63	GILBERTO PISELO DO NASCIMENTO	001/17	17/01/2017	Despesa com alugı	6000	2057638
2017	1	ACIR GURGACZ	Aluguel de imóve	05.423.963/0001-11	OI MÓVEL S.A.	744526352	18/01/2017	Despesa com paga	418,04	2060285

Figura 2 – Fragmento do conjunto de dados do ano de 2017.

do conjunto de dados para o ano de 2017.

3.2 Etapas do KDD

3.2.1 Pré-Processamento

O conjunto de dados inicial é composto por 5 arquivos referentes aos 5 anos na qual a legislatura 55 possui algum período vigente, o que origina em 125.988 registros. Na etapa de seleção, foi necessário a criação de um algoritmo em Python para extrair os dados que compreende apenas o período da legislatura, ou seja, os dados de 01/02/2015 até 31/01/2019 e junta-los em um conjunto de dados único e com os dados relevantes para o problema.

Após a definição dos dados relevantes, foi feita uma limpeza no conjunto de dados, onde foram eliminados todos os registros que possuíam alguma de suas colunas vazias e também algumas colunas que não eram úteis para o problema. Essa etapa também contemplou a detecção e eliminação de possíveis inconsistências nos dados, um exemplo de inconsistência tratada foi que, em alguns registros, o campo VALOR_REEMBOLSADO possuía números negativos, o tratamento escolhido foi transformar o valor para o seu módulo. Depois dessas etapas, o conjunto de dados tratado ficou reduzido em 104.298 registros.

A próxima etapa, foi fazer a transformação dos dados para facilitar a análise e aplicação do algoritmo de Mineração de Dados. Com isso, campos como SENADOR, TIPO_DESPESA, CPF_CNPJ foram normalizados, de forma que toda ocorrência de um valor x em TIPO_DESPESA, por exemplo, foi substituído para um número inteiro começando do 111, essa mudança foi feita para todos os outros campos citados. Para cada mudança desse tipo foi gerado um arquivo CSV novo com as identificações de cada valor anterior. Dessa forma, pode-se criar um Diagrama de Entidade-Relacionamento (MER) que representasse de que forma os arquivos gerados foram organizados para formar o arquivo original. A Figura 3 apresenta o MER de conexão dos arquivos gerados pelo pré-processamento .

Por exemplo, a Tabela 1 mostra o resultado do pré-processamento do campo TIPO_DESPESA. É possível perceber que em 2017 houve gastos com 7 tipos de despesas diferentes. No arquivo de análise final, todas as ocorrência de , por exemplo, “Serviços de Segurança Privada” foram substituídas por “117”. O mesmo foi feito para Senadores,

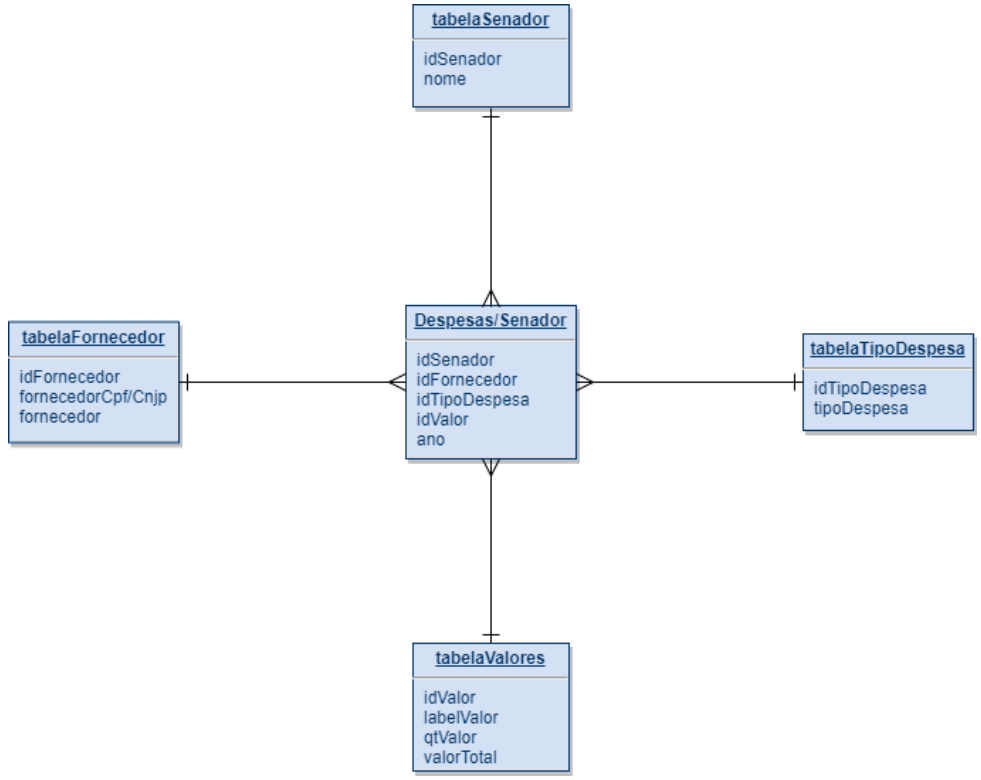

Figura 3 – MER de conexão dos arquivos gerados pelo pré-processamento.

Tabela 1 – Resultado do pré-processamento do campo TIPO_DESPESA

tabelaTipoDespesas	
idTipoDespesa	tipoDespesa
111	Aluguel de imóveis para escritório político, compreendendo despesas concernentes a eles
112	Passagens aéreas, aquáticas e terrestres nacionais
113	Locomoção, hospedagem, alimentação, combustíveis e lubrificantes
114	Aquisição de material de consumo para uso no escritório político, inclusive aquisição ou locação de software, despesas postais, aquisição de publicações, locação de móveis e de equipamentos
115	Divulgação da atividade parlamentar
116	Contratação de consultorias, assessorias, pesquisas, trabalhos técnicos e outros serviços de apoio ao exercício do mandato parlamentar
117	Serviços de Segurança Privada

Fornecedores e Valores.

3.2.2 Mineração de Dados

Nessa etapa do trabalho foi utilizado o algoritmo de Mineração de Dados Apriori, encontrado na biblioteca apyori do Python. O Apriori é um dos algoritmos de regra de

associação mais conhecidos e utilizados. Regras de associação são utilizadas para mostrar relacionamentos entre os itens existentes em um conjunto de dados. A tarefa de associação tem como objetivo descobrir elementos que aparecem junto a outro elemento em um mesmo registro [14].

As regras de associação possuem o formato X => Y que significa que se X está presente em um registro então Y também estará, com um certo grau de frequência. No contexto de regras de associação X e Y são conjuntos de itens, também chamados de itemsets. Na estrutura apresentada, X é o antecedente ou item_base enquanto Y é o consequente ou item_add.

As regras de associação dispõem de um grau de incerteza definido por diversos índices, os mais significativos são os índices de suporte e confiança. O índice de suporte representa a porcentagem de registros que possui os itens de X e Y, representando uma fração dos registros que satisfazem tanto o antecedente quanto o consequente da regra, indicando a relevância da mesma. Já o índice de confiança representa a porcentagem dos registros que possuem o item Y dentro dos registros que possuem o item X, ou seja, representa a quantidade de registros que satisfazem o antecedente e o consequente em relação as que satisfazem apenas o antecedente. A confiança é uma representação da validade da regra, uma confiança baixa indica que poucos registros contendo X também contém Y [15].

Para esse trabalho, o índice de incerteza essencial foi a confiança, afinal o intuito era detectar um perfil de gastos onde a associação entre um fornecedor específico e um senador fosse sempre 100%, ou seja, que um fornecedor só possua um único senador como cliente. O suporte não foi um índice considerado importante por não ter grande influência para análise do trabalho, portanto foi escolhido um valor fixo baseado na quantidade de registros total do banco tratado.

Nesse sentido, foi gerado apenas regras de associação onde o índice de confiança é 1 enquanto o suporte, por não influenciar no resultado final, foi escolhido um valor fixo de 0,0025 afinal em um universo com 104.298 registros, isso irá trazer regras onde pelo menos cerca de 260 registros são contempladas com X e Y.

Após o processamento do algoritmo, foi gerado um conjunto de regras de associação. Elas foram salvas em um arquivo CSV, com as seguintes colunas: *number* que representa o número da regra, *items_base* que é o conjunto de itens antecessores da regra, *item_add* que é o conjunto de itens sucessores da regra, seguidos dos índices das regras *Support*, *Confidence* e *Lift*. A Tabela 2 representa uma amostra das regras de associação geradas.

Tabela 2 – Amostra da tabela de regras de associação geradas

number	items_base	items_add	Support	Confidence	Lift
1	Fornecedor = AZUL LINHAS AÉREAS	Tipo Despesa = Passagens aéreas, aquáticas e terrestres nacionais	0,00807	1,0	3,89360
30	Fornecedor = KAI TOUR AGENCIA DE VIAGENS LTDA	Tipo Despesa = Passagens aéreas, aquáticas e terrestres nacionais	0,00271	1,0	3,89360
57	Fornecedor = Posto São Carlos Ltda	Senador = EDUARDO AMORIM	0,00504	1,0	51,53063
58	Fornecedor = PROPAGTUR - Propag Turismo Ltda	Senador = EDUARDO AMORIM	0,00418	1,0	51,53063

3.2.3 Pós-Processamento

Nessa etapa do trabalho, foi utilizado o Power BI que é uma ferramenta de BI, desenvolvida pela *Microsoft*, lançada em 24 de julho de 2015 e que em 5 anos já dominava o mercado de BI. O Power BI é uma ferramenta completa, onde é possível se conectar com diversas bases de dados de diferentes formatos, fazer análise e criar visualizações de dados, fazendo assim que ele se torne uma ótima opção para auxiliar na tomada de decisão [16].

O PowerBI é divido em três plataformas diferentes: *desktop*, serviço e *mobile*. Cada uma dessas plataformas tem uma característica diferente. O Power BI Desktop, versão utilizada nesse trabalho, é a plataforma voltada para o desenvolvedor e possui algumas funcionalidades voltadas para o desenvolvimento que não estão disponíveis nas demais. O Power BI Serviço é uma coleção de funcionalidades adicionais que podem agregar na utilização da ferramenta em vários aspectos, por exemplo a disponibilidade de um ambiente em nuvem. A plataforma *mobile* possibilita acesso aos *dashboards* produzido através do celular [17]. A versão do Power BI Desktop utilizada nesse trabalho foi a 2.87.1061.0 64-bit (novembro de 2020), mais informações sobre o Power Bi podem ser encontradas no site da Microsoft [18].

O conjunto de dados utilizado nessa etapa do trabalho foi o arquivo CSV resultante da etapa de limpeza e também o arquivo CSV com todas a regras de associação geradas pelo algoritmo Apriori. Esse conjunto foi carregado no Power BI e então foi criado um *dashboard* que permite confirmar que as regras de associação geradas na etapa anterior estão coerentes com os dados originais do problema e também fazer uma série de análises nos dados.

4 Resultados

Neste trabalho analisou-se os dados da legislatura 55. Originalmente, esses dados são fornecidos por ano, portanto foi necessário fazer um filtragem para reunir os dados do período correto. Com esse intuito foi criado um algoritmo em Python com a utilização da biblioteca pandas, onde todos os anos que possuem algum período dentro da legislatura que se desejava analisar foi carregado. Posteriormente, foi extraído apenas os dados dos períodos que compreendiam a legislatura 55, logo depois, esses dados foram unidos em um único arquivo. O conjunto de dados gerado nessa filtragem foi utilizado em todas as etapas do trabalho.

Na etapa de Pré-Processamento, foi construído um conjunto de algoritmos em Python também com a biblioteca pandas, onde cada um tinha o intuito de fazer a transformação de uma coluna do conjunto de dados base, tornando-o adequado para ser utilizado pelo algoritmo de Mineração de Dados. Esses algoritmos geraram diversas saídas, pois cada transformação gerava um arquivo CSV que foi utilizado para relacionar os valores originais com os valores transformados e, além disso, cada transformação aplicada no conjunto de dados base gerava um conjunto de dados novo. Essa estratégia foi utilizada para que possíveis problemas nos algoritmos não acabassem atrapalhando todo o trabalho feito anteriormente, além de que esses conjuntos de dados parcialmente transformados poderiam ser úteis em algum outro momento do trabalho.

Devido a essa estratégia, foi necessário criar uma divisão de pastas para organizar todos esse conjuntos de dados. A Figura 4 mostra a estrutura utilizada, onde na pasta *Context* foi guardada todas as tabelas de relação entre valor original e transformado, a pasta *DbVersions* é onde foi guardada todas as versões do banco de dados após a aplicação de uma transformação e a pasta *Rules* foi gerada para manter as Regras de Associação geradas pelo algoritmo de Mineração de Dados. Essa divisão é feita de forma automática pelo algoritmo criado.

Para a fase de Mineração de Dados, o algoritmo Apriori recebeu como parâmetro de entrada o conjunto dados final da etapa de Pré-Processamento que era constituído por um total de 104.298 registros. Além disso, como mencionado anteriormente, o valor de confiança utilizado foi 1 para que as regras geradas fossem apenas aquelas em que um fornecedor esteja associado a apenas um senador, o parâmetro de suporte foi escolhido como sendo 0,0025 pois em um conjunto com essa quantidade de registros, essa escolha irá trazer regras onde pelo menos cerca de 260 registros são contempladas tanto com o antecedente e o consequente. O parâmetro de tamanho máximo e mínimo utilizado foi 2 para ambos.

Figura 4 – Estrutura de pastas para organização.

Após o processamento do algoritmo, foi gerado um total de 69 regras de associação. Elas foram salvas em um arquivo CSV, com as seguintes colunas: *number* que representa o número da regra, *items_base* que é o conjunto de itens antecessores da regra, *item_add* que é o conjunto de itens sucessores da regra, seguidos dos índices da regras *Support*, *Confidence* e *Lift*.

O conjunto de dados com as regras de associação e o conjunto de dados gerado após a etapa de limpeza foi carregado no Power BI e utilizado para construir o *Dashboard* 5. Através dele, é possível observar uma tabela com o antecedente e o consequente da Regras de Associação gerada na etapa de Mineração de Dados, que foi filtrado apenas para exibir regras que contenham informação Fornecedor x Senador, além disso, é possível visualizar informações de suporte e confiança ao passar o mouse em cima da regra.

É possível observar também informações relacionadas aos dados no geral, por exemplo: quantidade de registros do conjunto de dados utilizado, valor total reembolsado, valor reembolsado por tipo de despesa, valor reembolsado por ano, valor reembolsado por senador, fornecedor que mais recebeu através do CEAP e senador com maior reembolso do CEAP. Além disso, é possível fazer filtros dinâmicos ao interagir com algum dos gráficos existentes no *dashboard*, e trazer informações mais especificas ou utilizar um dos dois filtros de senador e de fornecedor que já estão disponíveis.

Com isso, o *dashboard* disponibiliza uma gama de informações pertinentes para analisar o comportamento do conjunto de dados. E, principalmente, é possível constatar que as regras de associação geradas pela etapa de Mineração de Dados são coerentes e, além disso, observar outras informações relacionadas a essa regra.

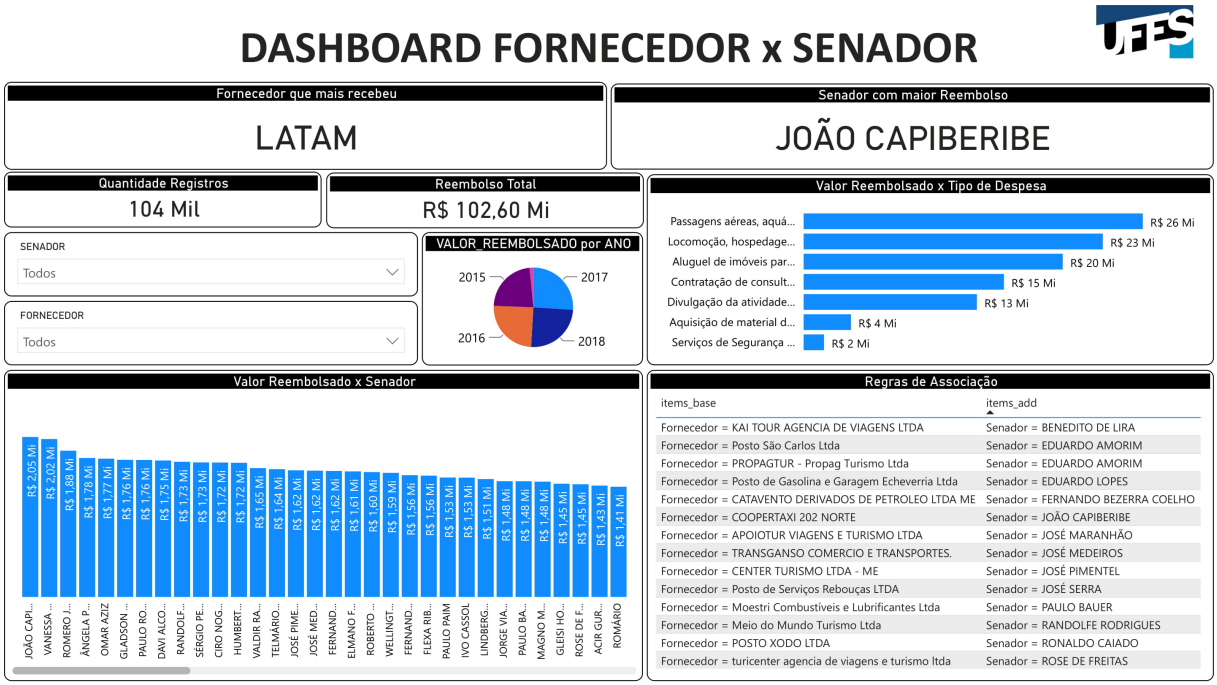

Figura 5 – *Dashboard* Fornecedor x Senador

5 Conclusão e Trabalhos Futuros

O desenvolvimento desse trabalho possibilitou perceber que a linguagem de programação Python, com o auxilio da biblioteca pandas, é uma ferramenta completa e eficiente para a construção de algoritmos voltados para a manipulação de dados. Além disso, a utilização do algoritmo Apriori nesse trabalho, através da biblioteca apyori, é um exemplo de que o Python possui uma diversidade de bibliotecas onde é possível encontrar algoritmos de Mineração de Dados prontos para serem utilizados, fazendo que essa etapa se torne muito menos árdua.

O algoritmo Apriori utilizado na etapa de Mineração de Dados, inicialmente, não era muito adequado ao conjunto de dados do Senado Federal. O algoritmo apresenta resultados mais interessantes em conjuntos de dados na qual os registros possuem uma gama de variáveis não correlacionadas e também que possuam uma quantidade grande de registros. Em algumas tentativas iniciais, as poucas regras geradas pelo algoritmo eram regras consideradas "óbvias", por exemplo: Se o Fornecedor é TAM então o Tipo de Despesa é Passagem Aérea.

Após algumas tentativas de adaptação do algoritmo para o problema, ficou claro uma forma de utilização do mesmo. Tal utilização consistia em explorar a relação Fornecedor x Senador através do índice de Confiança. Dessa forma, tornou-se possível gerar regras de associação que identifica fornecedores que tenham como cliente um único senador, fato que torna esse perfil atípico.

A aplicação do algoritmo como foi feito nesse trabalho, é apenas um ponto inicial para uma análise um pouco mais aprofundada que explorem outras fontes de informações, visto que o fato de um fornecedor ter como cliente apenas um senador não é informação suficiente para conclusões. Porém, foi verificado em uma reportagem, o caso de um Fornecedor que tinha como cliente apenas um Senador e que recebeu uma quantia significante da verba prevista para o CEAP, o que implica que tal linha de pensamento faz sentido e já foi verificado de outras formas. Logo, a descoberta dos perfis feita nesse trabalho é importante para orientar em pesquisas relacionadas a atividades atípicas.

O caso da reportagem mencionada anteriormente não foi constatado pela aplicação do algoritmo nesse trabalho, isso se dá ao fato de que a quantidade de registros dessa relação em especifico é muito pequena para o valor de suporte escolhido na aplicação do algoritmo Apriori. Caso o valor do Suporte seja reduzido, será possível captar tal informação, porém, para esse trabalho não foi escolhido um valor menor com o intuito de reduzir a quantidade de regras geradas pelo algoritmo. Durante o desenvolvimento do trabalho foi constatado que, para o tipo de análise feita, o valor do suporte não influencia

para definir que um perfil seja mais atípico que o outro, visto que pode existir um perfil com baixo suporte, porém com um valor reembolsado muito significativo.

A linguagem de programação Python possui bibliotecas voltadas para visualização de dados, sendo possível utilizar apenas o Python para aplicar o método do trabalho por completo. No entanto, no desenvolvimento deste trabalho foi utilizada a ferramenta Power BI, por ser de fácil compreensão e utilização, por permitir a construção de visualizações mais agradáveis, contendo uma vasta variação de possíveis visualizações, e também por ser uma ferramenta dinâmica.

Através do *dashboard* construído como o Power BI, foi possível verificar, na prática, que as regras de associação geradas pelo algoritmo eram consistentes com o conjunto de dados e também adquirir outras diversas informações, como as mencionadas anteriormente, sobre o conjunto de dados utilizado neste trabalho.

Para trabalho futuros, alguns caminhos podem ser seguidos:

1. Utilizar os algoritmos construídos, fazendo adaptações principalmente relacionadas ao índice suporte de entrada para o Apriori. E avançar na utilização das informações geradas para uma investigação mais aprofundada do perfil.

2. Aproveitar os outros campos que já foram Pré-Processados mas não foram utilizadas neste trabalho, para fazer uma outra abordagem de Mineração de Dados mais adequada para o conjunto de dados utilizado.

3. Fazer uma Seleção de dados diferente, baseado no campo Data e Senador. Poderia concatenar em um só registro todos os reembolsos de uma mesma data e de um mesmo Senador. Após isso, fazer a Mineração de Dados com Apriori e analisar as regras geradas pelo mesmo.

Referências

[1] Ronaldo. Goldschmidt and Emmanuel. Passos. *Data Mining um guia prático*. Elsevier Editora Ltda, 2005. Citado 2 vezes nas páginas 11 e 14.

[2] Sérgio. Côrtes, Rosa. Porcaro, and Sérgio. Lifschitz. Mineração de dados – funcionalidades, técnicas e abordagens. *PUC-RioInf.MCC10/02*, 2002. Citado na página 11.

[3] Salvador. García, Julián. Luengo, and Francisco. Herrera. *Data Preprocessing in Data Mining*, volume 72. Springer. Citado na página 11.

[4] Carla. Mereles. Como funciona o senado? https://www.politize.com.br/senado-como-funciona/, 2017. Politize!. Acesso em: 10 de out. de 2019. Citado na página 11.

[5] Senado. Federal. Entenda a atividade legislativa. https://www12.senado.leg.br/institucional/sobre-atividade. Institucional. Acesso em: 11 de out. de 2019. Citado na página 11.

[6] Agência. Senado. Subsídios, cotas e benefícios dos senadores. https://congressoemfoco.uol.com.br/especial/noticias/subsidios-cotas-e-beneficios-dos-senadores/, 2015. UOL. Acesso em: 11 de out. de 2019. Citado na página 11.

[7] Wikipédia. Lei da transparência. https://pt.wikipedia.org/wiki/Lei_da_Transpar%C3%AAncia. Wikipédia. Acesso em: 11 de out. de 2019. Citado na página 12.

[8] Alfredo. Boente, Ronaldo. Goldschmidt, and Vânia. Estrela. Uma metodologia de suporte ao processo de descoberta de conhecimento em bases de dados. *SEGeT – Simpósio de Excelência em Gestão e Tecnologia*, 2002. Citado na página 14.

[9] Daniel. Teófilo. Kdd – knowlegde discovery in database. https://danielteofilo.wordpress.com/2015/02/16/kdd-knowlegde-discovery-in-database/, 2015. Daniel Teófilo – Tecnologia. Acesso em: 11 de dez. de 2019. Citado na página 14.

[10] Alex. Camargo, Roger. Silva, Érico. Amaral, Milton Heinen, and Francisco. Pereira. Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do rio grande do sul. *Revista Brasileira de Computação Aplicada (ISSN 2176-6649)*, 2016. Citado na página 15.

[11] Vanessa. Formigoni, Icaro. Honorato, and Silvia. Rissino. Mineração de dados na base de dados aberta da câmara legislativa federal brasileira: Ênfase na análise dos dados da legislatura 54 (2011-2013). *Brazilian Journal of Production Engineering*, 2018. Citado na página 16.

[12] Open Knowledge Internacional. What is open? https://okfn.org/opendata/. Open Knowledge Internacional. Acesso em: 21 de nov. de 2019. Citado na página 17.

[13] Senado. Federal. Dados abertos. https://www12.senado.leg.br/dados-abertos. Dados Abertos . Acesso em: 11 de out. de 2019. Citado na página 17.

[14] Lívia. Vasconcelos and Cedric. Carvalho. Aplicação de regras de associação para mineração de dados na web. Technical report, Universidade Federal de Goiás, 2004. Citado na página 20.

[15] Carlos. Dias and Paulo. Rezende. Regras negativas de associaÇÃo em mineraÇÃo de dados. *Revista Eletrônica da Faculdade Metodista Granbery*, 2011. Citado na página 20.

[16] Mateus. Bittencourt. Lei de benford e regras de associação no power bi: Ferramentas estatísticas aplicadas à auditoria, 2020. Citado na página 21.

[17] Karine. Lago and Laennder. Alves. *Dominando o Power BI*. DATAB, 2018. Citado na página 21.

[18] Microsoft. O que é o power bi? https://powerbi.microsoft.com/pt-br/what-is-power-bi/. Citado na página 21.