

**UNIVERSIDADE DO VALE DO ITAJAÍ - UNIVALI**

Escola do Mar, Ciência e Tecnologia

Curso de Ciência da Computação

Disciplina de Trabalho de Conclusão de Curso

**Pré-proposta****FORMULÁRIO DE PRÉ-PROPOSTA**

<b>NOME DO ALUNO:</b> FÁBIO VOLKMANN COELHO	
<b>TÍTULO DO TRABALHO:</b> Extração de relações nas bases do PROCON utilizando KDD	
<b>ÁREA DE PESQUISA:</b> Mineração de dados, Exploração de dados	
<b>E-MAIL DO ALUNO:</b> fabiovcelho05@gmail.com	
<b>TELEFONE COMERCIAL:</b>	
<b>TELEFONE RESIDENCIAL:</b>	
<b>TELEFONE CELULAR:</b> 48996632804	
<b>ENDEREÇO RESIDENCIAL:</b> Amaro antonio vieira 2008 bloco 10 ap 403	
<b>ANO E SEMESTRE DE INGRESSO NO CURSO:</b> 2019-2	
<b>NOME DO ORIENTADOR:</b> Anita Maria da Rocha Fernandes	
<b>E-MAIL:</b> anita.fernandes@univali.br	
<b>TELEFONE DE CONTATO:</b>	
<b>TITULAÇÃO:</b>	<b>ÁREA:</b>
<b>CARGA HORÁRIA NA INSTITUIÇÃO:</b>	
<b>HORÁRIO DE ATENDIMENTO AO ALUNO:</b>	
<b>NOME DO CO-ORIENTADOR:</b>	
<b>E-MAIL:</b>	
<b>TELEFONE PARA CONTATO:</b>	
<b>INSTITUIÇÃO DE ORIGEM:</b>	
<b>TITULAÇÃO:</b>	<b>ÁREA:</b>
<b>HORÁRIO DE ATENDIMENTO:</b>	
<i>Considerando verídicas as Informações fornecidas neste formulário, encaminhamos a Pré-Proposta para avaliação</i>	
<b>ASSINATURA DO ALUNO:</b>	
<b>ASSINATURA DO PROFESSOR ORIENTADOR:</b>	
<b>ASSINATURA DO PROFESSOR CO-ORIENTADOR:</b>	
	DATA: ____/____/____

## **Extração de relações nas bases do PROCON utilizando KDD**

### **1. Introdução**

Os PROCONs são órgãos estaduais e municipais de proteção e defesa do consumidor, criados especificamente para este fim, com competências, no âmbito de sua jurisdição, para exercer as atribuições estabelecidas pela Lei nº 8.078, de 11 de setembro de 1990, e pelo Decreto nº 2.181/97. Os Procons são, portanto, os órgãos que atuam no âmbito local, atendendo diretamente os consumidores e monitorando o mercado de consumo local. Têm papel fundamental na execução da Política Nacional de Defesa do Consumidor (Sistema Nacional de Defesa do Consumidor – SNDC, s.d). O PROCON tem como objetivo, defender os consumidores de práticas abusivas dos fornecedores de bens e serviços nas relações de consumo.

O PROCON, órgão coordenador da Política Estadual de Defesa do Consumidor, tem a atribuição de auxiliar a criação e pleno funcionamento de Procons municipais. Para isso, possui servidores capacitados para prestar informações em quaisquer fases do processo de implantação do órgão municipal. Além disso, o PROCON treina, rotineiramente, integrantes de PROCONs municipais, capacitando-os em relação ao Direito do Consumidor, à instauração e tramitação de procedimentos administrativos (PROCON-MG, 2016).

Assim como outros órgãos do governo, o PROCON disponibiliza os dados coletados de reclamações dos usuários no portal de dados abertos assim como previsto na Constituição Federal e na Declaração Universal dos Direitos Humanos. Em 18 de novembro de 2011 foi sancionada a Lei de Acesso a Informação Pública (Lei 12.527/2011) que regula o acesso a dados e informações detidas pelo governo. Essa lei

constitui um marco para a democratização da informação pública, e preconiza, dentre outros requisitos técnicos, que a informação solicitada pelo cidadão deve seguir critérios tecnológicos alinhados com as “3 leis de dados abertos”. Dentro desse contexto o Portal Brasileiro de Dados Abertos é a ferramenta construída pelo governo para centralizar a busca e o acesso dos dados e informações públicas (Sobre o [dados.gov.br](http://dados.gov.br), s.d).

Em 2009 o especialista canadense em políticas públicas e ativista dos dados abertos David Eaves propôs as seguintes leis que regem os dados abertos governamentais e aceitas pela comunidade e pelo W3C (GALLO, 2012).

1. Se o dado não pode ser encontrado e indexado na Web, ele não existe
2. Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado
3. Se algum dispositivo legal não permitir sua replicação, ele não é útil.

Pessoas leigas no assunto tem dificuldade em extrair informações importante para as mesmas através dos dados disponibilizados, tais como, quais as empresas que mais receberam reclamações, qual cidade que mais reportou problemas, entre outras, portanto, este trabalho servirá para ajudar as pessoas a responderem algumas dessas perguntas.

Para responder tais perguntas, será utilizado o processo de KDD (*Knowledge Discovery in Databases*). O termo KDD – Knowledge Discovery in Databases, foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de base de dados (TEÓFILO,2015). O principal objetivo do processo KDD é extrair informações dos dados. Isso é feito utilizando algoritmos de mineração de dados, reconhecimento de padrões válidos, úteis e compreensivos,algoritmos de investigação de e dados, entre outros, para identificar o que é considerado conhecimento. O processo para extração de conhecimento possui várias etapas, não trivial, interativo e iterativo. Não trivial alerta para a complexidade normalmente presente na execução de processos de

KDD (TEÓFILO,2015), iterativo porque o analista pode intervir nas atividades, iterativo porque prevê uma sequência de atividades onde o resultado de uma etapa depende da outra. Cada etapa do processo pode ser repetida inúmeras vezes (MOURA,2019). O KDD consiste no conjunto de ações de exploração e análise de dados, envolvendo a seleção, pré-processamento, limpeza, transformação, data mining, interpretação e avaliação.

1. Seleção: esta etapa consiste em selecionar uma ou mais bases de dados que farão parte da análise. As fontes da base podem ser variados e de formatos diferentes.

2. Pré-processamento: esta etapa consiste em fazer a verificação da qualidade dos dados armazenados, nesta etapa é necessário tratar os dados disponibilizados, fazendo a limpeza dos mesmos, esse processo irá detectar e tratar dados incompletos, incorretos, imprecisos ou irrelevantes e, em seguida, substituir modificar ou excluir os dados. Afim de conseguir uma base de dados de qualidade que passe pelos 5 critérios de qualidade de dados:

Validade: O grau de em que os dados estão em conformidade com a regra de negócio, que possui restrição de tipo de dados, intervalo, obrigatoriedade, exclusivas, de associação de conjunto e de chave estrangeira.

Acuracidade: O grau de conformidade entre os dados e um valor padrão ou real.

Completude: O grau em que todas as medidas necessárias são conhecidas. A incompletude é quase impossível de corrigir com a metodologia de limpeza de dados: não se pode inferir fatos que não foram capturados quando os dados em questão foram inicialmente registrados.

Consistência: O grau em que um conjunto de medidas é equivalente em todos os sistemas. A inconsistência ocorre quando dois itens de dados no conjunto de dados se contradizem: por exemplo, um cliente é registrado em dois sistemas diferentes como

tendo dois endereços atuais diferentes e apenas um deles pode estar correto. A correção da inconsistência nem sempre é possível: requer uma variedade de estratégias - por exemplo, decidir quais dados foram registrados mais recentemente, qual fonte de dados provavelmente será mais confiável (o último conhecimento pode ser específico para uma determinada organização) ou simplesmente tentar encontrar a verdade testando os dois itens de dados (por exemplo, ligando para o cliente).

**Uniformidade:** O grau em que um conjunto de medidas de dados é especificado usando as mesmas unidades de medida em todos os sistemas. Em conjuntos de dados agrupados de diferentes localidades, o peso pode ser registrado em libras ou quilos e deve ser convertido em uma única medida usando uma transformação aritmética.

Existem diversas técnicas para realizar a limpeza dos dados a fim de ter uma base de qualidade, entre elas, apagar espaços em branco dos dados, tratar dados vazios, converter texto em número, remover dados duplicados, realçar erros, criar um padrão de texto, verificar ortografia, escalonar/transformar o dado, normalização, entre outras. Somente após todos os dados serem tratados será possível começar a análise de dados.

3. **Transformação:** esta etapa consiste em aplicar técnicas de transformação de dados, tais como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados.

4. **Data Mining:** esta etapa consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo (1) verificar uma hipótese, (2) descobrir novos padrões de forma autônoma. Além disso, a descoberta pode ser dividida em: preditiva e descritiva. Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto (MOURA, 2019). Alguns desses modelos são:

Algoritmo Genético, Redes Neurais, Árvores de Decisão, Regras de Associação, Análise de Vizinhaça, entre outros.

5. Interpretação e Avaliação: esta etapa consiste em interpretar o resultado obtido, avaliar padrões descobertos, visualização dos padrões extraídos, remoção de padrões irrelevantes ou redundantes e tradução de padrões úteis em termos inteligíveis pelos usuários e onde é feita a documentação dos resultados obtidos e repassadas para as partes interessadas. Algumas formas de visualização de dados são: gráficos, grafos, entre outros.

Este trabalho utilizará do modelo de KDD para a descoberta de dados relacionados ao PROCON, descobrindo padrões e relações entre os dados antes não conhecidos, ou difíceis de perceber, gerando mais valor ao conjunto de dados. Tais como a relação entre quais as empresas com mais reclamações entre as cidades registradas. Qual empresa teve mais reclamações resolvidas.

## **2. Objetivos do Trabalho**

### **2.1. Objetivo Geral**

Aplicar técnicas de KDD para analisar a base de dados o PROCON e inferir sobre as reclamações que este órgão recebe.

### **2.2. Objetivos Específicos**

- Descobrir novos conhecimentos utilizando o KDD nos dados do PROCON.
- Tratamento dos dados de CEP do Brasil.
- Descobrir novos conhecimentos utilizando o KDD nos dados do PROCON e CEPs juntos.
- Relação entre reclamações resolvidas e as empresas.
- Quais as empresas com mais reclamações por cidade

### **3. Escopo do Trabalho**

O trabalho irá abordar sobre os dados do PROCON em relação às suas reclamações gerando novos conhecimentos sobre os mesmos. Inicialmente será realizada a junção dos dados. Em seguida será utilizado o processo de KDD para gerar conhecimento com os dados. O mesmo será feito com o dados dos Ceps do Brasil, o qual os dados do procon possui, para juntar ambos e gerar um novo processo de KDD com esses novos dados.

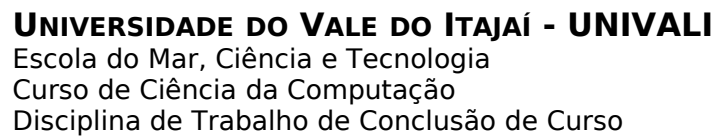
### **4. Plano de Trabalho**

Estudo sobre os dados do PROCON, como o mesmo está estruturado, quais o dados que possuem classificações, fazer a junção de todos os dados em uma única tabela para iniciar o processo de KDD.

Estudo sobre o estado da arte em relação ao KDD, quais os métodos mais utilizados, se os mesmos são possíveis de serem utilizados na minha base. Após a pesquisa sobre o mesmo, realizar os processos do KDD em relação a base de dados. Após concluir o processo de KDD dos dados, será feita uma busca pelos dados dos CEPs do Brasil para complementar os dados da base do PROCON. Com os dados dos CEPs será feita a limpeza dos dados caso seja necessária, fazendo a união das duas bases de dados. Após a união será feita outra rodada do processo de KDD para verificar se novos conhecimentos serão gerados.

### **5. Recursos Necessários (Fonte Times 14 em Negrito e Maiúscula)**

Será necessário um notebook/desktop com no mínimo 16Gb de memória RAM, pois para gerar os gráficos e tratar os dados é utilizada muita memória RAM.

[illegible]



Etapas	Meses				
	dez.	jan.	fev.	mar.	abr.
Redação do rascunho do TCC					
Entrega do rascunho do TCC					
Preparação da defesa pública					
Defesa pública					
Ajustes no relatório final do TCC					

## 7. Orçamento

Item	Quantidade	Valor unitário (R\$)	Valor Total (R\$)
<b>Estimativas para recursos não humanos</b>			
Impressão dos relatórios do TCC	300	0,15	45,00
Encadernação do TCC	4	5	20,00
<b>Reserva de contingência</b>			
Dano no equipamento de trabalho			500,00
<b>TOTAL</b>			565,00

Não haverá gasto com recurso permanente pois o autor já possui os recursos.

## 8. Referência Bibliográfica

\_\_\_\_\_. **Sistema Nacional de Defesa do Consumidor - SNDC.** Disponível em: <<https://www.justica.gov.br/seus-direitos/consumidor/a-defesa-do-consumidor-no-brasil/anexos/sistema-nacional-de-defesa-do-consumidor-sndc>>. Acesso em: 15 mar. 2020.

PROCON-MG, Programa Estadual de Proteção e Defesa do Consumidor do Estado de M.G. **PROCONS MUNICIPAIS Criação, Implementação e Atribuições.** 2016. Disponível em: <<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwjmnq7atp3oAhWELLkGHWBIB7AQFjAAegQIAxAB&url=https%3A%2F%2Fwww.mpmg.mp.br%2Fflumis%2Fportal%2Ffile%2FfileDownload.jsp%3FfileId%3D8A91CFA956B144440156B3D12BE50D78&usg=AOvVaw3gPFHVK35h9dLY8EtMgX9->>. Acesso em: 15 mar.2020.

\_\_\_\_\_. **Sobre o dados.gov.br.** Disponível em: <<http://www.dados.gov.br/pagina/sobre>>. Acesso em: 15 mar. 2020

GALLO, Fernando. **Dados Abertos: três leis e oito princípios.** 2012. Disponível em: <<https://politica.estadao.com.br/blogs/publicos/dados-abertos-tres-leis-e-oito-principios/>>. Acesso em: 15 mar. 2020.

SARFIN, Rachel Levy. **5 Characteristics of Data Quality.** 2019. Disponível em: <<https://blog.syncsort.com/2019/07/data-quality/5-characteristics-of-data-quality/>>. Acesso em: 06 abr. 2020.

ELGABRY, Omar. **The Ultimate Guide to Data Cleaning.** 2019. Disponível em: <<https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>>. Acesso em: 05 abr. 2020.

\_\_\_\_\_. **Data cleansing.** Disponível em: <[https://en.wikipedia.org/wiki/Data\\_cleansing](https://en.wikipedia.org/wiki/Data_cleansing)>. Acesso em: 05 abr. 2020.

\_\_\_\_\_. **Data Analysis - Process.** Disponível em: <[https://www.tutorialspoint.com/excel\\_data\\_analysis/data\\_analysis\\_process.htm](https://www.tutorialspoint.com/excel_data_analysis/data_analysis_process.htm)>. Acesso em: 05 abr. 2020.

TEÓFILO, Daniel. **KDD - Knowledge Discovery in Database.** Disponível em: <<https://danielteofilo.wordpress.com/2015/02/16/kdd-knowledge-discovery-in-database/>>. Acesso em: 26 abr. 2020.

MOURA. Karina. **Ciclo de vida dos dados #1.** Disponível em: <<https://medium.com/@kvmoura/kdd-process-9b8e3062142>>. Acesso em: 26 abr. 2020.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37, 15 Mar. 1996.

STOROPOLI, José Eduardo. **O uso do Knowledge Discovery in Database (KDD) de informações patentárias sobre ensino a distância: contribuições para instituições de ensino superior**. 2016. 194 f. Tese( Programa de Pós-Graduação em Administração) - Universidade Nove de Julho, São Paulo.

BARANAUSKAS, José Augusto. **Extração de Conhecimento & Mineração de Dados & Mineração de Dado**. Disponível em: <<http://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-KDD-DM.pdf>>. Acesso em: 26 abr. 2020.

\_\_\_\_\_. **Data Mining**. Disponível em: <[https://en.wikipedia.org/wiki/Data\\_mining](https://en.wikipedia.org/wiki/Data_mining)>. Acesso em: 26 abr. 2020.

\_\_\_\_\_. **KDD - Knowledge Discovery in Databases**. Disponível em: <<https://www.javatpoint.com/kdd-process-in-data-mining>>. Acesso em: 24 mai. 2020.

AMEDEO NAPOLI. Chapter 41 - A SMOOTH INTRODUCTION TO SYMBOLIC METHODS FOR KNOWLEDGE DISCOVERY. **Handbook of Categorization in Cognitive Science**, c. 41, p. 913-933, 9 Mai. 2007.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, **The KDD Process for Extracting Useful Knowledge from Volumes of Data**. Disponível em: <<https://dl.acm.org/doi/pdf/10.1145/240455.240464>>. Acesso em 24 mai. 2020.

Usama Fayyad, Paul Stolorz, Data mining and KDD: Promise and challenges, **Future Generation Computer Systems**, v. 13, 1.2-3 p. 99-115, 5 Mar. 1997.

Obs minha <https://sci-hub.tw/https://www.sciencedirect.com/science/article/abs/pii/S0167739X97000150>