UNIVERSIDADE FEDERAL DO ACRE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ADAPTAÇÃO E AVALIAÇÃO DO MÉTODO TRAINING-TEST SLIDING VALIDATION EM EXPERIMENTOS COM ALGORITMOS PREDITIVOS

RIO BRANCO
2019

MATEUS DA SILVA COSTA

ADAPTAÇÃO E AVALIAÇÃO DO MÉTODO TRAINING-TEST SLIDING VALIDATION EM EXPERIMENTOS COM ALGORITMOS PREDITIVOS

Projeto de monografia apresentado como exigência parcial para obtenção do grau de bacharel em Sistemas de Informação da Universidade Federal do Acre.

Prof. Orientador: Dr. Manoel Limeira de Lima Júnior Almeida.

RIO BRANCO

2019

LISTA DE FIGURAS

SUMÁRIO

1 APRESENTAÇÃO

A revolução tecnológica do final do século XX e início do século XXI propiciou um aumento considerável na capacidade de processamento e armazenamento de dados, isso aliado a diminuição do custo de aquisição possibilitou uma popularização do uso de computadores.

De acordo com Han, Kamber e Pei (2012), essa popularização da informática associada as grandes redes de computadores, resultou em um explosivo aumento na quantidade de dados disponíveis, logo a capacidade de analisar esses dados precisou crescer na mesma intensidade. Para isso, atualmente existem diversas ferramentas e algoritmos que facilitam esse processo.

No contexto de automatização da análise de bases de dados, o processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD) é utilizado em larga escala objetivando a descoberta de padrões de relacionamento ou de comportamento sobre esses dados. Segundo Hand, Mannila e Smyth (2001), no centro desse processo está a aplicação de métodos e algoritmos para descobrir padrões úteis dentro do conjunto de dados.

Um dos tipos de algoritmos mais utilizados em mineração de dados são os de classificação, principalmente quando se busca prever alguma característica, tal como o desenvolvedor mais apropriado, dentro de uma equipe de desenvolvimento, para avaliar uma solicitação de *pull request* (LIMA JÚNIOR, 2017). A aplicação de um determinado algoritmo pode resultar em um conhecimento válido e útil, todavia

isso depende dos acertos obtidos nos testes com base na aplicação do modelo obtido no treino. A taxa de acerto de um algoritmo é determinada por um valor chamado de acurácia, de modo que quanto maior esse valor melhor é seu desempenho.

Há outras métricas de se determinar a qualidade de um algoritmo de mineração de dados. As métricas são extraídas a partir dos métodos de avaliação dos algoritmos. No contexto de utilização de tarefas preditivas em bases de dados com características temporais, os atributos temporais, caso não sejam considerados, podem conduzir a previsões errôneas onde dados do futuro poderiam acabar sendo utilizados para prever dados do passado (LIMA JÚNIOR, 2017).

O método de avaliação *training-test sliding validation* (TTSV) proposto por Lima Júnior (2017), busca preservar a ordem cronológica de bases de dados com características temporais. Esse método já foi implementado por Lima (2017) na forma de um *plugin* para a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) e posteriormente evoluído por Costa (2018).

A partir da versão 3.2 o WEKA possui um modulo chamado *experimenter*, esse ambiente “permite ao usuário criar, executar, modificar e analisar experimentos de uma maneira mais conveniente do que é possível ao processar os esquemas individualmente” (BOUCKAERT et al., 2016, p. 61). Nesta área da ferramenta é possível selecionar vários algoritmos para testar o desempenho em uma ou mais bases de dados, todavia no caso de existência de atributos temporais o *plugin* que implementa o método de avaliação *training-test sliding validation* não pode ser utilizado, pois foi desenvolvido apenas para o ambiente *explore*.

2 PROBLEMA DA PESQUISA

A ferramenta WEKA possui a possibilidade de realizar experimentos com bases de dados no ambiente *experimenter*. O pesquisador de dados pode selecionar algoritmos dentre os que a ferramenta oferece e executar a tarefa de classificação sobre várias bases de dados de modo que possa, ao final da execução, avaliar o desempenho dos algoritmos sobre as bases de dados.

Os experimentos se tonam importantes porque a acurácia dos diferentes algoritmos pode ser facilmente comparada, tornando possível verificar qual o melhor algoritmo para uma ou várias bases. O *plugin* que implementa o método de avaliação TTSV não está disponível para esse ambiente, o que dificulta a realização de vários experimento.

Essa limitação torna difícil a realização de experimentos em bases de dados com características temporais, uma vez que na atual implementação do *plugin* só é possível executar um algoritmo por vez em uma única base de dados, tornando o trabalho do pesquisador de dados menos eficiente do que se estivesse utilizando o ambiente *experimenter*.

Com base nisso, questiona-se: Como tornar possível a realização de experimentos com várias bases de dados com atributos temporárias?

3 OBJETIVOS DA PESQUISA

Essa seção apresenta o objetivo geral do trabalho e os objetivos específicos.

3.1 OBJETIVO GERAL

O objetivo geral do trabalho é adaptar e avaliar o *plugin Training Test Sliding Validation* para permitir a comparação de diferentes algoritmos de classificação para várias bases com características temporais.

3.2 OBJETIVOS ESPECÍFICOS

Visando alcança o objetivo geral deste trabalho se define os seguintes objetivos específicos:

a) Analisar trabalhos anteriores relacionados ao método janela deslizante;
b) Levantar os requisitos para a implementação da variação do método de avaliação da janela deslizante;

c) Analisar o código fonte já existente do método de avaliação;

d) Implementar outros métodos de avaliação e testes estatísticos;

e) Realizar experimentos com a variação desenvolvida em bases de dados com características temporais.

4 JUSTIFICATIVA DA PESQUISA

De acordo com Han, Kamber e Pei (2012), o processo de KDD é dividido em sete etapas, sendo a mineração de dados a quinta. Nessa etapa é aplicado um algoritmo específico de forma a extrair padrões na base de dados (FAYYAD; PIATESTSKY-SHAPIRO; SMYTH 1996a), todavia existe uma grande quantidade de algoritmos disponíveis de modo que é possível que um pode ser melhor que outro para uma determinada base de dados.

É comum a utilização de vários algoritmos visando obter valores de acurácia melhores, em Rodrigues (2018) utiliza-se os algoritmos NB (*Naive Bayes*), IBK (*K-nearest neighbours classifier*) e J48 (*Decision Tree*) para prever ocorrências policiais na cidade de Rio Branco sendo que no contexto do trabalho o algoritmo que apresentou melhor desempenho foi o NB. Em Lima Júnior (2017) e Costa (2018) também são utilizados vários algoritmos, nos três trabalhos existe característica cronológica na base de dados e foi utilizado o *plugin training-test sliding validation*.

Como visto em trabalhos relacionados, a mineração de dados muitas vezes aplica muitos algoritmos visando melhorar os níveis de acurácia, dando assim maior credibilidade a modelo obtido. O ambiente *experimenter* da ferramenta WEKA facilita a utilização de vários algoritmos e várias bases, porém não há implementação do para o método janela deslizante.

Neste sentido, o trabalho justifica-se pela necessidade de evolução do *plugin* já existente, de modo a implementar um ambiente semelhante ao *experimenter,* o que seria um facilitador para a execução de trabalhos acadêmicos e científicos que precisem extrair conhecimento em bases de dados temporais.

5 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são contextualizados os conceitos que fundamental o trabalho.

2.1 *KNOWLEDGE DISCOVERY IN DATABASES* (KDD)

O grande número de sistemas operando nos mais diferentes contextos e lugares geram cada vez mais dados, o interesse de se extrair algum conhecimento sobre essas bases de dados fomenta uma série de esforços que vão desde novas teorias computacionais até novos aparatos ferramentais. Dentre esses esforços destaca-se o campo de KDD que se concentra na definição de técnicas e métodos para dar sentido aos dados (FAYYAD; PIATETSKI-SHAPIRO; SMYTH, 1996b).

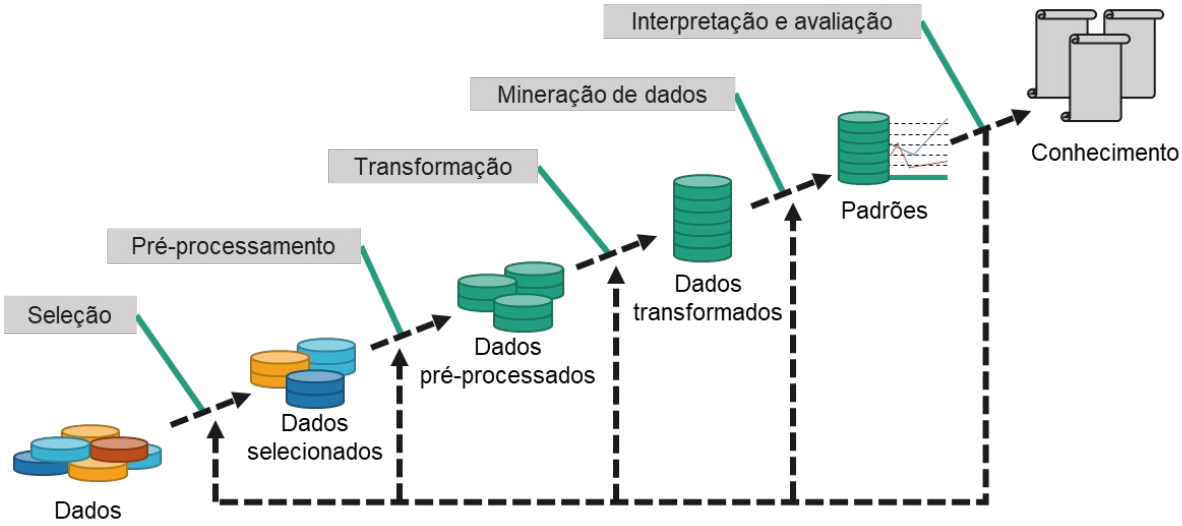
Para Dunham (2003) com um processo, o KDD deve ter como entrada dados e deve gerar como saída alguma informação útil. Fayyad, Piatetski-shapiro e Smyth (1996b) afirmam que esse processo pode ser dividido em cinco etapas e que podem ser definidas da seguinte maneira:

1. **Seleção:** Primeiramente se deve entender o domínio ao qual os dados estão atrelados, sendo isso de fundamental importância para a interpretação e geração de conhecimento. Na etapa de seleção deve-se obter e selecionar os dados, eles podem ser oriundos de diferentes repositórios, bases de dados ou arquivos.
2. **Pré-processamento:** Tendo em vista que os dados podem ser obtidos de diferentes repositórios e em diferentes formatos, alguns dados podem possuir métricas ou unidades de medida diferentes se tornando necessário corrigir ou remover dados errôneos;
3. **Transformação:** Nesta etapa os dados devem ser convertidos para um mesmo formato visando facilitar o processamento, a redução de dados é muito utilizada para reduzir o número de possibilidades para os valores;
4. ***Data Mining*:** Nesta fase ocorre a aplicação de algoritmos sobre a base dados já limpa e formatada, busca-se a obtenção de padrões;
5. **Interpretação ou avaliação:** A última etapa tem por objetivo a verificação do significado dos dados, busca-se determinar se eles contribuem para o problema e qual a qualidade dos padrões descobertos, são utilizadas métricas obtidas como resultado das aplicações dos algoritmos.

Uma representação grafica das etapas de KDD está apresentada na Figura 1, nela fica evidente a caracteristica incremental do processo, onde cada etapa serve de entrada para a próxima, outra caracteristica desse processo é apossibilidade de voltar a uma etapa anterior.

Os resultados obtidos da execução do processo se tornam relevantes se apresentarem vantagens para a organização (WITTEN; FRANK; HALL, 2011). O processo de KDD pode ser aplicado em diversos contextos e áreas diferentes desde que se tenha um grande volume de dados, conforme exemplifica Fayyad, Piatetsky-Shapiro e Smyth (1996b) o KDD pode ser aplicado a áreas relacionadas a ciência, *Marketing*, investimentos, detecção de fraude, indústria e telecomunicações.

Figura 1 - Representação das etapas que compõem o processo de KDD



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996b).

O processo de KDD se torna importante por facilitar a descoberta de conhecimento em grades bases de dados, o que se fosse realizado através de análise manual levaria uma dispendiosa quantidade de tempo e talvez não fosse realizado com a qualidade necessária.

2.2 MINERAÇÃO DE DADOS

Segundo Witten, Frank e Hall (2011) a mineração de dados pode ser definida como a busca por padrões em dados. Esse termo é tratado por muitos como sinônimo do processo de KDD porém a mineração de dados pode ser considerada como parte do processo de KDD sendo que nessa etapa busca-se a identificação de padrões válidos e que apresentem alguma utilidade compreensível (HAN; KAMBER; PEI, 2011; FAYYAD; PIATETSKI-SHAPIRO; SMYTH, 1996b).

Caracterização, discriminação, mineração de padrões, associações, correlações, classificação e regressão são exemplos de tarefas que podem ser

realizada com a mineração de dados, segundo Han, Kamber e Pei (2012) essas tarefas podem ser divida em um duas categorias: descritivas e preditivas.

As tarefas descritivas também chamadas de não supervisionadas focam na busca por relacionamento entre os dados. Esse tipo de tarefa não necessita de um atributo alvo, logo não há a necessidade de realizar categorização, ao invés disso a estratégia utilizada nesse tipo de mineração envolve a tentativa de medir a similaridade entre os dados, exemplos comumente utilizados são as tarefas de regra de associação e agrupamento (CIOS et al., 2007; HAN; KAMBER; PEI, 2012).

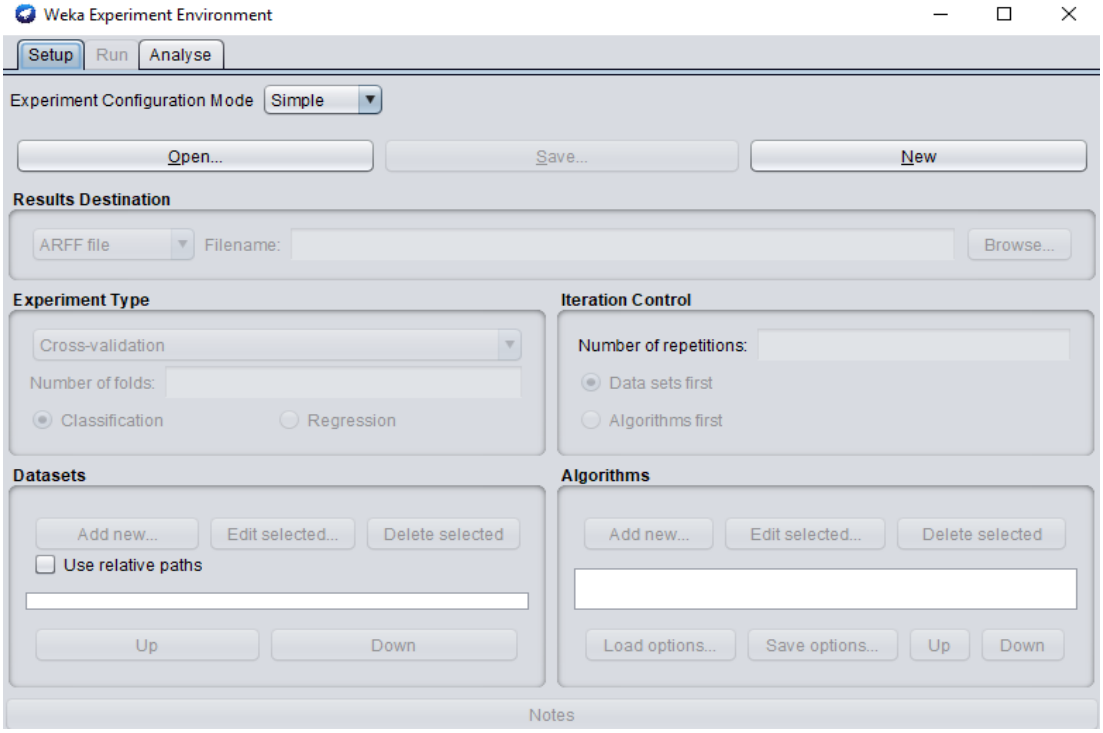
Segundo Hand, Mannila e Smyth (2001) a principal diferença entre as tarefas preditivas e descritivas é o fato de que a primeira tem como objetivo prever um único atributo. Para realizar essa previsão é necessário que haja um conjunto de dados associado ao atributo alvo, de modo que os dados possam ser classificados de acordo com o mesmo (CIOS et al., 2007). Pode-se citar como exemplos de tarefa preditiva classificação e regressão.

2.3 FERRAMENTA WEKA

O Waikato Enviroment for Knomledge Analysis (WEKA) é uma das mais populares ferramentas de descoberta de conhecimento, possui uma coleção de algoritmos de *machine learning* voltados para a mineração de dados. A ferramenta foi desenvolvida na Universidade de Waikato na Nova Zelândia e é distribuído nos termos da *General Public License* (GNU) sendo por tanto um software de código aberto, escrito na linguagem Java (WITTEN; FRANK; HALL, 2011; DEAN, 2014).

Os vários algoritmos de aprendizado disponibilizados pela ferramenta podem ser aplicados sobre uma base de dados com facilidade, além disso o WEKA permite a transformação das bases de dados assim como a análise dos resultados após a mineração (WITTEN; FRANK; HALL, 2011; HALL et al., 2009)

Figura 2 - Interface *Experimenter* da ferramenta WEKA



Fonte: elaboração própria.

A principal interface de usuário do WEKA é a denominada "*Explorer*" segundo Hall et al. (2009) nessa interface há vários painéis correspondentes a diferentes tarefas de mineração assim como ao pré e pós-processamento. Há outras duas interfaces na ferramenta: a "*Knowledge Flow*", voltada para processamento distribuído, e a "*Experimenter*" voltado para a realização de experimentos com as técnicas de classificação e regressão.

A Figura 2 trás a interface *Experimenter,* nela é possível ver três abas a "*Setup*", onde o experimento é ajustado definindo por exemplo as bases de dados e os algoritmos que serão usados, a segunda aba "*Run*" exibe o status da execução e a última aba denominada "*Analyse*" auxilia o pesquisador de dados na análise dos resultados obtidos. Ainda sobre a interface *experimenter* Witten, Frank e Hall (2011, p. 405) afirmam:

> A interface *Experimenter* permite que você automatize o processo, facilitando a execução de classificadores e filtros com diferentes configurações de parâmetros em um conjunto de conjuntos de

> dados, para coletar estatísticas de desempenho e executar testes de significância (tradução nossa).

A interface *Experimenter* é de fundamental importância pra esse trabalho devido ao fato de que sua junção com o *plugin* TTSV facilitaria o trabalho de pesquisadores de dados que trabalham com bases de dados com características temporais.

6 PROCEDIMENTOS METODOLÓGICOS

O trabalho a ser desempenhado pode ser classificado em diversas formas tal como pela sua natureza, abordagem, objetivo e delineamento (WAZLAWICK, 2009; GIL, 2008).

No que se refere a sua natureza, pode-se classificá-lo como um trabalho original, uma vez que se busca fazer algo novo de modo que o atual *plugin* seja expandido para possibilitar a realização de experimentos mais robustos.

No contexto de abordagem do problema, este trabalho pode ser considerado como uma pesquisa qualitativa, já que consiste em analisar de forma interpretativa os dados.

Quanto aos objetivos, o trabalho pode ser classificado como uma pesquisa exploratória, tendo em vista que o delineamento adotado é o de caso de uso, buscando-se assim expandir a ferramenta WEKA conforme os objetivos estabelecidos.

A pesquisa cujo qual esse projeto almeja será organizada em quatro etapas:

a) A primeira etapa do trabalho consiste no estabelecimento dos requisitos necessários para a implementação da evolução do *plugin*.

b) Na segunda etapa, será realizada a implementação da evolução do *plugin* seguindo o que foi estabelecido como requisito na etapa anterior, para tal será

analisado o atual código do *plugin* e da ferramenta WEKA, a codificação se dará através da linguagem de programação Java;

c) Na terceira etapa será realizada a escolha de uma ou mais bases de dados que apresentem características cronológicas;

d) Na penúltima e quarta etapa serão realizados experimentos utilizando a versão evoluída do *plugin* sobre as bases de dados escolhidas;

e) Por fim, a quinta etapa, consistirá na interpretação e análise dos resultados obtidos através do processo realizado.

7 ESBOÇO DOS CAPÍTULOS E SEÇÕES

O trabalho de conclusão de curso seguirá a seguinte estrutura:

1 INTRODUÇÃO

1.1 PROBLEMA DA PESQUISA

1.2 OBJETIVOS DA PESQUISA

1.2.1 OBJETIVO GERAL

1.2.2 OBJETIVOS ESPECÍFICOS

1.3 JUSTIFICATIVA DA PESQUISA

1.4 METODOLOGIA

1.5 ORGANIZAÇÃO DA PESQUISA

2 FUNDAMENTAÇÃO TEÓRICA

2.1 *KNOWLEDGE DISCOVERY IN DATABASES (KDD)*

2.2 MINERAÇÃO DE DADOS

2.3 FERRAMENTA WEKA

2.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

3 IMPLEMENTAÇÃO DA NOVA VERSÃO

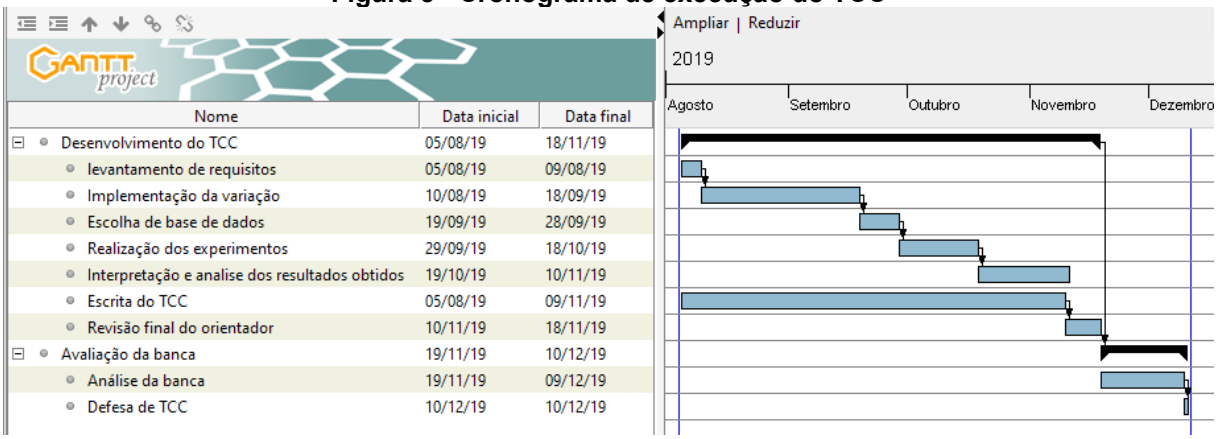
3.1 A EXTENSÃO

3.2 IMPLEMENTAÇÃO

8 CRONOGRAMA

A execução do TCC será dividida em duas grandes etapas representadas na Figura 3. A primeira etapa consiste no desenvolvimento do TCC e é composta por todas as etapas determinadas na seção 6 além de aspectos relacionados a escrita e revisão do trabalho. Na segunda etapa, a de avaliação da banca, foi organizado as atividades esperadas para a avaliação do trabalho por parte da banca, elas envolvem o tempo de análise e a defesa.

Figura 3 - Cronograma de execução do TCC



Nome	Data inicial	Data final
Desenvolvimento do TCC	05/08/19	18/11/19
levantamento de requisitos	05/08/19	09/08/19
Implementação da variação	10/08/19	18/09/19
Escolha de base de dados	19/09/19	28/09/19
Realização dos experimentos	29/09/19	18/10/19
Interpretação e analise dos resultados obtidos	19/10/19	10/11/19
Escrita do TCC	05/08/19	09/11/19
Revisão final do orientador	10/11/19	18/11/19
Avaliação da banca	19/11/19	10/12/19
Análise da banca	19/11/19	09/12/19
Defesa de TCC	10/12/19	10/12/19

Fonte: Elaboração própria.

9 REFERÊNCIAS BIBLIOGRÁFICAS

BOUCKAERT, Remco R.; FRANK, Eibe; HALL, Mark; KIRKBY, Richard; REUTEMANN, Peter; SEEWALD, Alex; SCUSE, David. **WEKA Manual for Version 3-8-1**. Hamilton, Nova Zelândia: University of Waikato, 2016.

CIOS, Krzysztof; SWINIARSKI Roman W; PEDRYCZ Wiltold; KURGAN, Lukasz A. **Data mining:** A Knowledge Discovery Approach. Oklahoma, EUA: 2007.

COSTA, Bruno. **Análise de variações do método de avaliação janela deslizante em modelos preditivos**: um estudo de caso no contexto de pull requests. Acre, Brasil: Universidade Federal do Acre, 2018.

DEAN, Jered. **Big data, data mining and marchine learning**. Wiley: New Jersey, EUA: 2014.

DUNHAM, Margaret H. **Data Mining**: Introductory and Advanced Topics. 1. ed. Dallas, EUA: Pearson, 2003.

FAYYAD, Usama; PIATESTSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **Knowledge Discovery and Data Mining**: Towards a Unifying Framework. 1996a.

FAYYAD, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **IA Magazine**. Califórnia, EUA, v. 17, n. 3, p. 37-54, 1996b.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Editora Atlas SA, 2008.

HALL, Mark; WITTEN, Ian; FRANK, Eibe. **The WEKA Data Mining Software:** An Update. Hamilton, Nova Zelândia: Department of Computer Science, University of Waikato, 2009.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining**: concepts and techniques. Elsevier, 2012.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. **Principles of data mining**. ISBN: 026208290x. Londres, Inglaterra: Massachusetts Institute of Technology, The Mit Press, 2001.

LIMA JÚNIOR, Manoel Limeira de. **Previsão de Integradores e Tempo de Vida de Pull Requests**. Niterói: Universidade Federal Fluminense, 2017.

LIMA, Max Wilian. **Uma extensão da ferramenta weka para avaliação de tarefas preditivas**. Acre, Brasil: Universidade Federal do Acre, 2017.

RODRIGUES, Alan. **Previsão da Natureza de Ocorrências Policias na Cidade de Rio Branco**. Acre, Brasil: Universidade Federal do Acre, 2018.

WAZLAWICK, Raul Sidnei. **Metodologia de Pesquisa para Ciência da Computação**. 6. Ed. Rio de Janeiro: Elservier, 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining: practical machine learning tools and techniques**. 3. ed. Burlington, MA: Morgan Kaufmann, 2011.