

Tarefa 4 - 16/07/2021

1. O que é mineração de dados?

Existem várias definições para Mineração de dados:

1. [Data Mining](https://www.cetax.com.br/blog/data-mining/) ou Mineração de Dados é o processo de explorar grandes quantidades de dados à procura de padrões consistentes. Como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.
2. A mineração de dados é um dos passos de um grande processo chamado de descoberta de conhecimento em banco de dados. A tarefa de mineração se preocupa com a aplicação de algoritmos para analisar os dados ou extrair padrões em categorias específicas dos dados com ou sem a supervisão humana (KLOSGEN; ZYTKOW, 2002).
3. [Data Mining](https://www.cetax.com.br/blog/data-mining/) ou Mineração de Dados consiste em um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e / ou relacionamentos sistemáticos entre variáveis e suas correlações e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados.
4. Data Mining ou Mineração de Dados é o processo de pesquisa em grandes quantidades de dados para extração de conhecimento, utilizando técnicas de Inteligência Computacional para procurar relações de similaridade ou discordância entre dados, com o objetivo de encontrar padrões, irregularidades e regras, com o intuito de transformar dados, aparentemente ocultos, em informações relevantes para a tomada de decisão e/ou avaliação de resultados.
5. Mineração de Dados é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como Inteligência Artificial, especialmente Aprendizado de Máquina, Base de Dados e Estatística. O foco central de Mineração de Dados é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados sem o uso de qualquer técnica, entretanto, existe também muito conhecimento que está de certa forma “embutido” na Base de Dados, na forma de relações existentes entre itens de dados que, para ser extraído, é necessário o desenvolvimento de técnicas especiais.
6. A mineração de dados é a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos.
7. A definição de Mineração de Dados aceita por diversos pesquisadores foi elaborada por Fayyad, Piatetsky-Shapiro, & Smyth (1996a) como sendo: “Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados”.

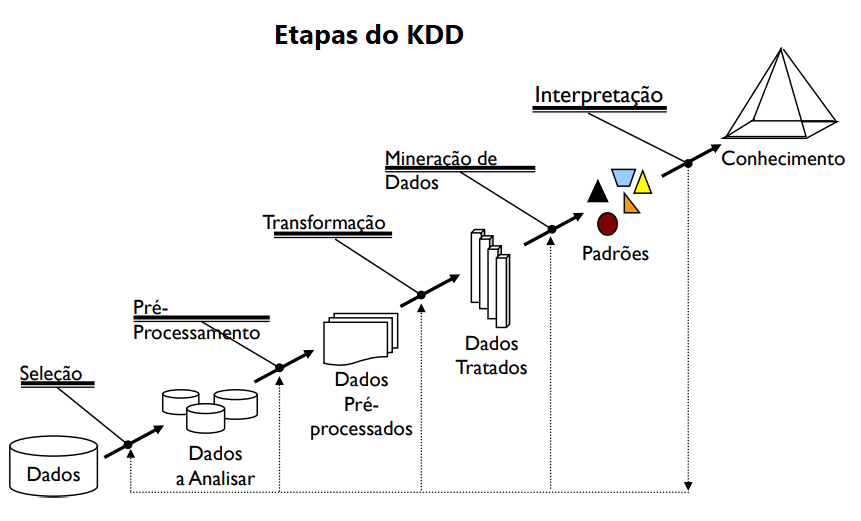
**Data mining** é formada por um conjunto de ferramentas e técnicas que através do uso de [algoritmos](https://www.cetax.com.br/blog/algoritmos-de-vendas-para-melhoria-de-processos-comerciais/) de aprendizagem ou classificação baseados em redes neurais e estatística. Estes são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões da base de dados e auxiliando na descoberta de conhecimento. A premissa do Data Mining é uma argumentação ativa, isto é, em vez do usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas do Data Mining pesquisam automaticamente os mesmos a procura de anomalias e possíveis relacionamentos, identificando assim problemas que não tinham sido identificados pelo usuário. Em outras palavras, as ferramentas de Data Mining analisam os dados, descobrem problemas ou oportunidades escondidas nos relacionamentos dos dados, e então diagnosticam o comportamento dos negócios, requerendo a mínima intervenção do usuário. Assim, ele se dedicará somente a ir em busca do conhecimento e produzir mais vantagens competitivas. O conhecimento em Data Mining pode ser apresentado por essas ferramentas de diversas formas: **agrupamentos, hipóteses, regras, árvores de decisão, grafos ou dendrogramas.** Considerando que minerar dados é um processo de transformar dados em informações úteis (dados mais valiosos a partir de dados complexos). Para atingir esse objetivo, alguns passos são realizados, como: encontrar padrões, associações e anomalias gerais nos dados. Entre as várias tarefas destacam-se associação, classificação, regressão, *clusterização* e sumarização.

1. Existe diferença entre mineração de dados e o processo de descoberta de conhecimento (KDD)?

A expressão Mineração de Dados surge inicialmente, como um sinônimo de KDD, mas é **apenas uma das etapas da descoberta de conhecimento em bases de dados no processo global do KDD**, ou seja, Data Mining **é parte de um processo maior de conhecimento** denominado *Knowledge Discovery in Database* (KDD). A descoberta de conhecimento é um processo que procura por informações relevantes presentes nos dados, sendo a mineração um de seus passos.

O termo KDD surgiu no final da década de 80, com objetivo de procurar conhecimento em bases de dados. O primeiro processo desenvolvido foi apresentado por Fayyad et al. (1996) em seu livro, onde o modelo não focava em uma técnica particular, mas provia um processo genérico para a geração de conhecimento presente nos dados. Muitas são as definições para este conceito, sendo a seguinte definição a mais utilizada: “KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados” (FAYYAD, 1996). A expressão “não trivial”, demonstra a complexidade na execução e manutenção dos processos de KDD, o termo “interativo” indica a relevância de se ter um elemento controlando o processo, o termo “iterativo” sugere a possibilidade de repetições em qualquer uma das etapas do processo e finalmente o “conhecimento útil” que é aquele onde o objetivo foi alcançado, trazendo consigo benefícios as aplicações de KDD. A extração de conhecimento em bases de dados é um processo dinâmico e evolutivo, que envolve relacionamento com outras áreas como estatística, inteligência artificial, máquina de estado e banco de dados. Os padrões extraídos devem ser úteis, gerando um conhecimento que poderá tirar alguma vantagem, seja cientifica ou comercial. O KDD, como um processo, consiste de uma sequência interativa e iterativa das seguintes etapas: Seleção, Pré-processamento, Transformação, Garimpagem (Mineração) e Análise e Assimilação dos Resultados (Cabena et al. 1998) conforme figura 01, sendo assim este processo dividido em 5 fases:

1. Seleção de dados;
2. Pré-processamento;
3. Transformação;
4. Mineração;
5. Análise e assimilação de resultados.



Após a realização das fases I, II e III, a Mineração de Dados (Data Mining) é iniciada. Esta fase é considerada a mais importante do KDD para a extração de conhecimento, baseando-se em técnicas da estatística, inteligência artificial, computação paralela, máquina de estado. Sendo realizada através da escolha do método e do algoritmo mais compatível com o objetivo da extração, a fim de encontrar padrões, relacionamentos entre dados, anomalias e regras, com objetivo de encontrar informações ocultas, que possam ser relevantes a tomada de decisão e/ou avaliação de resultados.

1. Explique as atividades preditivas e as atividades descritivas da mineração de dados.

De forma geral as tarefas mineração de dados podem ser classificadas em duas categorias principais: **Descritivas** e **Preditivas** (HAN et al. 2001).

As Atividades de descrição consistem na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada. Ou seja, as tarefas descritivas de mineração de dados caracterizam propriedades gerais dos dados em um modelo. Reliza a identificação dos segmentos com caracteristicas semelhantes, agrupando os casos ou as ocorrências de acordo com as informações. É empregado em modelagens não supervionadas. Possibilita que sejam encontrados padrões de comportamento e conhecimento que dificilmente um ser humano encontraria. Algumas das tarefas de descrição são clustering, regras de associação e sumarização.

Já as atividades de predição consistem na generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de reconhecer a classe de um novo exemplo. Ou seja as tarefas preditivas efetuam inferências nos dados em analise, visando o delineamento de predição. Realiza predição de eventos ou acontecimentos com previa ciência dos resultados. É empreagado na modelagem supervisionada. Na Preditiva são criados modelos formais que permitem estimar, classificar ou agrupar dados, a partir de determinados comportamentos que foram aprendidos e encontrados na Análise Descritiva. Com esse tipo de análise, é possível prever o que acontece com base nos dados históricos. O ponto mais crucial é a qualidade dos dados para que a previsão seja o mais precisa possível. Os dois principais tipos de tarefas para predição são classificação e regressão. A classificação consiste na predição de um valor categórico como, por exemplo, predizer se o cliente é bom ou mau pagador. Na regressão, o atributo a ser predito consiste em um valor contínuo como, por exemplo, predizer o lucro ou a perda em um empréstimo (Weiss & Indurkhya 1998).

Como as tarefas possíveis de um algoritmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas a escolha da tarefa deve ser feita de acordo com os objetivos desejáveis para a solução a ser encontrada.

1. Explique como os modelos de redes neurais artificiais podem ser aplicados para mineração de dados.

As redes neurais artificiais são utilizadas na mineração de dados em tarefas como classificação (supervisionado) e agrupamento (não supervionado).

Normalmente, quando se configura uma rede neural para realizar classificação de padrões, a camada de entrada tem tantos neurônios quanto atributos preditivos, e a camada de saída possui tantos neurônios quanto classes no problema. Dessa forma, a partir de um padrão apresentado à rede, a saída que for mais ativada determinará a classe correspondente.

Em previsão de séries temporais, normalmente a rede possui tantos neurônios quanto forem os valores dentro da “janela” de previsão escolhida. A camada de saída possui um único neurônio que irá gerar o valor da previsão para o instante futuro desejado. Geralmente, tanto em classificação quanto em previsão as redes neurais utilizadas são do tipo feed-forward com aprendizado baseado no algoritmo backpropagation. O treinamento é realizado de forma supervisionada, ou seja, é preciso que existam padrões com atributos objetivos definidos.

As Redes Neurais Artificiais (RNA) tem sido muito utilizadas, como classificadores no processo de Mineração de Dados, no que diz respeito predição de séries temporais em vários ramos de negócio tais como: econômico-financeiro, predição de séries caóticas , consumo de energia dentre outras. As RNAs podem aprender a partir de exemplos, reconhecer um padrão escondido em observações históricas e usá-los para predizer valores futuros. Além disso, são capazes de lidar com informação incompleta ou dados ruidosos e podem ser muito eficazes, especialmente em situações onde não é possível definir as regras ou etapas que levam à solução de um problema.

Na tarefa de clusterização ou agrupamento, o treinamento é não supervisionado, ou seja, não existem atributos objetivos definidos. Nesses casos, são utilizadas redes neurais auto-organizáveis, onde o Mapa de Kohonen [KOHO97], baseado em aprendizado competitivo, vem apresentando resultados muito satisfatórios. Uma das classes de redes neurais mais conhecidos são os Mapas Auto Organizáveis de Kohenen. Essas classes são caracterizadas por serem baseados em uma forma de aprendizado competitivo, onde estruturas denominadas de neurônios tendem a aprender a distribuição estatística dos dados de entrada. O seu campo de aplicação é designado principalmente para reconhecimento de padrões (DONI, 2004). As Redes de Kohonen Auto Organizáveis (SOM) são comumente e eficazmente utilizadas para fins de Mineração de Dados em tarefas de Clusterização. Além disso, podem ser utilizadas para facilitar a remoção de “outliers” de base de dados.

As Redes Neurais Recorrentes (RNAR) ainda não foram totalmente exploradas, no processo de Mineração de Dados, principalmente no que diz respeito à predição de séries temporais, devido ao seu tedioso treinamento e por conta de suas estruturas complexas.

Porém o problema indicado por especialistas na utilização de Redes Neurais Artificiais em tarefas de classificação e previsão é o fato dessas serem estruturas do tipo “caixa-preta”, ou seja, de difícil interpretação de seus resultados. Isso se deve ao fato de existirem camadas escondidas com possivelmente vários neurônios contendo vários pesos, o que se torna praticamente impossível de interpretar. Recentemente, muitos trabalhos foram realizados para extrair regras a partir de redes neurais previamente treinadas. Com isso, o problema de interpretação dos resultados foi minimizado, encorajando a utilização desta técnica. Existem três vertentes de algoritmos de extração de regras:

• Decomposição: Utilizam informações dos pesos internos na geração das regras [THRU93] [NEAL96];

• Black-Box: Não utilizam informações internas, apenas fazem correlações das entradas com as saídas geradas [SETI96][SETI97];

• Ecléticos: Utilizam princípios dos dois anteriores [CRAV94], [TICK96].

1. Defina o modelo de classificação de árvore de decisão (o que são os nós? O que são as arestas?).

A árvore de Decisão é um tipo de algoritmo de aprendizagem de máquina supervisionado (com uma variável alvo pré-definida) que se baseia na ideia de divisão dos dados em grupos homogêneos, muito utilizada em problemas de classificação. As arvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Funciona para as variáveis categóricas e contínuas de entrada e de saída. Na árvore de decisão, dividimos a população ou amostra em dois ou mais conjuntos homogêneos (ou sub-populações) com base nos divisores/diferenciadores mais significativos das variáveis de entrada. Em outras palavras, em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas. Estas últimas são as classes. Estes modelos utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema (Gama, 2004). As árvores de decisão estão entre os mais populares algoritmos de inferência e tem sido aplicado em várias áreas como, por exemplo, diagnóstico médico e risco de crédito (Mitchell, 1997), e deles pode-se extrair regras do tipo “se-então” que são facilmente compreendidas. A capacidade de discriminação de uma árvore vem da divisão do espaço definido pelos atributos em sub-espaços e a cada sub-espaço é associada uma classe.

As árvores de decisão representam uma técnica de classificação onde as regras de decisão estão dispostas na forma de uma árvore. A representação segue o seguinte padrão:

* Cada nó de decisão contém um teste num atributo
* Cada ramo descendente corresponde a um possível valor deste atributo
* Cada folha está associada a uma classe
* Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação

Assim, uma Árvore de Decisão é composta por nó folha (ou nó resposta) que contém o nome de uma classe ou o símbolo nulo (nulo indica que não é possível atribuir nenhuma classe ao nó por não haver nenhum exemplo que corresponda a esse nó); ou um nó interno (ou nó de decisão) que contém o nome de um atributo; para cada possível valor do atributo, corresponde um ramo para outra árvore de decisão. Cada nó de decisão contém um teste para algum atributo, cada ramo descendente corresponde a um possível valor deste atributo, os conjuntos de ramos são distintos, cada folha está associada a uma classe e, cada percurso da árvore, da raiz à folha corresponde uma regra de classificação. No espaço definido pelos atributos, cada folha corresponde a um hiper-retângulo onde a interseção destes é vazia e a união é todo o espaço (Gama, 2004).

No momento em que se deseja classificar um novo padrão, percorrem-se os nós da árvore iniciando pela raiz e tomam-se decisões em cada nó a partir do valor assumido pelo atributo. Assim como a criação da árvore, a classificação é feita de forma top-down; logo a classificação termina quando a busca através da árvore atinge uma folha. O mesmo atributo pode aparecer mais de uma vez na árvore e podem existir caminhos na árvore onde alguns atributos não apareçam.

O objetivo da árvore de decisão é encontrar o atributo que gera a melhor divisão dos dados, subconjunto com maior pureza. Existem algumas métricas para a definição de pureza, ou seja, qual será a métrica utilizada para decidir qual é o melhor atributo que divide os dados gerando a partição mais pura. Essas métricas são o Indice Gini, Chi-Square, Information Gain e a redução da variância.

Os fatores que devem ser levados em consideração para utilizar uma árvore de decisão:

* Possui um fácil entendimento, pois não requer nenhum conhecimento estatístico para a sua interpretação.
* Aceita tanto dados categóricos quanto numéricos diminuindo a necessidade da limpeza de dados em comparação com outros modelos.
* Ela é propensa a sofrer overfitting (sobreajuste), se ajustar muito aos dados de treino e não ter uma performance muito boa com os dados de teste.
* São instáveis, pequenas alterações nos dados de treino produzem novas árvores.

1. Como os algoritmos de árvore de decisão podem selecionar o nó raiz da árvore? Com base em que esta seleção deve ser realizada?

**Construir uma árvore de decisão se trata de encontrar regras sobre as variáveis do modelo (ou pontos de corte) que retornam o maior ganho de informação, isto é, que tornam os ramos da árvore mais homogêneos, com menor entropia**. Na construção da árvore de decisão, procura-se associar a cada nó de decisão o atributo “mais informativo” entre aqueles ainda não utilizados no caminho desde a raiz da árvore. A escolha central no algoritmo ID3 está em selecionar qual atributo de teste será usado em cada nodo da árvore. É interessante selecionar o atributo que é mais útil para classificar exemplos. Assim, é definida uma propriedade estatística chamada ganho de informação, que mede como um determinado atributo separa os exemplos de treinamento de acordo com a classificação deles. O ID3 usa o ganho de informação para selecionar, entre os candidatos, os atributos que serão utilizados a cada passo, enquanto constrói a árvore. O algoritmo básico, ID3, constrói árvores de decisão a partir da raiz e começa com a pergunta “que atributo deveria ser testado na raiz da árvore?”. Para responder esta pergunta, cada atributo da instância é avaliado usando um teste estatístico para determinar como este classifica os exemplos de treinamento. O melhor atributo é selecionado e é usado como o teste no nodo raiz da árvore. Um descendente do nodo raiz é criado então para cada possível valor deste atributo, e os exemplos de treinamento são particionados e associados a cada nodo descendente para selecionar o melhor atributo para testar naquele ponto na árvore. Isto forma uma procura para uma árvore de decisão aceitável na qual o algoritmo nunca retrocede para reconsiderar escolhas feitas anteriormente. **O melhor atributo é aquele com o ganho de informação maior.** O ID3 é um algoritmo que constrói a árvore de forma descendente (“top -down”), a cada nodo seleciona-se o atributo que melhor classifica os exemplos de treinamento locais. Após a definição de entropia como uma medida da impureza em uma coleção de exemplos de treinamento, pode-se definir a medida da efetividade de um atributo para classificar os dados de treinamento chamada ganho de informação, que é simplesmente a redução esperada na entropia causada pelo particionamento dos exemplos por este atributo. O ID3 determina o ganho de informação para cada atributo candidato, então seleciona aquele com o ganho de informação maior.

 A seleção dos nodos a serem utilizados na árvore é baseada na Teoria da Informação de Shannon, mais especificamente nos conceitos de entropia e ganho de informação. Onde entropia é quantidade necessária de informação para identificar a classe de um caso. **Ganho de Informação é** a redução esperada da entropia ao utilizarmos um atributo na árvore.

O ganho de informação é dado por:

Ganho (S, A) = Entropia (S) - ∑ ((|Sv| / |S|) \* Entropia (Sv))

Alguns dos métodos mais utilizados para definir o no da arvore:

* **Entropia:** através da entropia o algoritmo verifica como os dados estão distribuídos nas variáveis preditoras de acordo com a variação da variável target. Quanto maior a entropia, maior a desordem dos dados; e quanto menor, maior será a ordem destes dados, quando analisados pela ótica da variável target. Partindo da entropia, o algoritmo confere o ganho de informação de cada variável. **Aquela que apresentar maior ganho de informação será a variável do primeiro nó da árvore**. Podemos entender o ganho de informação como a medida de quão bem relacionados os dados da variável preditora estão com os dados da variável target (ou o quanto a variável target pode ser explicada a partir da variável preditora), sendo que a variável com melhor desempenho será a escolhida para iniciar a árvore.
* **Índice GINI**: com o cálculo do índice GINI, assim como na Entropia, será verificada a distribuição dos dados nas variáveis preditoras de acordo com a variação da variável target, porém com um método diferente. **A variável preditora com o menor índice Gini será a escolhida para o nó principal da árvore**, pois um baixo valor do índice indica maior ordem na distribuição dos dados.
* **Regressão:** nos [problemas de regressão](https://didatica.tech/problemas-de-classificacao-e-regressao/) nosso objetivo é prever um valor, e não uma classe. Para isso a árvore utilizará os conceitos de média e desvio padrão, que possibilitarão um resultado final numérico. Para definir as variáveis preditoras dos nós principais em um problema de regressão, será calculado o desvio padrão dos valores da variável target para cada variável preditora, de acordo com suas variações. Desta forma teremos um valor de desvio padrão para cada variável preditora e, comparando-o com o desvio padrão da variável target completa, chegaremos à redução de desvio padrão que a variável preditora em questão aplicou sobre a variável target. Lembrando que o desvio padrão indica o quão distante os valores estão da média, podemos entender que uma variável com grande redução de desvio padrão indica que através dela a variável target se aproxima da média, mostrando uma grande relação entre a variável preditora e a variável target. Portanto, **a variável preditora com maior redução de desvio padrão será escolhida para o nó principal da árvore**.

1. Como é possível obter regras do tipo “se-então” através de uma árvore de decisão gerada a partir de um modelo de classificação?

Através dos proprios caminhos descritos por um arvore de decisão é possivel extair as regras. As arvores e as regras são geralmente utilizadas em conjunto. Devido ao fato de as arvores tenderem a crescer muito de acordo com as aplicações, elas são muitas vezes substituidas pelas regras. Isso acontece em virtude de as regras poderem ser facilmente modularizadas [ING2000]. Portanto, a partir de uma árvore de decisão são geradas regras, onde cada caminho possível da árvore corresponde a uma regra do tipo SE-ENTÃO. Os antecedentes são formados pelos atributos preditivos que aparecem ao longo do caminho percorrido, testando os valores que os definem. Os conseqüentes são formados pelo atributo objetivo com a classe correspondente à folha. Na arvore de decisão os nos são representados pelos atributos que estao dispostos conforme seu nivel informativo. Nos arcos são testados os valores do atributo designado ao nodo a que pertencem. Os testes são realizados utilizando-se operadores lógicos que realizão a avalição logica entre o no e o arco (Se no1= arcoX entao classe I). E a folha indicara entao à classe associada ao no folha. Os testes a serem realizados em um nodo de uma arrvore de decisão dependem das caracteristicas do atributo designado a este nodo, sendo utilizado apenas um atributo por nodo na realização de cada teste, tornando assim, a estrutura da arvore de facil compreensão. Com base nos testes definidos e em um conjunto de exemplos, é decidido qual o caminho a percorrer na arvore durante o processo de classificação.

Algumas possiveis regras que podem ser extraidas da arvore de decisão:

**Se** Valor\_atributo\_no **=** Valor\_arco **entao** classe 1

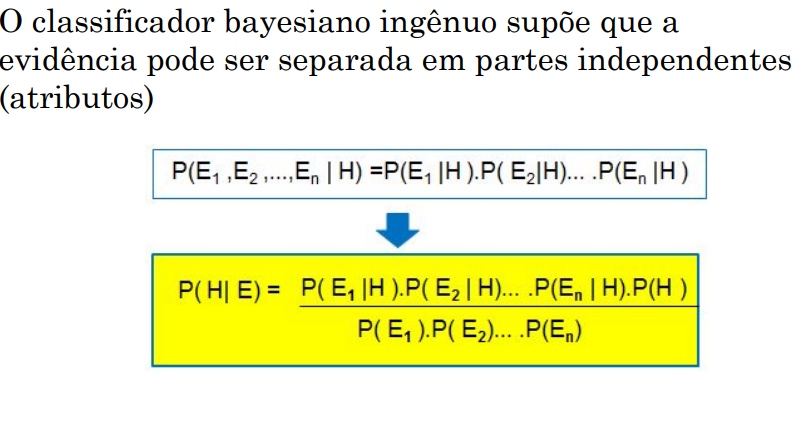
**Se** Valor\_atributo\_no **>=** Valor\_arco **entao** classe 2

**Se** Valor\_atributo\_no <=Valor\_arco **entao** classe 3

1. No modelo de classificação baseado em probabilidade chamado naïve bayes, os atributos são considerados independentes. Demonstre isto através das fórmulas que regem o modelo.

O teorema de Bayes fornece uma maneira de calcular a probabilidade posterior, P ( c | x ) , de P ( c ) , P ( x ) e P ( x | c ) . O classificador Naive Bayes assume que o efeito do valor de um preditor ( x ) em uma determinada classe ( c ) é independente dos valores de outros preditores. Essa suposição é chamada de independência condicional de classe.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| https://www.saedsayad.com/images/Bayes_rule.png  P ( c | x ) é a probabilidade posterior da classe ( alvo ) dado preditor ( atributo ).  P ( c ) é a probabilidade anterior da classe .  P ( x | c ) é a probabilidade que é a probabilidade do preditor dada classe .  P ( x ) é a probabilidade anterior do preditor . |  |  |
|  |  |  |



1. Descreva o algoritmo a priori.

O algoritmo a priori usa conjuntos de itens frequentes para gerar regras de associação. Baseia-se no conceito de que um subconjunto de um conjunto de itens frequentes também deve ser um conjunto de itens frequentes. O Conjunto de itens frequentes é um conjunto de itens cujo valor de suporte é maior do que um valor limite (suporte). O algoritmo Apriori é um dos algoritmos mais conhecidos quando o assunto é mineração de regras de associação em grandes bancos de dados centralizados. Ele encontra todos os conjuntos de itens freqüentes, denominados itemsets freqüentes (Lk). O algoritmo principal (Apriori) normalmente faz uso de duas funções principais: a função uma para gerar os candidatos e eliminar aqueles que não são freqüentes, e uma função para extrair as regras de associação.

O algoritmo APRIORI é considerado um clássico na extração de Regras de Associação. Ele foi proposto pela equipe de pesquisa QUEST da IBM que deu origem ao Software Intelligent Miner. Este algoritmo faz recursivas buscas no Banco de Dados à procura dos conjuntos freqüentes (conjuntos que satisfazem um suporte mínimo estabelecido). Alem disto, possui diversas propriedades que otimizam o seu desempenho, como por exemplo, a propriedade de antimonotonia da relação, que diz que para um itemset ser freqüente, todos os seus subconjuntos também devem ser, além de utilizar recursos da memória principal e estrutura hash. As três fases que compõem o APRIORI são: Geração dos conjuntos Candidatos; Poda dos conjuntos Candidatos e Contagem do Suporte (Nesta fase é necessário visitar o BD)

A este algoritmo é aplicada a propriedade de Antimonotonia da Relação ou Propriedade Apriori que é descrita a seguir: Se X está contido em Y e X não é freqüente, logo Y também não é freqüente. Isto implica uma diminuição do tempo de execução, pois se X não é freqüente, então não será necessário calcular o suporte de Y, e o BD não precisará ser varrido.

**Funcionamento do Algoritmo Apriori**: Esse algoritmo gera um conjunto de itens freqüentes a cada uma de suas passagens. Com base nestes Conjuntos será gerado um outro conjunto Ck, conjunto de itens candidatos, que consta os itens do conjunto freqüente (Lk) com minsup maior que o estabelecido. O conjunto candidato é resultado do produto cartesiano do conjunto de freqüentes da passagem anterior com ele mesmo. Posteriormente o conjunto candidato é podado, seu suporte é contado e os itens que tem suporte acima do estabelecido serão os itens freqüentes da próxima passagem (Lk+1). O Algoritmo Apriori utiliza os itens freqüentes obtidos pelo comando executado em SQL, sendo a primeira passagem k=1. Para k=2, enquanto o conjunto obtido na passagem anterior não for vazio então k será incrementado e o conjunto de candidatos receberá os itens retornados pela função apriori\_gen. A Função Apriori-gen é responsável pela união dos conjuntos freqüentes a fim de formar o conjunto candidato com k itens. Para isso tem os itens freqüentes da passagem anterior como parâmetro. Ela faz também a poda dos candidatos. Então para todas as transações t contidas no conjunto de transações é adicionado um contador de suporte, verificando assim quais itens do conjunto candidato estão contidos em cada uma das transações. Para o processo de contagem do suporte dos candidatos, os conjuntos são dispostos em uma árvore Hash. Esse é um método de espalhar os elementos de um conjunto seguindo uma dada função (função hash) com ela é possível realizar uma busca direta pelo elemento desejado, evitando a principio buscas seqüenciais em todo conjunto, acarretando em um ganho significativo em tempo de execução.



*Figura 1 – Algoritmo Apriori*

Um nó em uma árvore hash ou contém uma lista de conjuntos de itens (nó folha), ou contém uma tabela hash (nó interno) essa é usada quando o número máximo de elementos em uma folha excede o limite estabelecido. Quando um conjunto candidato é adicionado, inicia-se da raiz da árvore até alcançar uma folha, a definição do caminho a ser seguido é dada pela função hash calculada para este anteriormente. Inicialmente cada nó é criado como sendo uma folha. A poda é realizada se algum subconjunto do conjunto candidato não estiver presente no conjunto de itens freqüentes da passagem anterior. Como meio de otimização, a poda dos conjuntos também pode ser feita através de uma árvore hash, mas no algoritmo original ela é feita através da função Apriori\_Gen.

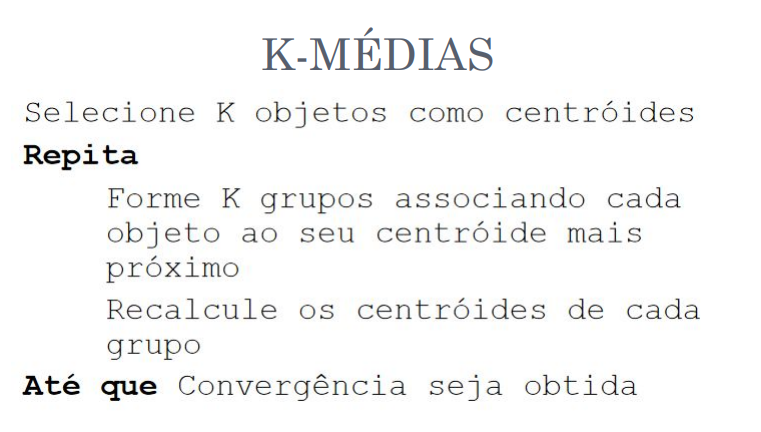
A Função Subset é encarregada de contar o suporte dos itens candidatos, ela toma como parâmetros o conjunto candidato (Ck) e o conjunto de Transações (T). Primeiramente ela faz as combinações entre os itens da transação da seguinte forma: um item é combinado com todos os outros que estão imediatamente a sua frente e assim até o ultimo elemento.

Posteriormente verificam-se quais destes estão presentes na árvore hash, da seguinte forma: É calculada a função hash dos subconjuntos obtidos com a transação e então esses são comparados com a árvore, sendo que se presentes na árvore haverá um contador de suporte que incrementará o suporte deste itemset.

Feito isso teremos o suporte dos itens candidatos, os itens que possuírem suporte maior que o estabelecido formarão o conjunto de itens freqüentes desta passagem.

O Algoritmo Apriori termina quando o conjunto de itens freqüentes da passagem anterior for igual a zero, e retorna como resultado de sua execução a união de todos os itens freqüentes de todas as passagens.

1. Considere o conjunto de dados X, formado pelos seguintes pontos: A(1,1), B(2,3), C(4,1), D(3,2), E(7,7), F(8,5) e G(6,5).
   1. Represente graficamente os dados.
   2. Apresente passo a passo o algoritmo k-médias



Passos Para execução do K-médias:

1. Primeiramente escolhem-se k centróides, chamados de sementes ou protótipos, para se inicializar o processo de partição;

2. Cada elemento do conjunto de dados é comparado com cada centróide inicial através da distância desejada (usualmente Euclidiana). O elemento é alocado ao cluster de menor distância

3. Após aplicar o passo 2 para todos os n elementos amostrais, atualiza-se os valores dos centróides de todos os grupos formados, e repete-se o passo 2 considerando os centróides desses novos grupos.

4. Os passos 2 e 3 são repetidos até que nenhum dos elementos amostrais seja realocado.

* 1. Empregando o algoritmo descrito no item b) agrupe os pontos do conjunto X em dois grupos cujos centros são: centro1(2,3) e centro2(7,7). A métrica de distância a ser empregada é dada pela Equação1:

D(a,b) = |x1-x2|+|y1-y2| , onde a(x1,y1) e b(x2,y2) (1)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Primeira Interação** | | **Distancia dos Pontos** | | | | | | | |
| **Grupo** | **Centroide** | **A** | **B** | | **C** | **D** | **E** | **F** | **G** |
| **G1** | **B (2; 3)** | 3 | 0 | | 4 | 2 | 9 | 8 | 6 |
| **G2** | **E (7; 7)** | 12 | 9 | | 9 | 9 | 0 | 3 | 3 |
| **Agrupamento** | | **DG1<DG2** | **DG1<DG2** | | **DG1<DG2** | **DG1<DG2** | **DG1>DG2** | **DG1>DG2** | **DG1>DG2** |
| **G1** | **G1** | | **G1** | **G1** | **G2** | **G2** | **G2** |
| **Definindo novo centroide** | | | A definição do Centroide é dada pela média dos elementos que compõem cada um dos grupos | | | | | | |
| **G1** | **X =** | **2,5** |
| **Y =** | **1,75** |
| **G2** | **X =** | **7** |
| **Y =** | **5,67** |
| **Segunda Interação** | | **Distancia dos Pontos (D)** | | | | | | | |
| **Grupo** | **Centroide** | **A** | | **B** | **C** | **D** | **E** | **F** | **G** |
| **G1** | **(2.5; 1.75)** | 2,25 | | 1,75 | 2,25 | 0,75 | 9,75 | 8,75 | 6,75 |
| **G2** | **(7; 5.67)** | 10,67 | | 7,67 | 7,67 | 7,67 | 1,33 | 1,67 | 1,67 |
| **Agrupamento** | | **DG1<DG2** | | **DG1<DG2** | **DG1<DG2** | **DG1<DG2** | **DG1>DG2** | **DG1>DG2** | **DG1>DG2** |
| **G1** | | **G1** | **G1** | **G1** | **G2** | **G2** | **G2** |
| **Definindo novo centroide** | | | | Na segunda interação não houve alterações do agrupamento e nem uma definição de centroide diferente. Encerrando assim a interação e o algoritmo K-médias. | | | | | |
| **G1** | **X =** | **2,5** | |
| **Y =** | **1,75** | |
| **G2** | **X =** | **7** | |
| **Y =** | **5,67** | |

Webgrafia: http://www.each.usp.br/lauretto/cursoR2017/04-AnaliseCluster.pdf

1. Explique a abordagem estatística dos dados na Mineração de texto.

A abordagem Estatística dos dados na Mineração de texto baseia-se principalmente na **frequência dos termos dentro do texto**, ignorando qualquer informação semântica. **É independente do idioma**. A importância de um termo é dada pelo número de vezes que este aparece no texto. Basicamente, seu processo envolve aprendizado estatístico a partir de dados, que normalmente inclui as etapas de codificação dos dados, estimativa dos dados e modelos de representação de documentos.

Na **Codificação dos Dados** uma codificação inicial dos dados é escolhida com base em indicações de especialistas. Também pode ser feita de acordo com critérios que representem propriedades interessantes dos dados em relação aos objetivos da seleção dos mesmos. Se informações relevantes forem descartadas nesta etapa, não poderão ser recuperadas depois. Entretanto, se a codificação inicial dos dados contém muita informação irrelevante ou ruídos, a busca por uma seleção adequada pode se tornar difícil ou consumir muito tempo. Além disso, propriedades importantes destes dados podem ser perdidas em meio ao ruído.

Na **Estimativa dos Dados** a etapa envolve a procura por um modelo adequado a partir de um conjunto de modelos (espaço de modelos). Um modelo pode ser obtido a partir da aplicação de um algoritmo de aprendizado ou de um método de estimativa.

Nos **Modelos de Representação de Documentos** os documentos podem ser vistos como “containers” de palavras. Esta abordagem, também conhecida como bag of words, ignora a ordem que as palavras aparecem nos textos, assim como qualquer informação de pontuação ou de estrutura, mas retém o número de vezes que a palavra aparece. Provê informações para agrupamento e recuperação de informação a partir de grandes coleções de dados. Esta técnica é considerada uma simplificação de toda a abundância de informações que um texto pode expressar, não fornecendo, portanto, uma descrição fiel de seu conteúdo. O desenvolvimento de modelos mais ricos, que sejam computacionalmente viáveis e possíveis de serem estimados, continua sendo um problema desafiador para a computação. Entretanto, apesar desta técnica não ser suficiente para interpretação completa a respeito dos textos, ela provê uma quantidade considerável de informações sobre associações entre palavras e documentos que tem se apresentado suficiente para clustering e para recuperação de informações a partir de grandes coleções de textos.

1. Apresente uma aplicação de mineração de dados especificando a tarefa de mineração empregada e um algoritmo que pode ser empregado.

A aplicação escolhida para estudo foi proposta por Baldomir(2017), com objetivo de auxiliar auditores da Controladoria-Geral da União na tarefa de encontrar relações entre as empresas participantes de licitações públicas e possíveis fraudes.

A aplicação conta com uma interface web, com um campo de busca para inserção de um CNPJ. Com essa informação, todas as compras relacionadas ao CNPJ são buscadas na base de dados. A partir de cada compra é gerada uma lista de todos os CNPJs que participaram da licitação. Esse conjunto de dados é enviado para o componente **Apriori**.

Procurando identificar as relações entre as empresas, a aplicação usou das **regras de associação nos dados** das licitações disponíveis no ComprasNet. Para um conjunto ser considerado frequente, ele deve aparecer nas transações em, pelo menos, uma quantidade igual ao suporte mínimo. O conjunto de itens analisado pelo algoritmo foi dos CNPJs retornados pelo Banco.

A partir da aplicação do Apriori, são definidos os valores de suporte e confiança. A partir dos valores a aplicação retorna as regras de associação resultantes entre os CNPJs.

**Referência Bibliográfica:**

1) CASTANHEIRA, L. C. Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões. Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais. Belo Horizonte. 2008. Disponível em: <https://www.ppgee.ufmg.br/defesas/349M.PDF>. Acesso: 21 Jul. 2021.

2) MELO, M. D. Um processo de mineração de dados para a predição de níveis criminais de áreas geográficas urbanas. Mestrado acadêmico de Ciência da Computação, Universidade Estadual do Ceará. Fortaleza. 2010. Disponível em: <http://www.uece.br/ppgcc/wp-content/uploads/sites/51/2020/02/MARCELO-DAMASCENO-DE-MELO.pdf>. Acesso: 21 Jul. 2021.

CONDE, G. A. B et al. Data Mining. Laboratório de Suporte à Decisão, 2013. Disponível em: <http://www.ufopa.edu.br/lsd/index.php?option=com_content&view=article&id=8:beginners&catid=19&Itemid=260>. Acesso: 21 Jul. 2021.

GALVÃO, N. D, MARIN H. F. Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem, 2009. Disponível em: <https://www.scielo.br/j/ape/a/Lzj9vW6Fp4QVdXyNKhmtfvv/?lang=pt>. Acesso: 21 Jul. 2021.

DANTAS, E. R. G. et al. O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões. SEGeT – Simpósio de Excelência em Gestão e Tecnologia, 2010. Disponível em: <https://www.aedb.br/seget/arquivos/artigos08/331_331_Artigo_SEGET_EJDR_Versao_Final_010808.pdf>. Acesso: 21 Jul. 2021.

BERNARDINI, F. C. Introdução ao Processo de KDD e MD. 81 slides. Disponível em: <https://www.professores.uff.br/fcbernardini/wp-content/uploads/sites/68/2017/08/01-Introdução-a-KDD-e-DM.pdf>. Acesso: 21 Jul. 2021.

BARANAUSKAS, J. A. Extração de Conhecimento Extração de Conhecimento

& Mineração de Dados. 42 slides. Disponível em: <https://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-KDD-DM.pdf>. Acesso: 21 Jul. 2021.

PÁDUA, S. I. D. Sistemas de Apoio à Tomada de Decisão. 57 slides. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4450337/mod_resource/content/1/Aula%205%20SAD%20%20Dataming.pdf>. Acesso: 21 Jul. 2021.

SILVA, M. A. O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica. Programa de Pós Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins. 2014. Disponível em: https://docs.uft.edu.br/share/s/VbZ4G1KXQLqWbTuCPJqv7Q. Acesso: 21 Jul. 2021.

PASTA, A. Aplicação da Técnica de Data Mining na Base de Dados do Ambiente de Gestão Educacional: Um Estudo de Caso de Uma Instituição de Ensino Superior de Blumenau-SC. Mestrado em Computação Aplicada. São José. 2011. Disponível em: <http://www.uniedu.sed.sc.gov.br/wp-content/uploads/2013/10/Arquelau-Pasta.pdf>. Acesso: 21 Jul. 2021.

3) CAMPOS, C. M. Análise inteligente de dados em um banco de dados de procedimentos em cardiologia intervencionista. Doutora em Ciências. Univeridade de São Paulo. São Paulo. 2016. Disponível em: <https://www.teses.usp.br/teses/disponiveis/98/98131/tde-18102016-085650/publico/TeseCantidiodeMouraCamposNeto_Corrigida.pdf>. Acesso: 21 Jul. 2021.

PETERMANN, R. J. Modelo de Mineração de Dados para Classificação de Clientes em Telecomunicações. Programa de Pós-Graduação em Engenharia Elétrica. Pontíficia Universidade Católica do Rio Grande do Sul. Porto Alegre. 2006. Disponível em: <https://tede2.pucrs.br/tede2/bitstream/tede/3044/1/388093.pdf> Acesso: 21 Jul. 2021.

PÁDUA, S. I. D. Sistemas de Apoio à Tomada de Decisão. 57 slides. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4450337/mod_resource/content/1/Aula%205%20SAD%20%20Dataming.pdf>. Acesso: 21 jul. 2021.

4) WERMERSCH, Fábio Glauco. Uso de redes neurais artificiais para descoberta de conhecimento sobre a escolha do modo de viagem. 2002. Dissertação (Mestrado em Transportes) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2002. doi:10.11606/D.18.2017.tde-10072017-165016. Acesso: 21 jul. 2021.

BUENO, L. F. Inteligência artificial aplicada à melhoria da acurácia do mapeamento de redes de drenagem. Tese (Doutor em Geografia) - Universidade Federal do Paraná. Curitiba, 2016. <https://acervodigital.ufpr.br/handle/1884/44459>. Acesso: 21 jul. 2021.

ARRAES, D. et al. Arquiteturas de Redes Neurais Aplicadas a Data Mining no Mercado Financeiro Uma Aplicação para a Geração de Credit Ratings. V Congresso Brasileiro de Redes Neurais. Rio de Janeiro, 2001. Disponível em: <http://abricom.org.br/wp-content/uploads/2016/03/5cbrn_029.pdf>. Acesso: 21 jul. 2021.

PALMA, L. F. Agrupamento de Dados: K- Médias. Monografia (Bacharel em Ciências Exatas e Tecnológicas). Universidade Federal do Recôncavo da Bahia. Cruz das Almas, 2018. Disponível em: <https://www2.ufrb.edu.br/bcet/components/com_chronoforms5/chronoforms/uploads/tcc/20190604200511_2018.2_TCC_Luann_Farias_Palma-_Agrupamento_de_dados_-_K_medias.pdf>. Acesso: 21 jul. 2021.

5) KLEINA, M. Árvores de Decisão. 43 slides. Disponível em: <https://docs.ufpr.br/~marianakleina/ARVORE_DECISAO.pdf>. Acesso: 21 jul. 2021.

QUEIROZ, S. et al. Árvores de Decisão. Universidade Federal de Pernambuco. 51 slides. Disponível em: <https://www.cin.ufpe.br/~if684/EC/aulas/Aula-arvores-decisao-SI.pdf>. Acesso: 21 jul. 2021.

6) GARCIA, S. C. O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde. Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós Graduação em Computação. Porto Alegre, 2003. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/4703/000503532.pdf?sequence=1>. Acesso: 21 jul. 2021.

D. R. CARVALHO, Árvore de Decisão / Algoritmo Genético para Tratar o Problema de Pequenos Disjuntos Em Classificação de Dados. Tese (doutorado em Ciências em Engenharia Civil.) – Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2005. Disponível em: <http://www.ipardes.pr.gov.br/sites/ipardes/arquivos_restritos/files/documento/2019-09/deborah_carvalho_tese_2005.pdf>. Acesso: 21 jul. 2021.

GUARDA, A. Inteligência Artificial em Controle e Automação. Disponível em: <https://www.cin.ufpe.br/~pacm/SI/ArvoreDecisaoIndutiva.pdf>. Acesso: 21 jul. 2021.

MENOTTI, D. Árvores de Decisão, 2018. 29 slides. Disponível em: <https://www.inf.ufpr.br/menotti/ci171-182/slides/ci171-arvoresdecisao.pdf>. Acesso: 21 jul. 2021.

7)

GARCIA, S. C. O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde. Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós Graduação em Computação. Porto Alegre, 2003. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/4703/000503532.pdf?sequence=1>. Acesso: 21 jul. 2021.

D. R. CARVALHO, Árvore de Decisão / Algoritmo Genético para Tratar o Problema de Pequenos Disjuntos Em Classificação de Dados. Tese (doutorado em Ciências em Engenharia Civil.) – Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2005. Disponível em: <http://www.ipardes.pr.gov.br/sites/ipardes/arquivos_restritos/files/documento/2019-09/deborah_carvalho_tese_2005.pdf>. Acesso: 21 jul. 2021.

8) MENOTTI, D. Aprendizagem Bayesiana, 2018. 28 slides. Disponível em: <https://www.inf.ufpr.br/menotti/ci171-182/slides/ci171-bayes.pdf>. Acesso: 22 jul. 2021.

KOERICH, A. L. Aprendizagem Bayesiana. 61 slides. Disponível em: <http://www.eletrica.ufpr.br/ufpr2/professor/36/TE808/5-NaiveBayes-AM.pdf>. Acesso: 22 jul. 2021.

9)

ROMÃO, W. Extração de Regras de Associação em C&t: o Algoritmo Apriori. Programa de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina. Florianópolis. Disponível em: [http://www.abepro.org.br/biblioteca/ENEGEP19 99\_A0901.PDF](http://www.abepro.org.br/biblioteca/ENEGEP19%0999_A0901.PDF). Acesso: 22 jul. 2021.

ALMEIDA, D. C. Descoberta Automatizada de Associações com o Uso do Algoritmo Apriori como Técnica de Mineração de Dados. Dissertação (Mestrado em Engenharia da Computação) – Universidade Federal de Goiás. Goiânia, 2011. Disponível em: <https://repositorio.bc.ufg.br/tede/bitstream/tde/966/1/Dissertacao%20Derciley%20Cunha%20de%20Almeida.pdf>. Acesso: 22 jul. 2021.

VASCONCELOS, L. M. R, CARVALHO, C. L. Aplicação de Regras de Associação para Mineração de Dados na Web. Universidade Federal de Goiás. Goiânia, 2004. Disponível em: <https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf>. Acesso: 22 jul. 2021.

11) BECKER, K. ; TUMITAN, D. Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios. In: Joao Eduardo Ferreira. (Org.). Lectures of the 28th Brazilian Symposium on Databases. 1ed.Pernambuco: CIN - UFPE, 2013, v. , p. 27-52. Disponível em: <https://www.inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf>. Acesso: 22 jul. 2021.

<https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>

12)

https://bdm.unb.br/bitstream/10483/19987/1/2017\_RebecaAndradeBaldomir\_tcc.pdf

