FUNÇÕES DE NORMALIZAÇÃO DE DADOS

CREDENCIADA JUNTO AO MEC PELA PORTARIA
N 3.455 DO DIA 19/11/2003

SUMÁRIO

NOSSA HISTÓRIA 3

1. Análise de dados 4

1.1 Dados sujos 4

1.1.1 Tempo de preparação 5

1.1.2 Dificuldade na procura 5

1.1.3 Garbage in, garbage out 5

1.1.4 Correlação não implica causalidade 6

1.1.5 É fácil fazer a análise de dados de forma errada 6

1.2 Processo de análise de dados 7

1.3 Preparando os dados 9

1.4 Limpeza dos dados 9

1.5 Manipulação de dados ausentes 11

1.6 Identificação de anomalias 12

1.7 Transformação dos dados 14

1.8 Redução dos dados 15

1. Construção do modelo e mineração dos dados 17

2.1 Classificação 18

2.2 Regressão 19

2.3 Análise de séries temporais 20

2.4 Sumarização 21

2.5 Padrão de agrupamento 22

2.6 Associações 23

2.7 Descoberta de sequências 24

2.8 Validando modelo de dados 24

3 Normalização em banco de dados estruturado 27

3.1 Vantagens 27

3.2 Definição e características 28

3.3 Formas normais 29

3.4 Caso de estudo 29

3.5 Primeira forma normal (1FN) 30

3.6 Segunda forma normal (2FN) 31

3.7 Terceira forma normal (3FN) 35

3.8 Quarta forma normal (4FN) 36

3.9 Quinta forma normal (5FN) 38
3.9.1 Finalização 38
Referências Bibliográficas 41

NOSSA HISTÓRIA

A nossa história inicia com a realização do sonho de um grupo de empresários, em atender à crescente demanda de alunos para cursos de Graduação e Pós-Graduação. Com isso foi criado a nossa instituição, como entidade oferecendo serviços educacionais em nível superior.

A instituição tem por objetivo formar diplomados nas diferentes áreas de conhecimento, aptos para a inserção em setores profissionais e para a participação no desenvolvimento da sociedade brasileira, e colaborar na sua formação contínua. Além de promover a divulgação de conhecimentos culturais, científicos e técnicos que constituem patrimônio da humanidade e comunicar o saber através do ensino, de publicação ou outras normas de comunicação.

A nossa missão é oferecer qualidade em conhecimento e cultura de forma confiável e eficiente para que o aluno tenha oportunidade de construir uma base profissional e ética. Dessa forma, conquistando o espaço de uma das instituições modelo no país na oferta de cursos, primando sempre pela inovação tecnológica, excelência no atendimento e valor do serviço oferecido.

1. Análise de dados

Os pilares do Big Data estão nos Vs, mas toda inteligência está na análise dos dados. Sem uma análise correta e criteriosa, é impossível gerar insights e direcionar o caminho mais acertado. Por isso ela é uma das etapas mais importantes do processo em que o Big Data está inserido. O processo da análise passa por inspecionar os dados e criar hipóteses para realizar testes com o objetivo de melhorar ou entender um determinado cenário e seus padrões.

Todo e qualquer trabalho em marketing e vendas existe um padrão de comportamento. Com a análise dos dados encontramos esses padrões que nos permitem monitorar qualquer desvio para algo positivo ou negativo. O que quero dizer é que a análise de dados cuida de encontrar esses padrões de comportamento para então monitorá-los e, quando houver qualquer mudança, somos alertados para tomar uma decisão baseada em conhecimentos adquiridos com as análises realizadas.

Como é de se imaginar, as análises são realizadas por analistas, cientistas de dados, growth hackers, entre muitos outros cargos em uma empresa. Profissionais com cargos analíticos estão muito requisitados no mercado em diversos setores como marketing, vendas, tecnologia da informação, e-commerce, advocacia entre outros.

Caso você nunca tenha se aventurado a realizar a análise de dados com o objetivo de extrair informações úteis, você pode não saber que existem algumas particularidades nessa prática.

1.1 Dados sujos

Embora seja comum encontrarmos em livros de análises de dados exemplos que utilizam bases de dados estruturadas e prontas para serem analisadas, no cenário real é muito raro isso acontecer.

Provavelmente a base de dados que você deseja analisar terá dados incompletos, inconsistentes, corrompidos, duplicados, em formatos inadequados, com caracteres indesejados, entre tantas outras questões. Por esse motivo, é necessário

um profissional com habilidades para realizar o tratamento dos dados, antes de a análise ser efetivamente realizada.

1.1.1 Tempo de preparação

Há uma estimativa de que, no processo de análise de dados, 80% do tempo é gasto para limpar e preparar os dados. Parece muito tempo, não? Mas é o que acontece na maioria das análises.

Como cada base de dados possui sua peculiaridade, muitas tarefas de tratamento precisam ser avaliadas e definidas manualmente, não existindo muitos meios para automatizar completamente esse processo. Então, não se espante se você demorar muito tempo nessa etapa. Embora seja oneroso, o tratamento dos dados evita inconsistências nos resultados das análises.

1.1.2 Dificuldade na procura

Analisar uma grande base de dados em busca de padrões pode significar muitas vezes um processo análogo ao de procurar uma agulha em um palheiro. Essa analogia existe pelo fato de que encontrar um padrão diante de uma infinidade de dados é uma tarefa muitas vezes complexa e demorada.

Entretanto, no contexto de Big Data, em que se trabalha com uma avalanche de dados, alguns pesquisadores dizem que o desafio da análise de dados não é somente encontrar a agulha em um palheiro, mas encontrar o que de fato é a agulha. Ou seja, identificar qual pergunta é possível se responder a partir dos dados.

1.1.3 Garbage in, garbage out

Aqui voltamos à importância da qualidade dos dados durante o processo de análise. No contexto de Big Data, é muito comum a utilização de dados em sua forma

bruta, que não passaram por um processo de refinamento. O problema é que, sem um processo de inspeção, pode ocorrer que dados incorretos não sejam descartados ou corrigidos.

Uma vez que esses dados sejam usados na construção de um modelo analítico, o resultado obtido pode não representar a realidade dos fatos. Se uma organização faz a tomada de decisão orientada por esses resultados, ela pode desencadear uma série de ações baseadas em fatos inconsistentes.

1.1.4 Correlação não implica causalidade

Esse é um dos principais fundamentos da estatística: correlação não implica em causalidade! Enquanto, na causalidade, você prova que "o acontecimento A causa o acontecimento B", a correlação apenas indica que "A" e "B" tendem a ser observados no mesmo tempo, mas não há necessariamente uma causalidade entre eles. Pode ser que a correlação seja apenas uma coincidência.

Para inferir uma causalidade, é preciso a realização de testes estatísticos e experimentos controlados que façam essa validação. Se a correlação sempre implicasse causalidade, poderíamos identificar algumas tendências um tanto quanto estranhas, como por exemplo de que, sempre que a venda de sorvetes aumenta, aumenta também o número de afogamentos. Por esse motivo, tenha sempre muito cuidado na interpretação dos dados.

1.1.5 É fácil fazer a análise de dados de forma errada

Isso é um perigo alertado por muitos pesquisadores. As ferramentas de análise de dados disponíveis atualmente facilitaram a construção de inúmeros algoritmos utilizando uma diversidade de dados. Entretanto, um erro cometido ou uma interpretação errada dos dados durante esse processo pode gerar resultados que nos deixam animados, mas que na verdade não condizem com a realidade.

Por esse motivo, é extremamente necessária a validação das respostas obtidas, principalmente quando utilizamos bancos de dados de grande volume, em que as incoerências podem não ser claramente perceptíveis.

1.2 Processo de análise de dados

Quando falamos em Big Data e em análise de dados, é comum ouvirmos palavras como identificação de padrões, modelagem dos dados, detecção de grupos, classificação de dados. Essas atividades são possíveis por meio da utilização de técnicas há muito tempo desenvolvidas, como técnicas estatísticas, matemáticas, de aprendizado de máquina e de mineração de dados.

Pense em uma solução em que um sistema computacional receba informações de sensores instalados em uma fábrica e consiga identificar automaticamente que uma das máquinas usadas está prestes a falhar, mesmo antes de ela ter apresentado problemas perceptíveis ao olhar humano. Imagine também em um sistema computacional que tenha a capacidade de diagnosticar uma doença de forma automatizada, com base na análise dos dados coletados sobre o paciente.

O que podemos perceber de comum nessas duas soluções? Ambas utilizavam meios para fazer predições de forma automatizada. É para soluções como estas que se aplicam as técnicas de aprendizado de máquina.

O foco principal dessa área de estudo é permitir que o computador aprenda, ou seja, que ele seja capaz de organizar seu conhecimento, sem que isso seja explicitamente programado. Quando aplicado à análise de dados, o aprendizado de máquina é utilizado para automatizar a construção de um modelo analítico.

Embora essa técnica tenha sido adotada com maior ênfase somente nos últimos anos pelas organizações, já temos exemplos inspiradores resultantes dessa adoção, tais como:

- Detecção de fraude em transações com cartão de crédito;
- Diagnóstico de doenças;
- Identificação de atividades criminosas;

- Segmentação de clientes;
- Descoberta de genes.

Vamos ao exemplo de uma varejista Grandes Compras, imagine que a equipe de analistas deseja identificar se existe um padrão na compra dos clientes em relação a escolha dos produtos. Será que existem produtos que sempre (ou na maioria das vezes) são adquiridos na mesma compra?

Saber isso é importante, pois gera insights para a criação de campanhas e definição de ofertas. Mas eis o problema. Para realizar essa análise, será necessário observar dados históricos de 5 milhões de registros de compras.

Uma alternativa para esse problema é a adoção de técnicas de mineração de dados. Utilizando técnicas estatísticas, matemáticas e de aprendizado de máquina, a mineração de dados é um campo de estudo com foco na extração de informações úteis e padrões ocultos em conjuntos massivos de dados.

Embora seja similar a uma relação, para se obter sucesso na análise de dados, é preciso estabelecer e seguir um processo sistemático. Existem diversas definições de processos de análise de dados na literatura, tais como o SEMMA (Sample, Explore, Modify, Model, and Assess) e CRISP-DM (Cross Industry Standard Process for Data Mining).

Embora cada processo tenha definições distintas, em geral, eles envolvem as seguintes etapas:

1. **Entendimento do negócio:** aqui são definidas as perguntas, o objetivo da análise de dados e o plano a ser seguido;
2. **Compreensão dos dados**: etapa utilizada para coletar e explorar os dados, aumentando a compreensão sobre sua estrutura, atributos e contexto;
3. **Preparação dos dados:** após a análise exploratória, inicia-se o processo de limpeza, filtragem, estruturação, redução e integração dos dados;

4. **Modelagem dos dados:** envolve as tarefas de seleção dos dados, definição e construção do modelo;
5. **Validação do modelo:** os resultados gerados pelo modelo são avaliados, para verificar se a precisão obtida está satisfatória e coesa;
6. **Utilização do modelo:** após serem validados, os resultados dos modelos são utilizados e monitorados.

1.3 Preparando os dados

Sabe aquele mundo ideal, no qual acessamos um software de análise de dados, inserirmos nossa base, pressionarmos um botão e rapidamente nosso modelo é gerado e os padrões ocultos são revelados? Pois é, infelizmente esse mundo ainda não existe.

Conforme já descrito, a fase de preparação, tratamento ou pré-processamento dos dados é essencial na análise de dados, sendo a tarefa que demanda maior tempo e trabalho. Quando falamos de análise dos dados no contexto de Big Data, essa fase se tornou ainda mais importante, uma vez que muitas vezes os dados usados estão em seu formato original, sem nenhuma "lapidação" realizada sobre eles.

Mas por que será que preparar os dados é algo tão demorado? Confira a seguir algumas das atividades realizadas nessa fase e a resposta para essa pergunta.

1.4 Limpeza dos dados

Está lembrado do termo "garbage in, garbage out"? O processo de limpeza de dados é necessário exatamente para minimizar essa ocorrência, de gerar resultados incorretos devido às "sujeiras" existentes nos dados de entrada. O processo de limpeza requer uma inspeção minuciosa dos dados, bem como a realização de operações de correção e remoção, conforme a necessidade.

A limpeza é feita por meio de um processo de inspeção dos dados coletados. Para isso, é possível aplicar alguns métodos estatísticos que avaliam desvios e, com base em alguns critérios, definem a sua relevância para a análise a ser feita.

Isso significa, na prática, que dados considerados anômalos (valores nulos, inconsistentes, duplicados etc.) serão removidos ou tratados para evitar que causem algum tipo de viés nos insights gerados. No entanto, vale destacar que a limpeza permite também o enriquecimento da base de dados, sugerindo novos parâmetros para a coleta.

Como essa transformação pode afetar inúmeros registros da base de dados, é preciso ter cuidado para não aplicar uma regra que realize a transformação incorretamente. Para evitar essa situação, é indicado testar a operação de limpeza em uma pequena parte dos dados primeiramente, para somente depois aplicá-la em toda a base de dados.

As seguintes perguntas podem auxiliar na identificação de quais operações devem ser realizadas:

- Existem dados duplicados?
- Existem dados com informações incompletas?
- Existem dados com erros de digitação?
- Existem dados iguais representados de diferentes formas?
- Existem dados que violam a regra de negócio?

Embora em alguns casos seja possível realizar uma inspeção manual desses dados, isso pode ser muito custoso, principalmente no contexto de Big Data. Linguagens como R e Python podem ajudar nessas operações, dado que elas possuem pacotes com funções específicas para tratamento de dados, facilitando consideravelmente esse processo.

1.5 Manipulação de dados ausentes

Ao avaliarmos uma base de dados sendo ela específica do seu trabalho ou produzida em áreas externas, como na internet, podemos identificar a ausência de dados, ou seja, registros com informações incompletas. Para exemplificar, vamos criar a situação abaixo onde será exposto pontos relevantes:

Registros podem estar com campos vazios, nulos, terem dados de tipos diferentes e informações incompletas que só atrapalham em uma. Em um sistema de vendas online, se um registro não possuir informação sobre o campo frete, e um registro não possui informação sobre o campo data de vencimento e pagamento. O que fazemos com esses registros em nossa análise?

A opção mais simples nesse caso é eliminar os registros com informações incompletas da análise, pois assim não teremos problemas, correto? Não exatamente.

Imagine se, além desses dois registros, a quantidade de registros com informações incompletas fosse superior a 20% do conjunto total de dados? Parece um desperdício não utilizá-los, não?

Embora seja comum descartar registros com dados ausentes, a adoção dessa prática oferece riscos de gerar estimativas viesadas e inconsistentes, uma vez que os registros descartados podem conter padrões significativos para a análise. Para não descartar os registros com dados ausentes em nossa análise, vamos adotar algumas medidas e técnicas para obtermos o melhor resultado possível neste contexto, lembrando que vamos utilizar algumas abordagens dentre outras existentes:

- Substituir o dado ausente com alguma constante, especificada pelo analista;
- Substituir o dado ausente pela média ou moda do campo;
- Substituir o dado ausente com um valor gerado aleatoriamente a partir de uma distribuição observada;
- Substituir o dado ausente a partir de valores baseados em outras características do registro.

Embora essas abordagens sejam indicadas, é preciso muito cuidado para selecionar qual a mais apropriada, evitando que a substituição gere informações inapropriadas ao conjunto de dados e, consequentemente, à análise.

Leitura complementar

http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:marilia:workshop_estudos_longitudinais:sessao3.pdf

1.6 Identificação de anomalias

Para darmos início à explicação de identificação de anomalias, considere o seguinte cenário a seguir, ainda referente a uma empresa de vendas

Na parte financeira, e na coluna de valores pagos nos produtos, foi encontrado um dado bem diferente de todos os outros registros restantes. Enquanto que os outros registros ficaram com valores entre R$ 100,00 e R$ 400,00, esse teve o valor de compra de R$ 50000,00. Como esse registro apresenta um valor que desvia significativamente do padrão normal do restante dos dados, ele é considerado uma anomalia (do inglês outlier). Mas por que identificar anomalias é uma tarefa importante na preparação de dados?

A detecção de anomalias é importante porque ela permite identificar se existe algum erro na entrada de dados numéricos, bem como nos ajuda a perceber a existência de valores extremos que influenciarão alguns métodos estatísticos, mesmo em casos em que as anomalias correspondam a dados válidos.

Estes registros acabam atrapalhando as métricas e resultados apresentados, influenciando diretamente ou indiretamente, no nosso exemplo caso fizéssemos uma média dos valores recebidos, por produto, categoria, região ou ordenação de mais

vendidos e menos vendidos teríamos dados incoerentes, um ticket médio de venda do produto acabaria sendo mascarado por este valor encontrado a mais.

Quando temos um grande volume de dados, identificar uma anomalia em dados apresentados em formato tabular não é uma tarefa fácil. Como solução, os gráficos podem auxiliar bastante esse processo, como por exemplo, o diagrama de caixa (boxplot) e o gráfico de dispersão (scatterplot), conforme veremos no capítulo seguinte.

Em estatística descritiva, diagrama de caixa, diagrama de extremos e quartis, boxplot ou box plot é uma ferramenta gráfica para representar a variação de dados observados de uma variável numérica por meio de quartis.

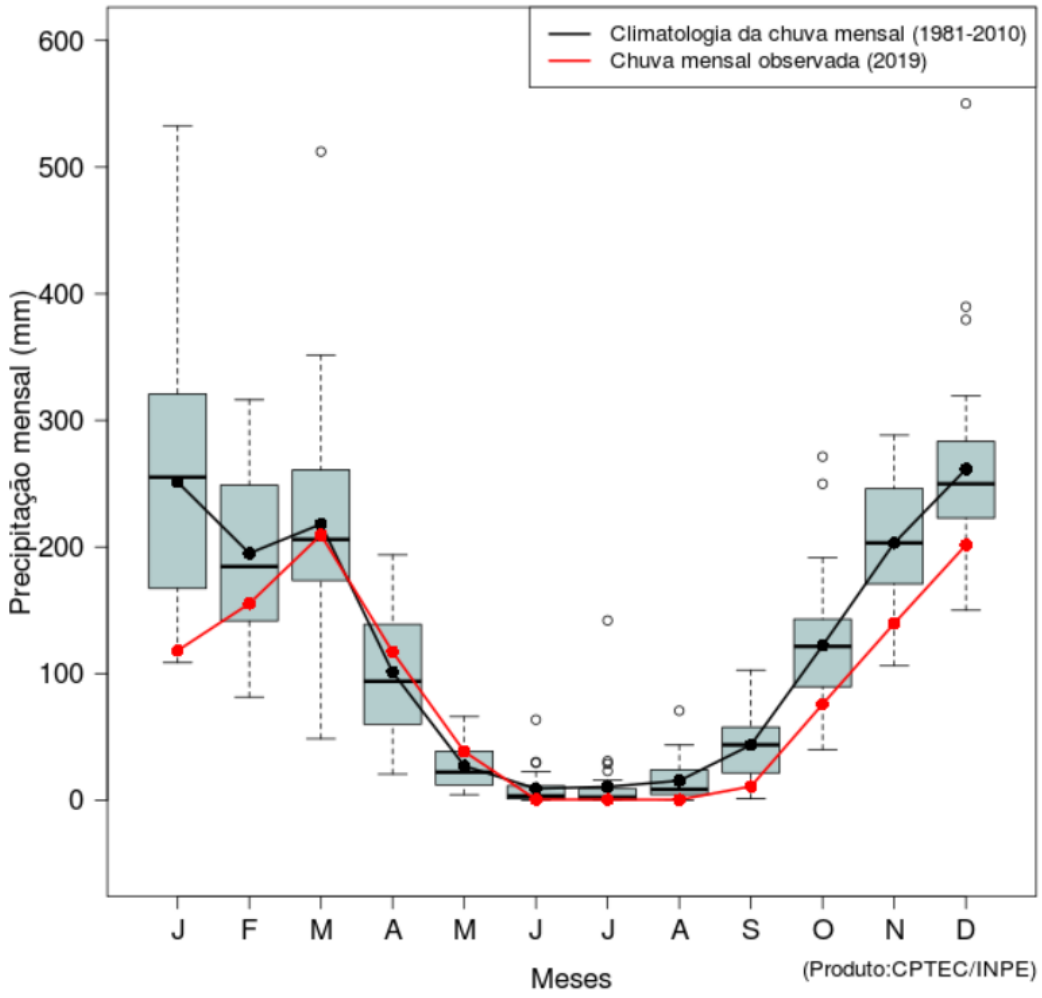

Figura 1.6 – Boxplot de precipitação mensal (1981-2010) : Região 92

Os diagramas de dispersão ou gráficos de dispersão são representações de dados de duas ou mais variáveis que são organizadas em um gráfico. O gráfico de dispersão utiliza coordenadas cartesianas para exibir valores de um conjunto de dados.

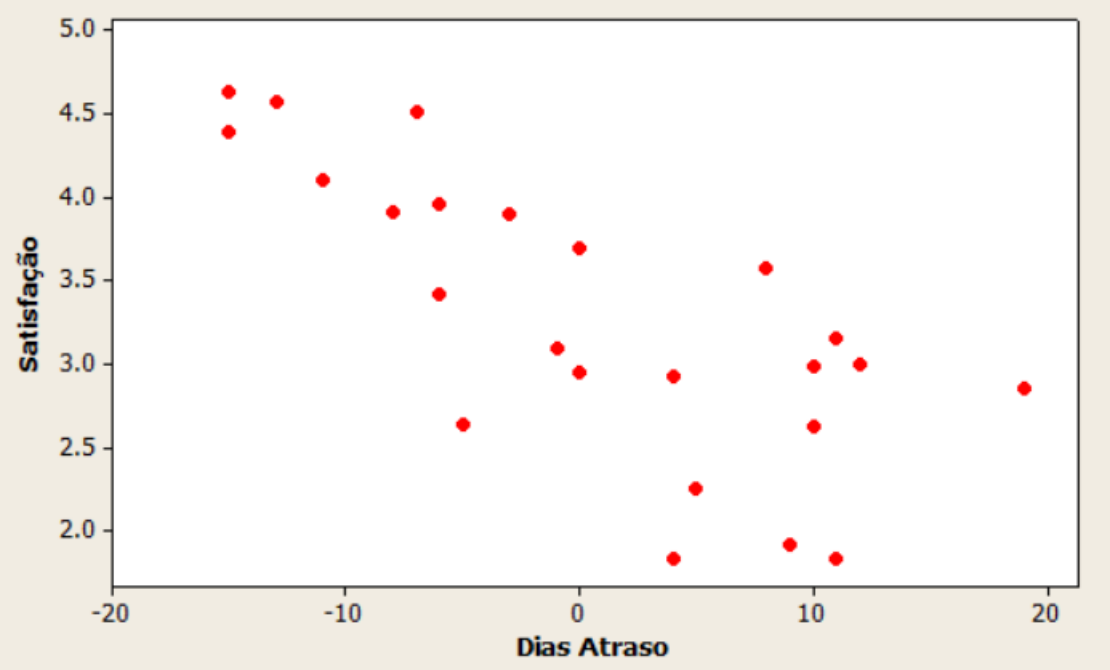

Figura 1.6.1 – Gráfico de dispersão da satisfação em relação aos dias de atraso

1.7 Transformação dos dados

Mesmo em situações nas quais os dados usados para a análise já estejam limpos e sem informações ausentes, pode ser necessário aplicar técnicas de transformação sobre eles.

Digamos que o campo preço apresenta valores bem distintos, utilizando como unidade de medida a moeda Real. Para evitar que essa diferença influencie de forma tendenciosa a construção do modelo, uma transformação muito adotada é a normalização dos dados. O processo de normalização de variáveis numéricas é aplicado para ajustar a escala dos valores das variáveis. Uma das formas de normalização é a transformação linear, também conhecida como normalização min-max, dado que o cálculo é feito com base nos valores mínimo e máximo de cada

atributo no ajuste da escala. Aplicando essa normalização, os registros teriam os seguintes valores:

- Transformação de dados numéricos para categóricos;
- Transformação de dados categóricos para numéricos;
- Agregação de dados, por meio da combinação de dados de diferentes conjuntos em uma única fonte, de forma coerente;
- Criação de novos atributos.

1.8 Redução dos dados

Mesmo com as possibilidades oferecidas pelas tecnologias de Big Data para processar um grande volume de dados, é possível que o processamento de uma base de dados com centenas de variáveis e milhões de registros seja muito caro computacionalmente, resultando em um gargalo de desempenho em alguns algoritmos. Para casos como esses, são aplicadas técnicas de redução e sintetização de dados em busca de reduzir a dimensionalidade dos dados.

Mas ora, se preciso reduzir a base de dados, não basta apenas selecionar uma parte do conjunto de dados? Não é bem assim. Caso façamos a redução dessa forma, não temos garantia de que registros significativos não foram descartados do modelo.

Na verdade, a técnica de redução de dados tem como objetivo gerar uma representação reduzida do conjunto de dados, porém mantendo os mesmos (ou próximo a isso) resultados da análise. Para isso, essa prática requer uma fase de seleção de atributos, identificando quais são irrelevantes para a análise e podem ser removidos da base. Além de reduzir a complexidade do processamento, a eliminação dos atributos irrelevantes também evita que eles atrapalhem o resultado final do modelo.

Uma técnica muito conhecida para a prática de redução de dados é a de Análise de Componentes Principais (Principal Component Analysis — PCA). Essa técnica tem como objetivo detectar a correlação entre as variáveis. E caso seja

detectado uma forte correlação entre elas, cria-se um conjunto menor de combinações lineares dessas variáveis, reduzindo assim a dimensionalidade dos dados.

Conseguiu perceber quantas etapas são necessárias, você até pode conseguir seguir adiante, porém, as possibilidades de encontrar problemas na execução do algoritmo ou nos resultados obtidos são muito grandes. Ou seja, preparar os dados para a análise é um "mal necessário".

1. Construção do modelo e mineração dos dados

Com os dados preparados para a análise, damos início à fase de modelagem dos dados. É nessa etapa que utilizamos um algoritmo para gerar a resposta que estamos procurando. A figura a seguir apresenta uma lista de tarefas comuns em mineração de dados para obtenção dessas respostas. Em geral, essas tarefas podem ser divididas em duas categorias: descritiva e preditiva.

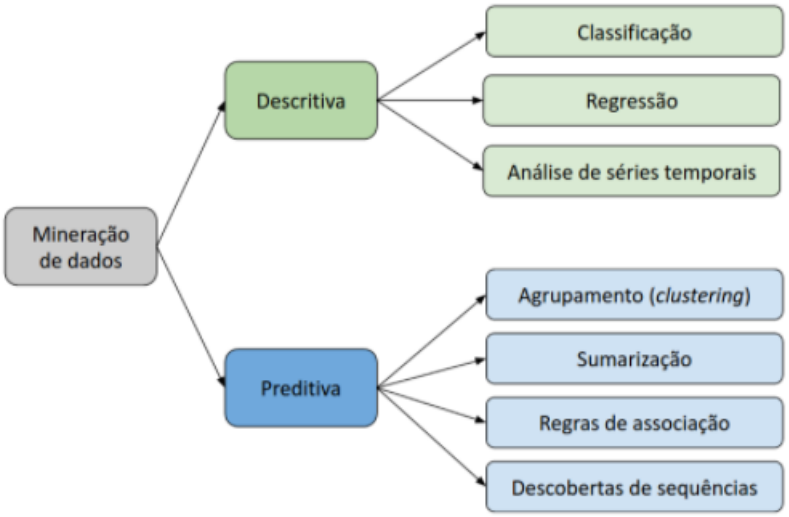

Figura 2.0 – Tarefa de mineração de dados

Enquanto que o foco principal das tarefas da categoria descritiva é caracterizar e apresentar as propriedades de um conjunto de dados de maneira concisa e informativa, o objetivo das atividades preditivas é construir um modelo para prever as propriedades e/ou tendências de um conjunto de dados desconhecido. Mas qual a diferença entre cada uma dessas tarefas? Veja um resumo sobre cada uma a seguir.

2.1 Classificação

Considerado por muitos pesquisadores a tarefa mais comum em mineração de dados, a classificação tem como objetivo utilizar atributos de um objeto para determinar a qual classe ele pertence. Imagine, por exemplo, que uma empresa de vendas online deseja avaliar as transações de compras dos clientes pelo aplicativo e identificar se alguma transação online de cartão de crédito é fraudulenta.

A cada transação é gerado um conjunto de atributos, tais como: data e horário da transação, valor da transação, localização, lista de produtos comprados. A partir desses atributos, o objetivo é classificar a transação como fraudulenta ou idônea. Esse objetivo pode ser alcançado com uso de algoritmos de classificação.

Os algoritmos de classificação necessitam de um conjunto de dados rotulados para gerar o modelo preditivo. Por exemplo, para o cenário de detecção de fraude, devemos utilizar como entrada do algoritmo um conjunto de dados históricos de transações, tendo para cada transação um conjunto de atributos da transação e um atributo especial, que indique se a transação foi rotulada (classificada) como fraudulenta ou não.

A partir desse conjunto de dados, o algoritmo de classificação vai "aprender" quais combinações dos atributos estão associados com cada rótulo, gerando assim o modelo. Após essa etapa, novos registros de transações, agora não rotulados, são enviados ao modelo, que deverá gerar como resultado a predição do rótulo de cada uma delas.

Algoritmos que utilizam dados rotulados na fase de treinamento do modelo são categorizados como algoritmos de aprendizado supervisionado, conforme ilustrado adiante.

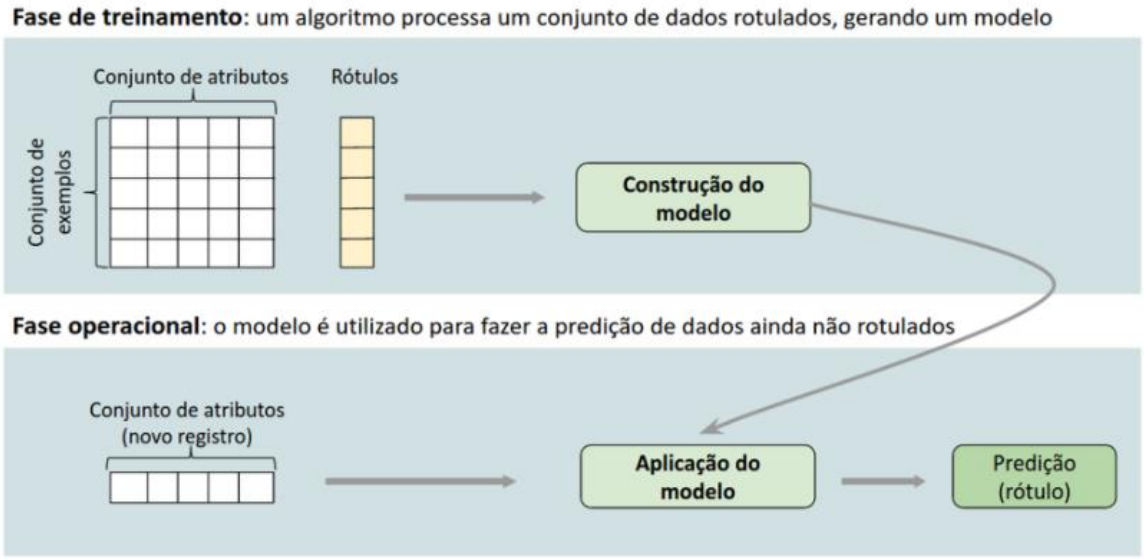

Figura 2.1 – Exemplo de aprendizado supervisionado

São exemplos de algoritmos de classificação: árvores de decisão, classificação Bayesiana, classificação baseada em regras, máquinas de vetores suporte (support vector machines) e redes neurais.

Leitura complementar

https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69

2.2 Regressão

Além da classificação, outra técnica de aprendizado supervisionado é a regressão. A diferença entre essas técnicas é que, enquanto a classificação tenta prever à qual classe pertence uma nova instância, a regressão busca prever um valor numérico contínuo.

Por exemplo, imagine que, em vez de prever a adesão a uma oferta de cartão, a equipe de análise de dados estivesse interessada em prever o total de vendas nos próximos meses. Perceba que aqui a resposta desejada é um valor contínuo, e não um rótulo do tipo "sim/não". Esse valor será obtido com base na análise de valores passados de um conjunto de dados.

São exemplos de algoritmos de regressão: regressão linear simples e múltipla, regressão não linear simples e múltipla.

Regressão múltipla é uma coleção de técnicas estatísticas para construir modelos que descrevem de maneira razoável relações entre várias variáveis explicativas de um determinado processo. A diferença entre a regressão linear simples e a múltipla é que na múltipla são tratadas duas ou mais variáveis explicativas.

2.3 Análise de séries temporais

Essa tarefa é aplicada a bancos de dados de séries temporais, ou seja, bancos de dados que contenham sequências de valores ou eventos armazenados sucessivamente em função do tempo. Tais valores são normalmente obtidos em um mesmo intervalo de tempo, como a cada dia, hora ou minuto.

Por exemplo, no caso de vendas online, esse banco poderia ser o histórico de vendas de uma categoria de produtos ao longo do tempo. A partir da análise de série temporal, torna-se possível observar o comportamento desses dados em relação ao tempo, podendo assim fazer estimativas como a previsão de vendas, controle de estoque, lucro mensal, entre outras.

A tendência de uma série temporal é definida como um padrão de crescimento/decrescimento da variável em um certo período de tempo. Existem testes específicos para a identificação da tendência, como o Teste de Wald e o de Cox-Stuart. Entretanto, uma técnica muito utilizada é o ajuste de uma Regressão Linear Simples para a identificação da inclinação da reta de tendência.

A sazonalidade pode ser definida como padrões de comportamento que se repetem em específicas épocas do ano. Por exemplo, o número de passageiros que

utilizam o transporte aéreo geralmente é maior em períodos de férias escolares do que nos demais meses do ano.

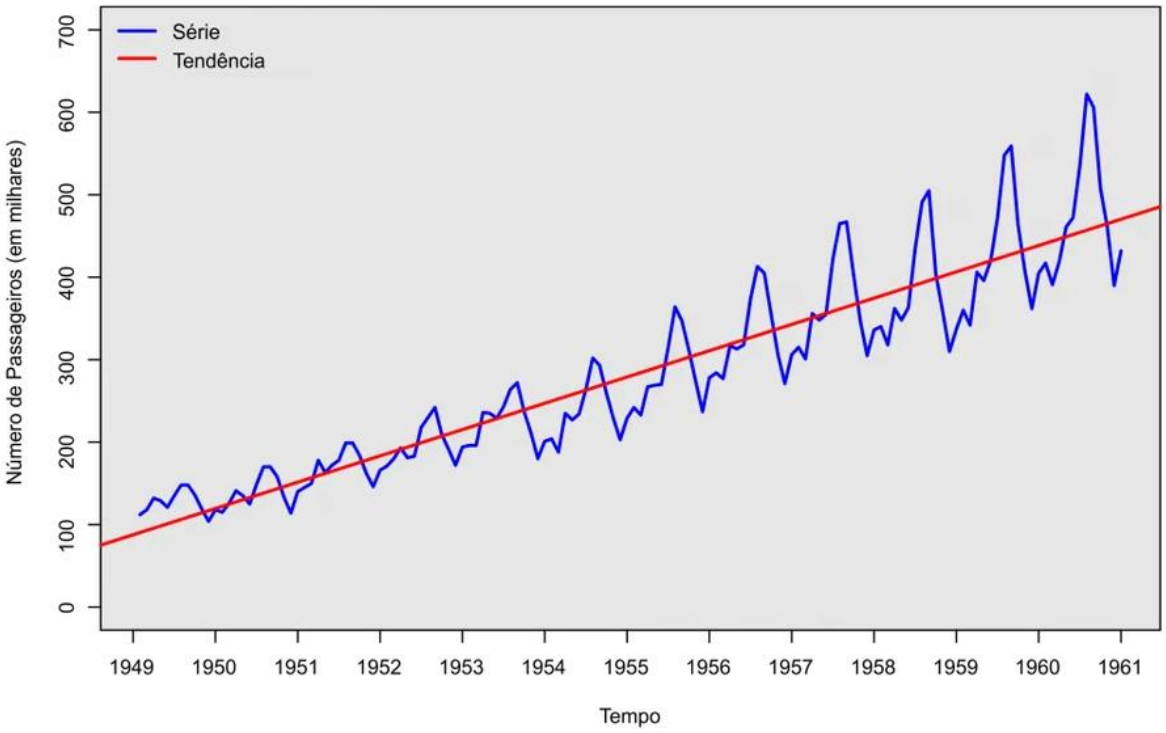

Figura 2.3 – Gráfico linear entre série e tendência

2.4 Sumarização

Essa tarefa descritiva tem como objetivo mapear os dados em subconjuntos, podendo ocorrer em diversos níveis, para fazer uma descrição compacta sobre eles. Aqui são utilizadas desde operações estatísticas básicas (como média, mediana, moda e desvio padrão) até operações mais complexas (como a derivação de regras de sumarização).

Se pensarmos no caso da loja online, por exemplo, a sumarização pode ser útil para analisar dados relacionados à navegação dos clientes no aplicativo. Isso gera informações como a média de minutos permanecidos no aplicativo, de produtos pesquisados e produtos comprados em uma escala diária, semanal e anual.

2.5 Padrão de agrupamento

Lembra-se de que, na tarefa de classificação, é necessário enviar um conjunto de dados rotulados para que o modelo seja treinado? Mas como fazer em situações nas quais não sabemos antecipadamente esse rótulo?

Por exemplo, imagine que a equipe da sua empresa tivesse como objetivo realizar campanhas de marketing e precisasse segmentar seus clientes com base em comportamentos ou características similares. O problema é que a equipe não sabe como "rotular" esses clientes, pois ela não conhece os padrões existentes nos dados para fazer essa inferência. Para situações como essa, em que o objetivo é que um algoritmo seja capaz de detectar padrões ocultos nos dados, utiliza-se a tarefa de agrupamento.

Também conhecido como clustering ou segmentação, nessa tarefa um algoritmo de agrupamento analisa um conjunto de exemplos não rotulados, com foco em determinar se alguns deles podem ser agrupados de acordo com uma medida de similaridade, gerando assim os grupos (ou clusters). Dessa forma, um algoritmo de agrupamento poderia segmentar clientes de acordo com os padrões encontrados, tais como: nível de renda, faixa de idade, preferências de marca, etc. Essa mesma estratégia pode ser adotada em inúmeras outras aplicações, tais como o agrupamento de pacientes com sintomas similares e a classificação de documentos.

Conforme apresentado na figura a seguir, os algoritmos que não utilizam conjuntos de dados rotulados no processo de aprendizado são denominados algoritmos de aprendizado não supervisionado. Isso porque eles não recebem nenhuma indicação em relação aos padrões que devem ser detectados.

Durante a fase de treinamento, um modelo é criado para identificar os grupos com base nas similaridades. Estando o modelo construído, na fase operacional novos registros são enviados ao modelo, que deverá identificar a qual grupo esse registro pertence.

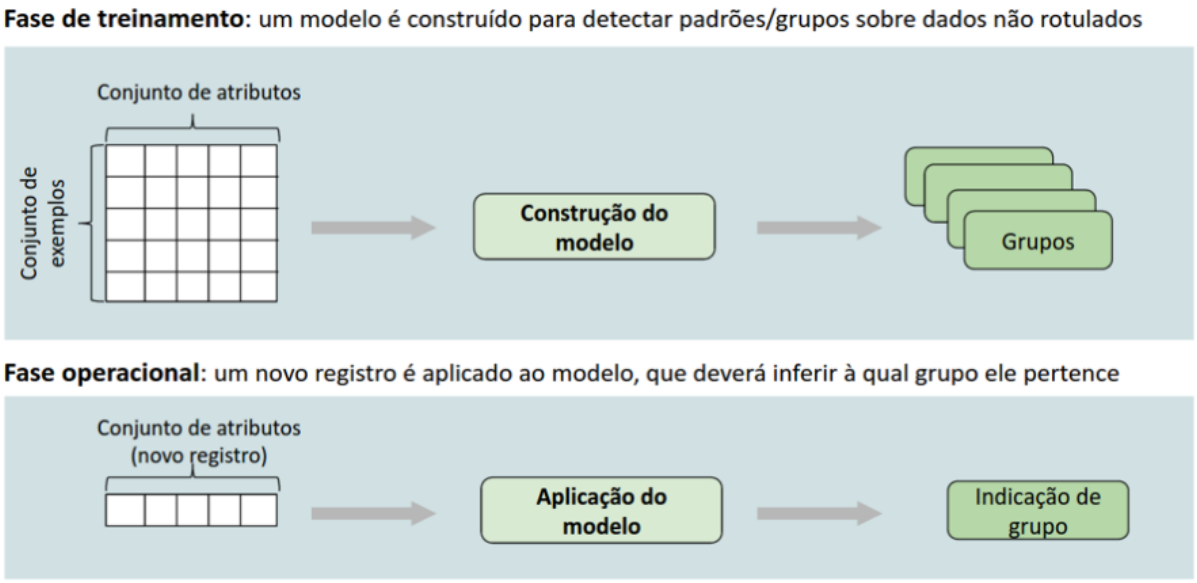

Figura 2.5 – Exemplo de fluxo de aprendizado não supervisionado

São exemplos de algoritmos de agrupamento: k-means, fuzzy c-means e redes neurais.

2.6 Associações

Essa atividade tem como objetivo identificar afinidades em um conjunto de atributos. Ou seja, avaliar como os atributos estão relacionados, gerando como resultado um conjunto de regras de associação.

Um dos problemas mais conhecido nessa tarefa é a análise do cesto de compras. Como exemplo dessa análise aplicado ao cenário de vendas, pode-se citar a análise de itens comprados em uma mesma compra pelos clientes. Como resultado, pode-se obter a seguinte regra de associação: 80% dos clientes que compram leite também compram pão e margarina, sendo o 80% denominado confiança da regra.

Descobrir informações similares a essa pode gerar insights para a organização das prateleiras e definição de itens em promoção, aumentando as chances de vendas casadas.

2.7 Descoberta de sequências

Podendo ser utilizado como uma extensão das tarefas de regras de associação, o objetivo das tarefas de descoberta de sequências é também identificar itens frequentes, porém considerando um determinado período de tempo. Ou seja, de acordo com a maneira com que os dados estão alinhados, com essa tarefa pode-se descobrir uma sequência cronológica em que aconteceram os eventos.

A descoberta de sequências pode revelar que, quando um cliente compra uma cama, ele costuma comprar itens de roupa de cama na sequência. Tal informação pode ser útil para a realização de campanhas de marketing direcionadas.

Essas são apenas algumas das possibilidades para se obter valor a partir da análise de dados. Empresas brasileiras de diversos setores já estão adotando essas técnicas para obter vantagem competitiva no mercado.

Um exemplo notório é a empresa varejista Magazine Luiza. Com um histórico de dados de clientes desde 1998, a empresa realiza a análise de dados para diversas ações, desde segmentação de clientes, modelagem estatística e ações de comunicação dirigida.

Além dos dados transacionais, a empresa também utiliza informações como dados de navegação, informações sobre presença em loja física e informações de crédito. A análise desses dados permite que a empresa consiga ter maior inferência sobre o comportamento do consumidor, para que assim eles possam ser mais assertivos nas abordagens, realizadas por meio do site da loja, e-mail marketing, mala direta e rede social. Como resultado, a empresa obtém maior satisfação do cliente, maior retorno em vendas e diluição nos investimentos de marketing.

2.8 Validando modelo de dados

Já ouviu falar que uma informação errada é pior que nenhuma informação? Essa frase também se aplica à análise de dados.

Imagine o caos que pode ser gerado em uma empresa da área médica que utiliza resultados de um modelo que faz diagnósticos errados sobre seus pacientes.

Ou então, uma empresa que utiliza um modelo preditivo que reconhece grande parte das transações idôneas como sendo fraudulentas? Ou até mesmo o contrário, que considera muitas transações fraudulentas como sendo idôneas.

Quanto mais serviços forem realizados com base em informações obtidas da análise de dados, maior a importância de se validar os modelos e assim ter resultados mais assertivos. Dessa forma, após ter realizado o tratamento dos dados e construído o modelo de acordo com a análise desejada, deve ser iniciado a fase de validação do modelo.

Essa fase tem como objetivo avaliar o desempenho do modelo por meio de dados reais, ou seja, dados que não foram utilizados na fase de treinamento. Existem diversas formas para medir a qualidade de um modelo, dependendo da tarefa e do algoritmo adotado. Entre as possibilidades, as mais comuns são:

- Utilização de medidas estatísticas para validar se os dados de treinamento e o modelo foram corretamente utilizados;
- Separação da base de dados em treinamento e teste, permitindo avaliar o desempenho do modelo antes de usá-lo em um ambiente de produção;
- Avaliação perante profissionais especializados em análise de dados e na área de negócio em que o modelo foi aplicado, para que eles possam determinar se a descoberta ou predição foi condizente e significativa.

Para se ter uma medida mais precisa da qualidade do modelo, é muito comum que mais de uma alternativa seja utilizada. Durante esse processo, diversos aspectos sobre os resultados obtidos são validados, sendo mais comuns as verificações da acurácia, confiabilidade e utilidade do modelo.

A medida de acurácia é utilizada para avaliar quão bem o modelo faz a correlação de um resultado com os atributos dos dados de entrada. Já a confiabilidade tem como objetivo avaliar como o modelo é executado em diferentes conjuntos de dados. Caso ele gere a mesma predição ou encontre os mesmos padrões, independente dos dados testados, ele é considerável confiável. Por fim, mas não menos importante, a utilidade do modelo é uma medida que avalia o quanto o modelo oferece informações significativas ao propósito da análise.

Uma técnica existente para validar a acurácia do modelo é a validação cruzada (cross validation), muito utilizada em algoritmos de classificação. Nessa técnica, omite-se uma observação da base de dados durante as iterações, e a função de classificação é realizada com os dados restantes.

Por exemplo, sendo k = 10, o classificador será treinado 10 vezes. Na primeira iteração, o grupo 1 é utilizado para teste e o restante para treinamento. Na segunda iteração, o grupo 2 é usado para teste e o restante para treinamento, e assim sucessivamente.

Para cada iteração é calculada a taxa de erro de classificação. E ao fim de todas as iterações, calculam-se a média e o desvio padrão das taxas de erro sobre esses grupos.

3 Normalização em banco de dados estruturado

Para se ter um bom conhecimento de banco de dados, não basta apenas saber como criar um banco de dados e saber utilizar comandos SQL, é necessário entender que armazenar dados de uma aplicação requer muito mais que isso. Um banco de dados bem modelado e normalizado, é uma das partes muito importantes na construção de um novo sistema, desde a parte de conceitos e como fazer uma modelagem de dados utilizando Diagrama Entidade x Relacionamento (DER) até os métodos mais avançados.

3.1 Vantagens

Primeiramente, antes de entrarmos na parte prática, veremos algumas das vantagens de despender tempo neste tipo de atividade:

- Se o banco não receber a devida atenção, pode comprometer todo o desenvolvimento de um sistema.
- Um dos principais problemas relacionados a banco de dados é redundância de informações. Quando existirem duas informações que deveriam ser iguais e estão diferentes, em qual confiar?
- Eliminar redundâncias no banco de dados, significa também que o banco de dados será menor.
- Um banco de dados bem modelado permite qualidade das informações, e mais segurança, pelo aumento dos dados num futuro.
- Facilita para que novas pessoas possam entender facilmente quando necessitarem consultar o banco.
- Geração de aplicações mais estáveis.

3.2 Definição e características

O uso do modelo de entidade e relacionamento, apesar de ser prático e usual pode deixar dúvidas quando ao modelo, logo, uma forma mais “científica” de realizar o trabalho é utilizando normalização de dados.

Segundo Oliveira (2002), “a normalização de dados é uma sequência de etapas sucessivas que, ao final, apresentará um modelo de dados estável com um mínimo de redundância”.

Normalização é o processo de organização de dados em um banco de dados. Isso inclui a criação de tabelas e o estabelecimento de relações entre essas tabelas de acordo com as regras projetadas para proteger os dados e tornar o banco de dados mais flexível, eliminando a redundância e dependência inconsistente.

Os dados redundantes desperdiçam espaço em disco e criam problemas de manutenção. Se os dados existentes em mais de um local precisarem ser alterados, os dados devem ser alterados exatamente da mesma maneira em todos os locais. Uma alteração de endereço do cliente é muito mais fácil de ser implementada se esses dados são armazenados apenas na tabela clientes e em outro lugar no banco de dados.

O que é uma "dependência inconsistente"? Embora seja intuitivo que um usuário procure a tabela clientes para obter o endereço de um cliente específico, talvez não seja bom procurar o salário do funcionário que faz a chamada no cliente. O salário do funcionário está relacionado, ou dependente, do funcionário e, portanto, deve ser movido para a tabela funcionários. Dependências inconsistentes podem dificultar o acesso dos dados porque o caminho para localizar os dados pode estar ausente ou quebrado.

Há algumas regras para normalização do banco de dados. Cada regra é chamada de "normal Form". Se a primeira regra é observada, o banco de dados é considerado como "primeira forma normal". Se as três primeiras regras forem observadas, o banco de dados é considerado como sendo "terceiro formato normal". Embora outros níveis de normalização sejam possíveis, a terceira forma normal é considerada o nível mais alto necessário para a maioria dos aplicativos.

Como muitas regras e especificações formais, cenários reais nem sempre permitem conformidade perfeita. Em geral, a normalização requer tabelas adicionais e alguns clientes acham isso complicado. Se você decidir violar uma das três primeiras regras de normalização, certifique-se de que o aplicativo prevê qualquer problema que possa ocorrer, como dados redundantes e dependências inconsistentes.

Um banco de dados dentro dos padrões de normalização reduz o trabalho de manutenção e ajuda a evitar o desperdício do espaço de armazenamento. Se tivermos cadastrado no banco um cliente e tivermos o seu telefone registrado em mais de uma tabela, havendo uma alteração no seu número de telefone, teremos que fazer essa atualização em cada tabela. A tarefa se torna muito mais eficiente se tivermos seu telefone registrado em apenas uma tabela.

3.3 Formas normais

Como mencionado anteriormente, temos conjuntos de regras para determinar com qual forma normal o banco é compatível. Primeiramente, precisamos verificar se encontramos compatibilidade com a primeira forma normal. Caso esteja tudo conforme, analisamos se a segunda forma normal se encaixa e assim sucessivamente.

É importante lembrar que para uma relação atender as exigências de uma forma normal, se faz necessário que esta obedeça as regras da forma normal anterior. A primeira forma normal é exceção pois não existe uma forma normal anterior a primeira.

3.4 Caso de estudo

Para nosso estudo será utilizado modelo de catálogo de CD abaixo, para assim podermos tratar os dados e avançar nas formas normais:

Cód. CD	Nome do CD	Gravadora	Preço	Nº Faixa	Música	Autor	Tempo	CD Indicado
1	Mais do Mesmo	EMI	15,00	1	Será	Renato Russo, Dado Villa e Marcelo Bonfá	2:28	2
				2	Ainda é Cedo	Renato Russo, Dado Villa e Marcelo Bonfá	3:55	
				3	Geração Coca-Cola	Renato Russo	2:20	
				4	Eduardo e Monica	Renato Russo	4:32	
				5	Tempo perdido	Renato Russo	5:00	
				6	Índios	Renato Russo	4:23	
				7	Que país é esse	Renato Russo	2:54	
				8	Faroeste Caboclo	Renato Russo	9:03	
				9	Há tempos	Renato Russo, Dado Villa e Marcelo Bonfá	3:16	
				10	Pais e Filhos	Renato Russo, Dado Villa e Marcelo Bonfá	5:06	
				11	Meninos e Meninas	Renato Russo, Dado Villa e Marcelo Bonfá	3:22	
				12	Vento no Litoral	Renato Russo, Dado Villa e Marcelo Bonfá	6:05	
				13	Perfeição	Renato Russo, Dado Villa e Marcelo Bonfá	4:35	
				14	Giz	Renato Russo	3:20	
				15	Dezesseis	Renato Russo, Dado Villa e Marcelo Bonfá	5:28	
				16	Antes das Seis	Renato Russo, Dado Villa e Marcelo Bonfá	3:09	
2	Bate-Boca	PolyGram	12,00	1	Meninos, Eu vi	Tom Jobim e Chico Buarque	3:25	1
				2	Eu Te Amo	Tom Jobim e Chico Buarque	3:06	
				3	Piano na Mangueira	Tom Jobim e Chico Buarque	2:23	
				4	A violeira	Tom Jobim e Chico Buarque	2:54	
				5	Anos Dourados	Tom Jobim e Chico Buarque	2:56	
				6	Olha, Maria	Tom Jobim, Chico Buarque e Vinícios deMo	3:55	
				7	Biscate	Chico Buarque	3:20	
				8	Retrato em Preto e Branco	Tom Jobim e Chico Buarque	3:03	
				9	Falando de Amor	Tom Jobim	3:20	
				10	Pois é	Tom Jobim e Chico Buarque	2:48	
				11	Noites dos Mascarados	Chico Buarque	2:42	
				12	Sabiá	Tom Jobim e Chico Buarque	3:20	
				13	Imagina	Tom Jobim e Chico Buarque	2:52	
				14	Bate-Boca	Tom Jobim	4:41	

Figura 3.4 – Catálogo de CD

3.5 Primeira forma normal (1FN)

Para normalizar os dados, a primeira regra é eliminar redundâncias, logo, dizemos que um modelo está na primeira forma normal quando não possuir nenhuma repetição. Isso quer dizer que cada dado não é repetido e é indivisível. Deve-se verificar se cada valor é único e aparece somente uma vez na entidade (chamado de tabela no banco de dados físico). Como é possível verificar nos dados do catálogo, há diversos registros (atributos) que se repetem: número da faixa, música, autor e tempo. A partir desta análise, vamos dividir os dados em duas entidades, entidade CD e entidade ITEM_CD:

Cód. do CD	Nome do CD	Gravadora	Tempo Total	Preço	CD indicado
1	Mais do Mesmo	EMI		15,00	2
2	Bate-Boca	PolyGram		12,00	1

Figura 3.5 – Entidade CD

Cód CD	Nº Faixa	Música	Autor	Tempo
1	1	Será	Renato Russo, Dado Villa e Marcelo Bonfá	2:28
1	2	Ainda é Cedo	Renato Russo, Dado Villa e Marcelo Bonfá	3:55
1	3	Geração Coca-Cola	Renato Russo	2:20
1	4	Eduardo e Monica	Renato Russo	4:32
1	5	Tempo perdido	Renato Russo	5:00
1	6	Índios	Renato Russo	4:23
1	7	Que país é esse	Renato Russo	2:54
1	8	Faroeste Caboclo	Renato Russo	9:03
1	9	Há tempos	Renato Russo, Dado Villa e Marcelo Bonfá	3:16
1	10	Pais e Filhos	Renato Russo, Dado Villa e Marcelo Bonfá	5:06
1	11	Meninos e Meninas	Renato Russo, Dado Villa e Marcelo Bonfá	3:22
1	12	Vento no Litoral	Renato Russo, Dado Villa e Marcelo Bonfá	6:05
1	13	Perfeição	Renato Russo, Dado Villa e Marcelo Bonfá	4:35
1	14	Giz	Renato Russo	3:20
1	15	Dezesseis	Renato Russo, Dado Villa e Marcelo Bonfá	5:28
1	16	Antes das Seis	Renato Russo, Dado Villa e Marcelo Bonfá	3:09
2	1	Meninos, Eu vi	Tom Jobim e Chico Buarque	3:25
2	2	Eu Te Amo	Tom Jobim e Chico Buarque	3:06
2	3	Piano na Mangueira	Tom Jobim e Chico Buarque	2:23
2	4	A violeira	Tom Jobim e Chico Buarque	2:54
2	5	Anos Dourados	Tom Jobim e Chico Buarque	2:56
2	6	Olha, Maria	Tom Jobim, Chico Buarque e Vinícios deMorais	3:55
2	7	Biscate	Chico Buarque	3:20
2	8	Retrato em Preto e Branco	Tom Jobim e Chico Buarque	3:03
2	9	Falando de Amor	Tom Jobim	3:20
2	10	Pois é	Tom Jobim e Chico Buarque	2:48
2	11	Noites dos Mascarados	Chico Buarque	2:42
2	12	Sabiá	Tom Jobim e Chico Buarque	3:20
2	13	Imagina	Tom Jobim e Chico Buarque	2:52
2	14	Bate-Boca	Tom Jobim	4:41

Figura 3.5.1 – Entidade ITEM_CD

Até esta etapa podemos pensar nos atributos Código do CD e Número da Faixa, na entidade ITEM_CD, como atributos chaves pois não irão se repetir.

3.6 Segunda forma normal (2FN)

Para normalizar os dados na segunda forma normal, precisamos que todos os atributos não chave dependam unicamente da chave. Deve-se verificar cada atributo

que não é chave se realmente depende da chave. Isso faz com que os dados sejam agrupados em grupos semelhantes (entidades).

Quando encontramos situação em que um atributo não chave não é dependente unicamente da chave, devemos separar os atributos ou criar uma nova entidade com uma nova chave, e essa chave deve ser mantida na entidade original.

No resultado da 1FN, a gravadora, o autor e a música são independentes das suas entidades CD e ITEM_CD (não dependem da chave). Há uma vantagem enorme em separar esses itens em uma entidade nova, visto que ao altera qualquer item seria necessário alterar todas as linhas. É muito mais fácil se precisar alterar em um único lugar, logo o melhor é separar.

Vamos então criar uma entidade para autor, gravadora e música, e alterar as entidades CD e ITEM_CD para vincular com as chaves das novas entidades:

Cód. Autor	Autor
1	Renato Russo
2	Tom Jobim
3	Chico Buarque
4	Dado Villa-Lobos
5	Marcelo Bonfá
6	Ico Ouro-Preto
7	Vinicius de Moraes

Figura 3.6 – Entidade Autor

Veja que ao criar uma nova entidade Autor, é muito mais fácil de adicionar outros autores.

Cód. Gravadora	Nome Gravadora	Enredeço	Site
1	EMI	Rod. Dutra Km 229,8	www.emi-music.com.br
2	PolyGram		

Figura 3.6.1 – Entidade Gravadora

Na entidade Gravadora, também fica mais fácil de adicionar novas informações que estão relacionadas somente a gravadora, como neste caso, foi adicionado os atributos endereço e o site. Agora que criamos uma nova entidade para a gravadora, na entidade CD substituímos o nome da gravadora apenas pela sua chave:

Cód. CD	Nome CD	Cód. Gravadora	Tempo Total	Preço	CD indicado
1	Mais do Mesmo	1		15,00	2
2	Bate-Boca	2		12,00	1

Figura 3.6.2 – Entidade CD

Criamos também uma nova entidade chamada Música com as informações de tempo e autores (Obs: vamos deixar os autores duplicados de propósito para analisar em outra etapa, apesar de que já poderíamos separar em uma nova entidade na 1FN que elimina as duplicidades).

Cód	Música	Tempo	Autor
1	Será	2:28	Renato Russo, Dado Villa e Marcelo Bonfá
2	Ainda é Cedo	3:55	Renato Russo, Dado Villa e Marcelo Bonfá
3	Geração Coca-Cola	2:20	Renato Russo
4	Eduardo e Monica	4:32	Renato Russo
5	Tempo perdido	5:00	Renato Russo
6	Índios	4:23	Renato Russo
7	Que país é esse	2:54	Renato Russo
8	Faroeste Caboclo	9:03	Renato Russo
9	Há tempos	3:16	Renato Russo, Dado Villa e Marcelo Bonfá
10	Pais e Filhos	5:06	Renato Russo, Dado Villa e Marcelo Bonfá
11	Meninos e Meninas	3:22	Renato Russo, Dado Villa e Marcelo Bonfá
12	Vento no Litoral	6:05	Renato Russo, Dado Villa e Marcelo Bonfá
13	Perfeição	4:35	Renato Russo, Dado Villa e Marcelo Bonfá
14	Giz	3:20	Renato Russo
15	Dezesseis	5:28	Renato Russo, Dado Villa e Marcelo Bonfá
16	Antes das Seis	3:09	Renato Russo, Dado Villa e Marcelo Bonfá
17	Meninos, Eu vi	3:25	Tom Jobim e Chico Buarque
18	Eu Te Amo	3:06	Tom Jobim e Chico Buarque
19	Piano na Mangueira	2:23	Tom Jobim e Chico Buarque
20	A violeira	2:54	Tom Jobim e Chico Buarque
21	Anos Dourados	2:56	Tom Jobim e Chico Buarque
22	Olha, Maria	3:55	Tom Jobim, Chico Buarque e Vinícios deMorais
23	Biscate	3:20	Chico Buarque
24	Retrato em Preto e Branco	3:03	Tom Jobim e Chico Buarque
25	Falando de Amor	3:20	Tom Jobim
26	Pois é	2:48	Tom Jobim e Chico Buarque
27	Noites dos Mascarados	2:42	Chico Buarque
28	Sabiá	3:20	Tom Jobim e Chico Buarque
29	Imagina	2:52	Tom Jobim e Chico Buarque
30	Bate-Boca	4:41	Tom Jobim

Figura 3.6.3 – Entidade Música

Por fim, nossa entidade ITEM_CD ficou apenas com as chaves contendo o código do CD, o número da Faixa e o código da música:

Cód. CD	Nº Faixa	Cód. Música
1	1	1
1	2	2
1	3	3
1	4	4
1	5	5
1	6	6
1	7	7
1	8	8
1	9	9
1	10	10
1	11	11
1	12	12
1	13	13
1	14	14
1	15	15
1	16	16
2	1	17
2	2	18
2	3	19
2	4	20
2	5	21
2	6	22
2	7	23
2	8	24
2	9	25
2	10	26
2	11	27
2	12	28
2	13	29
2	14	30

Figura 3.6.4 – Entidade ITEM_CD

3.7 Terceira forma normal (3FN)

Na segunda forma normal vimos, vimos que uma entidade possui todos os atributos não chave dependendo exclusivamente da chave.

Na terceira forma normal, uma entidade possui todos os seus atributos não chave não dependendo de nenhum outro atributo não chave, ou seja, um atributo não pode depender de outro.

É comum que um atributo dependa de outro em cálculos matemáticos ou atributos perdidos na entidade errada. Podemos citar uma nota fiscal com um valor total, o valor total depende de cada produto contido na nota fiscal, logo o valor total seria resultado de uma operação matemática (multiplicar o valor de cada item por sua quantidade e somar o total de todos os itens). Ao armazenar esses valores, segundo Oliveira (2002), estamos dando oportunidade para ocupar mais espaço no banco de dados e permitir a possibilidade de inconsistência de informações, ou seja, do total da nota ser um valor e o resultado da operação matemática ser outro. Em qual valor realmente podemos confiar?

Em alguns casos será possível identificar itens inter-relacionados, logo, deverá ser criado uma nova entidade. Para resolver esses problemas encontrados na 3FN deve-se analisar:

- Se o atributo for resultado de um cálculo matemático, deve ser removido, pois não acrescenta nada ao modelo de dados.
- Se for um grupo de informações relacionadas, deve-se aplicar a 2FN criando uma nova entidade.
- Se for um atributo “perdido” em uma entidade errada, deve-se move-se para a entidade certa.

No nosso exemplo, adicionamos o atributo “Tempo Total” na entidade CD que indica o tempo total de todas as músicas, mas esse atributo é o resultado da soma do tempo de cada música, logo, devemos removê-lo. Assim ficará a entidade CD:

Cód. do CD	Nome do CD	Gravadora	Preço	CD indicado
1	Mais do Mesmo	EMI	15,00	2
2	Bate-Boca	PolyGram	12,00	1

Figura 3.7 – Entidade CD

3.8 Quarta forma normal (4FN)

Pode ocorrer de após a 3FN ainda existir algum tipo de redundância, isso irá acontecer quando um atributo não chave conter diversos valores para uma mesma

chave, isso é chamado de dependência multivalorada, logo a 4FN é a ausência de dependências multivaloradas.

Em palavras mais simples, isso ocorre quando há a repetição de dois ou mais atributos não chave. Não há um caso para aplicar a 4FN com os dados utilizados até o momento, então, para entender, vamos supor que um intérprete pode cantar várias músicas e publicar em diversas gravadoras, teríamos os seguintes dados:

Música	Intérprete	Gravadora
Será	Renato Russo	EMI
Será	Simone	PolyGram
Será	Renato Russo	PolyGram
Imagine	John Lenon	EMI
Imagine	Simone	EMI
Imagine	John Lenon	PolyGram

Figura 3.8 – Intérprete de músicas em diferentes gravadoras

Veja que não é possível criar uma chave com música pois música se repete, não é possível também criar uma chave com música + intérprete pois também se repete, e caso seja colocado os três atributos como chave, até é uma solução, no entanto, redundante, pois haveria repetição de música e intérprete ou música e gravadora.

A solução para este caso é dividir a entidade em duas novas entidades, uma ficaria com a música e o intérprete, e a segunda com a música e a gravadora:

Música	Intérprete
Será	Renato Russo
Será	Simone
Imagine	John Lenon
Imagine	Simone

Figura 3.8.1 – Primeira entidade

Música	Gravadora
Será	EMI
Será	PolyGram

Figura 3.8.2 – Segunda entidade

Após isso, os nomes seriam substituídos pelas respectivas chaves.

3.9 Quinta forma normal (5FN)

Chegar nessa etapa é raro, a 5FN é utilizada após a 4FN quando divide-se uma entidade em duas ou mais entidades e o resultado ainda apresenta dependência multivalorada. Para solucionar as redundâncias que sobraram deve-se dividir novamente em novas entidades.

3.9.1 Finalização

Ao finalizar todas as etapas de normalização de dados, deve-se checar se alguma entidade permite ainda aplicar as mesmas regras de normalização. Deixamos de propósito os autores duplicados na entidade música para exemplificar que ao checar o modelo de dados podemos encontrar formas normais que ainda podem ser aplicadas, então retornaremos na 1FN e eliminaremos essa redundância na entidade Música criando uma nova entidade que faz o relacionamento entre a música e entre cada um dos seus autores:

Nº Faixa	Música	Tempo
1	Será	2:28
2	Ainda é Cedo	3:55
3	Geração Coca-Cola	2:20
4	Eduardo e Monica	4:32
5	Tempo perdido	5:00
6	Índios	4:23
7	Que país é esse	2:54
8	Faroeste Caboclo	9:03
9	Há tempos	3:16
10	Pais e Filhos	5:06
11	Meninos e Meninas	3:22
12	Vento no Litoral	6:05
13	Perfeição	4:35
14	Giz	3:20
15	Dezesseis	5:28
16	Antes das Seis	3:09
1	Meninos, Eu vi	3:25
2	Eu Te Amo	3:06
3	Piano na Mangueira	2:23
4	A violeira	2:54
5	Anos Dourados	2:56
6	Olha, Maria	3:55
7	Biscate	3:20
8	Retrato em Preto e Branco	3:03
9	Falando de Amor	3:20
10	Pois é	2:48
11	Noites dos Mascarados	2:42
12	Sabiá	3:20
13	Imagina	2:52
14	Bate-Boca	4:41

Figura 3.9.1 – Entidade Música

Cód. Música	Cód. Autor
2	1
3	4
3	5
4	1
4	4
2	5
2	6
3	1
4	1
5	1
6	1
7	1
8	1
9	1
9	4
9	5
10	1
10	4
10	5
11	1
11	4
11	5
12	1
12	4
12	5
13	1
13	4
13	5
14	1
14	4
14	5
15	1
15	4
15	5
16	1
16	4
16	5
17	2
17	3
18	2
18	3

Figura 3.9.2 – Entidade Música_Autor

Neste ponto chegamos ao final da modelagem dos dados por meio do processo de normalização de dados.

Considerando as dificuldades de elaborar um projeto para um sistema e planejar toda a modelagem de um banco de dados robusto, ágil e seguro, as regras para normalização de dados aplicadas da forma correta contribuem consideravelmente para a criação de uma boa estrutura das bases de dados relacionais, evitando anomalias de redundância ou perda de informação. Dessa forma, a pessoa que vai analisar a documentação de uma modelagem normalizada consegue abstrair com mais clareza, pois uma vez conhecendo os padrões, a compreensão é facilitada e agiliza todo o trabalho.

Dado o exposto, a aplicação das regras de normalização de dados é altamente recomendada, pois os ganhos são consideravelmente relevantes. Investir um pouco mais de dedicação e tempo trabalhando com um número maior de tabelas trás mais benefícios do que um banco de dados sem a devida organização.

CONTEÚDO COMPLEMENTAR

https://www.youtube.com/watch?v=eRaAMNjCFYw

https://www.youtube.com/watch?v=NpG1Xt8LB_c

Referências Bibliográficas

OLIVEIRA, Celso Henrique Poderoso de. SQL Curso Prático. São Paulo: Novatec Editora Ltda, 2002. 271 p.

CALÇADO, Bruno. Normalização de Banco de Dados. 2010. Disponível em: <http://wiki.sintectus.com/bin/view/SGBD/LicaoNormalizacaoDeBD#Varia_o_da_3FN_Forma_Normal_Boyc>. Acesso em: 11 out. 2020.

SILVA, Eduardo. Banco de Dados — Aula 10. Normalização de Dados. 2010. Disponível em: <http://www.conekti.com/site/index.php/apostilas/>. Acesso em: 10 out. 2020.

SILBERSCHATZ, Abraham; Korth, Henry; Sudarshan, S. Sistema de Banco de Dados. 5. ed. Rio de Janeiro. Elsevier, 2006. 781p.

BARLOW, Mike. Learning to Love Data Science. O'Reilly Media, Inc., 2015.

BRATH, Richard; JONKER, David. Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data. John Wiley & Sons, 2015