# TRABALHO FINAL DO BOOTCAMP

## ROTEIRO APRESENTAÇÃO

### Resumo

#### Faça uma contextualização do problema

Nas organizações contemporâneas, estamos testemunhando um aumento exponencial do volume de dados transacionais disponíveis. Esse vasto e crescente conjunto de informações se tornou um recurso valioso que, quando explorado corretamente através da mineração de dados, oferece uma riqueza de oportunidades e insights.

Entre as diversas técnicas de mineração de dados, a busca por regras de associação permite identificar conexões e padrões ocultos nos dados transacionais, abrindo portas para tomadas de decisões mais inteligentes, estratégias de marketing eficazes, melhorias em processos e, em última instância, maior vantagem competitiva para as organizações.

Nesta apresentação, exploraremos um conjunto de dados com mais de 3 milhões de transações de um supermercado (*o dataset será descrito mais a frente*), que servirá de exemplo das potencialidades do uso desta técnica na área de marketing e da análise de dados.

Nosso foco foi na descoberta de regras de associações que permitissem:

1. **Análise de cesta de compra**: entender a composição típica da cesta de compra dos clientes, identificando padrões ou associações entre itens frequentemente comprados juntos em uma base de dados de transações de compras. Com essa informação, as organizações podem criar promoções e descontos específicos para incentivar a compra de conjuntos de itens frequentemente adquiridos juntos;
2. **Recomendação de produtos**: analisar padrões de compra de clientes e criar sistemas de recomendação personalizados. Por exemplo, se um cliente costuma comprar um determinado produto X, o algoritmo pode identificar outros itens frequentemente comprados em conjunto com X e sugerir esses produtos como recomendações, aumentando as chances de venda cruzada.

Neste contexto, a análise da Cesta de Compras é um excelente ponto de entrada para pessoas e organizações que desejam explorar a ciência de dados. Utiliza aprendizado sem supervisão e exige pouco em termos de recursos e uma quantidade limitada de preparação e limpeza de dados. A barreira de entrada é baixa em termos de habilidade matemática. Na verdade, o cálculo não vai além da divisão simples e teoria da probabilidade básica.

Cabe observar que esses são apenas alguns exemplos de como a mineração de dados pode ser aplicada em marketing.

#### Pesquisa bibliográfica: apresentar sucintamente dois artigos que abordam o tema escolhido descreva a abordagem de cada um

Selecionamos dois artigos como referência para o nosso projeto:

O artigo “**Fast Algorithms for Mining Association Rules**”, escrito por Rakesh Agrawal e Ramakrishnan Srikant e publicado em 1994, introduz o algoritmo Apriori para mineração eficiente de regras de associação em grandes conjuntos de dados. Fonte: <https://www.vldb.org/conf/1994/P487.PDF>

O artigo aborda o desafio de descobrir regras de associação em conjuntos de dados transacionais de tamanho considerável. Regras de associação são declarações do tipo “**Se X, então Y**”, onde X e Y são conjuntos de itens, e o objetivo é identificar associações frequentes entre os itens que ocorrem em transações.

O algoritmo Apriori é apresentado como uma abordagem eficiente para encontrar conjuntos frequentes e derivar regras de associação. O Apriori baseia-se no princípio apriori, que sugere que, se um conjunto de itens é frequente, então todos os seus subconjuntos também serão frequentes. Isso permite reduzir o espaço de busca, evitando a exploração de conjuntos infrequentes.

O artigo “**Data Mining: Market Basket Analysis with Apriori Algorithm**”, escrito por Yenwee Lim e publicado em 2022, explica de forma didática o uso do algoritmo Apriori aplicado em supermercados e apresenta exemplos. Fonte: <https://towardsdatascience.com/data-mining-market-basket-analysis-with-apriori-algorithm-970ff256a92c>

#### Descreva a base de dados que você irá trabalhar

A base de dados trabalhada contém, como supracitado, mais de 3 milhões de registros. Assim, trata-se de uma base com uma ótima quantidade de dados. Apesar de robusta, apresenta alguns problemas, como no número de pedidos (orders), por exemplo: é possível perceber que há uma limitação neste dataset de 99 pedidos: se um consumidor tiver feito 120 pedidos, por exemplo, ele muda para 99. Qualquer valor acima de 99, ele considera como 99 pedidos.

#### Mostre a etapa de análise de negócios, apresentando inclusive a “Análise Exploratória de Dados”

A finalidade da Análise Exploratória de Dados (EDA - Exploratory data analysis - em Inglês) é examinar os dados previamente à aplicação de qualquer técnica estatística. Desta forma, o analista consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas. Neste projeto, cada base de dados (total de 5) passou por processamento usando pandas.  
Análise de negócio: através do uso do algoritmo apriori, conseguimos ver os hábitos de compras dos clientes, entender melhor esse histórico de vendas e tirar diversos insights para recomendação de produtos, facilitando a vida de quem compra e aumentando as vendas e o ticket médio para o varejista.

### Método

#### Qual o experimento proposto?

A partir do dataset selecionado desenvolvemos dois experimentos que permitissem:

1. **Análise de cesta de compra**: entender a composição típica da cesta de compra dos clientes, identificando padrões ou associações entre itens frequentemente comprados juntos em uma base de dados de transações de compras;
2. **Recomendação de produtos**: analisar padrões de compra de clientes e criar sistemas de recomendação personalizados.

#### Quais os resultados são esperados?

Os resultados esperados são a descoberta de padrões ou associações entre itens frequentemente comprados juntos na base de dados de transações de compras.

#### Como os dados são preparados?

Aqui temos uma etapa de preparação, onde tivemos que transformar os dados em categóricos para ser usado no algoritmo Apriori

O algoritmo Apriori segue uma abordagem iterativa e passo a passo para descobrir conjuntos de itens frequentes e gerar regras de associação. As etapas principais do algoritmo são:

1. **Preparação dos dados**: O primeiro passo é coletar os dados relevantes, como registros de compras de clientes contendo listas de itens adquiridos. Os dados devem ser organizados em uma estrutura que indique quais itens foram comprados por cada cliente em transações separadas.

2. **Definir o suporte mínimo**: O suporte mínimo é um parâmetro essencial no algoritmo Apriori. Ele representa a frequência mínima com que um conjunto específico de itens deve aparecer nas transações para ser considerado relevante. Escolha um valor apropriado para o suporte mínimo com base no tamanho do conjunto de dados e nos objetivos de análise.

3. **Encontrar os itens frequentes**: O algoritmo Apriori começa identificando os itens individuais que atendem ao suporte mínimo. Em seguida, combina esses itens para formar conjuntos maiores e verifica sua frequência nas transações. O processo continua até que nenhum conjunto novo atinja o suporte mínimo.

4. **Gerar regras de associação**: Após identificar os itens frequentes, o Apriori usa esses conjuntos para gerar regras de associação. As regras de associação são declarações do tipo "Se o cliente comprar X, então é provável que também compre Y." Essas regras são baseadas na frequência conjunta de itens nas transações.

5. **Avaliar e selecionar regras relevantes**: Nem todas as regras geradas pelo Apriori são igualmente úteis. É importante avaliar e selecionar aquelas que sejam relevantes para os objetivos de marketing específicos. Algumas métricas comuns para avaliar as regras incluem confiança, suporte e lift.

6. **Implementar as estratégias de marketing**: Com as regras de associação relevantes em mãos, o profissional de marketing pode implementar estratégias específicas.

#### Qual modelo computacional foi escolhido? Por quê?

A metodologia mais comum para realizar uma análise de cesta de compras é a mineração de regras de associação. Esse processo envolve a descoberta de padrões frequentes em conjuntos de transações de compras, onde cada transação é composta por um conjunto de itens.

O algoritmo Apriori é um método clássico para descobrir regras de associação, baseado na *suposição apriori*, que afirma que se um conjunto de itens está presente em uma transação, então todos os seus subconjuntos também estão presentes. Essa abordagem ajuda a reduzir o espaço de busca e torna a análise mais eficiente

O processo do algoritmo Apriori começa com a identificação dos itens individuais frequentes. Em seguida, ele gera conjuntos maiores de itens candidatos, combinando os conjuntos frequentes encontrados anteriormente. O algoritmo continua iterando e gerando conjuntos cada vez maiores até que nenhum conjunto novo atinja o suporte mínimo definido pelo usuário.

Uma vez que todos os conjuntos frequentes são identificados, o Apriori usa esses conjuntos para extrair regras de associação candidatas. As regras são geradas combinando os itens frequentes em diferentes conjuntos, formando antecedentes e consequentes. As regras candidatas são então avaliadas com base em métricas como confiança e suporte para identificar as regras mais relevantes.

Adotamos o algoritmo Apriori por ser uma técnica poderosa para a descoberta eficiente de regras de associação em grandes volumes de dados transacionais. Sua abordagem baseada no princípio apriori simplifica o processo de busca e contribui para sua eficiência na mineração de conjuntos frequentes e regras relevantes.

O artigo "Fast Algorithms for Mining Association Rules" apresenta experimentos e resultados que demonstram a eficiência do algoritmo Apriori em comparação com abordagens anteriores de mineração de regras de associação. O Apriori mostrou-se capaz de lidar com conjuntos de dados grandes de forma mais rápida e eficiente, tornando-o uma ferramenta valiosa para análise de mercado e tomada de decisões em marketing.

### Resultados: apresente os resultados de acordo com sua tarefa

#### Recomendação

##### Figuras de mérito para a qualidade de recomendação

Para avaliar a importância das associações descobertas, trabalhamos com as métricas de Suporte, Confiança e Lift.

* Suporte é a frequência absoluta de um conjunto de itens na base de dados.
* Confiança mede a probabilidade condicional de que um item esteja presente na transação, dado que outro item também está presente.
* Lift é uma medida que mostra o quão mais provável é que os itens estejam sendo comprados juntos em relação à probabilidade de serem comprados independentemente.

###### Suporte, Confiança e Lift

As métricas de Suporte, Confiança e Lift são usadas para avaliar a importância e relevância das regras de associação descobertas por algoritmos de mineração de dados, como o Apriori. Essas métricas são fundamentais para ajudar a identificar quais regras são mais interessantes e significativas para uma análise de mercado ou uma estratégia de marketing.

**Suporte**

O Suporte de uma regra de associação é a proporção de transações que contêm tanto o antecedente (X) quanto o consequente (Y) da regra em relação ao total de transações. Em termos matemáticos, o Suporte é calculado como:

Suporte(X → Y) = (Número de transações contendo X e Y) / (Número total de transações)

Essa métrica indica a frequência com que a regra ocorre no conjunto de dados. Quanto maior o suporte, maior é a frequência com que o antecedente e o consequente são comprados juntos, o que sugere uma associação mais forte. Geralmente, o suporte varia entre 0 e 1, sendo 1 uma associação perfeita, onde o antecedente e o consequente aparecem juntos em todas as transações.

**Confiança**

A Confiança de uma regra de associação mede a probabilidade de o consequente (Y) ocorrer em uma transação, dado que o antecedente (X) também ocorreu. Em termos matemáticos, a Confiança é calculada como:

Confiança(X → Y) = (Número de transações contendo X e Y) / (Número de transações contendo X)

Essa métrica representa a força da associação entre X e Y. Quanto maior a confiança, maior é a probabilidade de que, se o antecedente ocorrer, o consequente também ocorrerá. A confiança varia entre 0 e 1, sendo 1 uma associação perfeita, onde o consequente sempre ocorre quando o antecedente está presente.

**Lift**

O Lift (ou Índice de Elevação) é uma métrica que compara a confiança de uma regra com a frequência esperada do consequente (Y) sem a influência do antecedente (X). Em termos matemáticos, o Lift é calculado como:

Lift(X → Y) = Confiança(X → Y) / (Suporte(Y))

O Lift indica o quão mais provável é o consequente ocorrer quando o antecedente está presente, em comparação com sua probabilidade de ocorrer independentemente. Um Lift maior que 1 indica uma associação positiva, ou seja, o antecedente e o consequente têm uma relação mais forte do que o esperado aleatoriamente. Lift igual a 1 indica que a regra é independente, e Lift menor que 1 indica uma associação negativa, ou seja, o consequente ocorre menos frequentemente quando o antecedente está presente.

Em resumo, o Suporte, a Confiança e o Lift são métricas cruciais para avaliar a relevância e a força das regras de associação em análises de mercado e estratégias de marketing. O Suporte mede a frequência da regra, a Confiança avalia a força da associação e o Lift compara a associação entre o antecedente e o consequente com sua ocorrência esperada independentemente.

##### 

##### Interpretação dos resultados (pós-validação)

\*Interpretação de rules\_df--(insight):\*

\* rules\_df[0] ---> Se a pessoa está fazendo compra no horário de pico (peak\_hours) e é sua primeira compra, ela tem preferência de comprar no final de semana. Isso daria a oportunidade do mercado realizar uma promoção para esses novos clientes, pois essa tendência de comprar no final de semana e horário de pico é uma oportunidade.

\* rules\_df[1] ---> Se a compra esta sendo feita no final de semana e está entre a primeira ou terceira compra do cliente, ele fará a compra no horário de pico e é a primeira ordem.

\* rules\_df[32] ---> Se o cliente esta fazendo entre a 4º ou 5º compra, ela tem a tendência de dá um intervalo de 30 dias para comprar novamente. Outra oportunidade para tentar atrair o cliente para voltar ao mercado com mais frequência, talvez uma promoção ou e-mail(sms) marketing ajude.

\* rules\_df[34] ---> Se a pessoa está fazendo as compras no horário de pico e tem um intervalo médio de 30 dias, ela deve está na 4º ou 5º compra. Bem parecido com a regra acima. O Mercado poderia dar uma sugestão para ela vir num horário com menos movimento.

\* rules\_df[14] e [16], provam que para a primeira ordem as pessoas não tem preferência se é durante a semana ou final de semana e sim horário de pico.

\*\*Conclusão-----> temos diversas oportunidades, temos que observar bem cada regra. Podemos também, inclusive, mudar parâmetros como confiança e lift e ver o comportamento das regras.\*\*

### Conclusão

#### Os resultados esperados foram encontrados?

#### Sim, encontramos oportunidades de melhoria no mercado. Pontos iniciais de campanhas de marketing, mudança no layout de lojas físicas etc.

#### Quais são as possíveis melhorias no modelo?

#### Fazer um estudo mais aprofundado, mudando as métricas (suporte, confiança e lift) e migrar de algoritmo (o apriori) consome muita memória para processamento.

#### Próximos passos sugeridos?

**Experimentar com outros datasets**

**Experimentar outro algoritmo**

O **algoritmo FP-Growth (Frequent Pattern-Growth)** é uma alternativa ao Apriori que utiliza uma estrutura de árvore chamada FP-Tree para representar os padrões frequentes. Ele evita a geração explícita de conjuntos de itens frequentes, o que pode ser computacionalmente custoso em conjuntos de dados grandes.

xxxx

Desenvolva seu projeto seguindo as etapas do método CRISP-DM:

1. Análise do negócio
2. Compreensão dos Dados
3. Preparação dos dados
4. Modelagem (usando modelos de inteligência artificial)
5. Avaliação do modelo
6. Encapsulamento (Deployment) do Modelo (Opcional)