



ISEL
INSTITUTO SUPERIOR DE
ENGENHARIA DE LISBOA

Final Project B

Aprendizagem e Mineração de Dados

Grupo 3:

Duarte Valente | A47657

João Valido | A51090

Docente:

Eng. Paulo Trigo

Curso: MEIM

2023

Índice

Introdução	3
a) Descrição	3
b) Âmbito	3
c) Cenário	4
Desenvolvimento	5
a) Análise	5
b) Processo	5
c) Orange	8
Conclusão	10

Índice de ilustrações

Figura 1 - Distribuição de visitante e sessões	6
Figura 2 - Distribuição de sessões e visitantes	6
Figura 3 - Ficheiro .txt	7
Figura 4 - Combinação 1	8
Figura 5 - Combinação 2	9

Introdução

a) Descrição

Este trabalho teve como foco principal aprofundar a relação entre dados, informação e conhecimento, procurando inserir esses conceitos em cenários concretos e utilizando técnicas avançadas para converter dados brutos em percepções valiosas e práticas.

A iniciativa foca-se no aprofundamento e reforço das competências em mineração de dados, utilizando uma variedade de ferramentas e algoritmos especializados. Primeiramente, realizou-se uma análise detalhada na vertente conceptual, dedicando-se tempo ao estudo cuidadoso das definições e diferenças entre dados, informação e conhecimento, e analisando exemplos concretos das suas interações. Posteriormente, no contexto das técnicas de mineração de dados, recorreu-se a algoritmos reconhecidos de classificação, como J4.8/C4.5 e ID3, para a segmentação precisa dos dados; técnicas como Apriori foram utilizadas para identificar padrões e conexões frequentes nos conjuntos de dados; e métodos de agrupamento foram empregues para dividir os dados com base nas suas características distintivas. A análise estatística desempenhou um papel vital, usando a regra de Bayes e métodos de indução de regras para esclarecer e deduzir associações em conjuntos de dados diversificados. Em termos de operacionalização, a gestão de dados foi otimizada com o PostgreSQL, a ferramenta Orange foi essencial na identificação de tendências e a codificação de algoritmos em Python aproveitou as funcionalidades das bibliotecas, em particular scikit-learn.

Em conclusão, prevê-se que este trabalho consolide uma compreensão sólida dos conceitos chave e estimule a aplicação eficiente de técnicas de mineração de dados, promovendo o desenvolvimento contínuo de habilidades em análise e descoberta de conhecimento.

b) Âmbito

Este relatório insere-se no âmbito da realização do projeto final B da unidade curricular de Aprendizagem e Mineração de Dados, do Mestrado em Engenharia Informática e Multimédia do DEETC do ISEL.

c) Cenário

No cenário apresentado durante a reunião entre "SoftKnow" e "We-Commerce", foi destacada a importância e o potencial dos dados coletados pela "We-Commerce". Ao navegar pelo site da empresa, os visitantes são identificados por meio de cookies, associados a um identificador único global denominado "GUI".

Com o retorno desses visitantes, novos identificadores de sessão são gerados, permitindo rastrear diferentes atividades ao longo do tempo. Para os usuários registrados, um identificador adicional, "user_gui", é atribuído. Cada interação, ou evento, é meticulosamente registrado, incluindo detalhes como identificador do produto acessado, timestamp e informações sobre a página visitada. Embora os dados forneçam uma visão rica, também há registros irrelevantes que necessitam de ser excluídos. O dataset principal, "z_dataset_JAN_updated.csv", contém um volume considerável de aproximadamente 420.000 eventos, com uma amostra inicial, "z_datasetSample_JAN.csv", facilitando uma análise preliminar.

Em resumo, os dados do "We-Commerce" são uma fonte valiosa para insights sobre comportamento do usuário e eficácia de campanhas, requerendo abordagens analíticas avançadas para transformar informações brutas em decisões estratégicas eficazes no e-commerce.

Desenvolvimento

a) Análise

A análise detalhada do conjunto de dados fornecido pelo "We-Commerce" revela múltiplas dimensões de informação que podem ser exploradas para obter perceções valiosas sobre o comportamento dos visitantes e a eficácia das campanhas.

O **tracking_record_id** serve como um identificador único para cada evento, garantindo a integridade dos registos. O atributo **date_time** oferece uma visão temporal, permitindo análises sazonais ou diárias de atividades. A identificação da **company** ajuda a entender quais empresas ou marcas têm maior interação com os visitantes. O **product_gui** e **link** permitem uma análise profunda dos produtos mais visitados e das páginas que geram mais tráfego. O **refer** fornece insights sobre as fontes externas que direcionam o tráfego para o site, enquanto o **browser** ajuda a compreender as preferências tecnológicas dos visitantes. O **campaign_id** destaca o desempenho das campanhas promocionais. Os identificadores **cookie_id** e **session_id** são cruciais para rastrear a jornada do visitante, desde a primeira visita até as interações subsequentes, mesmo que sejam em sessões diferentes ou após um intervalo de tempo. Finalmente, o **user_gui** permite uma análise focada nos usuários registrados, enquanto o **ip_address** pode oferecer perceções sobre a geolocalização dos visitantes.

Em resumo, cada atributo deste conjunto de dados possui um papel vital na construção de uma narrativa detalhada sobre o engajamento do usuário e a eficácia das estratégias de marketing do "We-Commerce".

b) Processo

A escolha de regras de associação é uma técnica que identifica padrões frequentes, associações, correlações ou relações causais entre conjuntos de itens. Escolher a regra de associação correta é fundamental para obter perceções significativas e tomar decisões informadas. Este capítulo descreve o processo de escolha de uma regra de associação, desde a obtenção dos dados até a análise e eleição da regra final.

Primeiramente, a base de dados foi criada no PostgreSQL, depois as views para calcular estas regras de associação. Assim foi criada uma tabela com o nome de Track que regista todos os eventos ao armazenar todos os parâmetros explicados no ponto anterior. De seguida importámos os dados a partir do ficheiro z_dataset_JAN_updated.csv para a base de dados através do comando \COPY do PostgreSQL.

Para efetuar uma primeira análise dos dados acabámos por fazer as algumas interrogações e por criar as seguintes views:

- Calcular o número total de eventos ao contar o número total de instâncias da tabela Track, obtendo um total de 415863 eventos;
- Calcular o número de visitantes, através da contagem dos distintos cookie_id da tabela Track, obtendo um total de 263137;
- Criação de uma view, que permite analisar a relação entre a quantidade de sessões e a quantidade de visitantes das mesmas, tendo verificado que existiram 240911 utilizadores que apenas iniciaram 1 sessão e que existiu um utilizador que iniciou 59 sessões;

Numero de sessões	Numero de visitantes
1	240911
2	15695
3	2529
...	...
48	1
51	1
59	1

Figura 1 - Distribuição de visitante e sessões

- Criação de uma view que permite analisar o número de eventos por sessão em relação ao número de visitante;

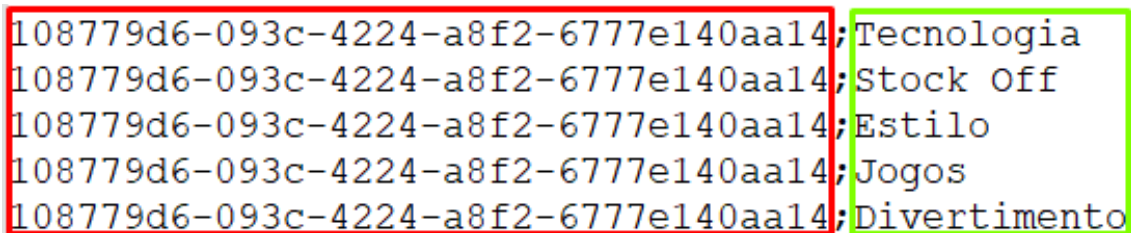
Eventos por sessão	Número de visitantes
1	250708
2	8924
3	3962
...	...
367	1
541	1
2522	1

Figura 2 - Distribuição de sessões e visitantes

Os resultados foram comparados com os do ficheiro z01_aggregatedView.txt e verificámos que estes estavam corretos.

Assim que verificados os resultados, foi criada uma view que filtra o dataset original de forma a reduzir a sua complexidade. Para efetuar esta filtragem foi sugerido escolher apenas os eventos dos visitantes com um número de sessões entre 5 e 30. Para tal foi ainda calculado o número de visitantes que se encontravam nestas condições, tendo chegado a um total de 2911 visitantes, o que contabilizou um dataset reduzido com 47374 eventos.

Assim, de maneira a exportar a view do PostgreSQL, foi utilizado o comando COPY do PostgreSQL, que armazena num ficheiro .txt os dados dos atributos cookie_id e product_gui do dataset reduzido, como representado na figura 3, onde a vermelho estão as cookie_id e a verde os product_gui.



```
108779d6-093c-4224-a8f2-6777e140aa14;Tecnologia
108779d6-093c-4224-a8f2-6777e140aa14;Stock Off
108779d6-093c-4224-a8f2-6777e140aa14;Estilo
108779d6-093c-4224-a8f2-6777e140aa14;Jogos
108779d6-093c-4224-a8f2-6777e140aa14;Divertimento
```

Figura 3 - Ficheiro .txt

Partindo do ficheiro .txt, foi criado um script com a linguagem Python, que possibilita normalizar e transformar o ficheiro em .basket, formato específico utilizado pelo Orange para conseguir obter os dados da melhor maneira possível. A execução do script conta com os seguintes passos:

- Recebe como input o ficheiro z_dataset_sample_OUT.txt;
- Executa a função generateBasket():
 - Inicia uma estrutura vazia basket;
 - Recebe e lê o ficheiro z_dataset_sample_OUT.txt;
 - Processa cada linha de forma a obter o cookie_id e o product_gui;
 - Normaliza os valores obtidos;
 - Adiciona os valores normalizados à estrutura basket caso estes valores ainda não estejam presentes na mesma;
 - Retorna a estrutura basket;
- Executa a função generateDataFile_basket():
 - Constrói o ficheiro zz_dataset_2012_01.basket com a estrutura criada pela função anterior;
 - Guarda na diretoria presente o ficheiro;

c) Orange

Após obtermos o ficheiro .basket, fizemos o upload do mesmo no Orange, o que nos possibilitou de utilizar a sua extensão “associate” para realizar uma análise de associação dos produtos com o objetivo de compreender padrões no comportamento dos utilizadores durante as suas interações na plataforma.

Para identificar diferentes padrões, optamos por utilizar diferentes configurações nesta análise. Logo para uma primeira tentativa, foram utilizados os seguintes parâmetros:

- Suporte de 4%, pois é um valor elevado, que permite obter os produtos mais acedidos pelos visitantes.
- Confiança de 90%, que indica que um utilizador que visite este produto, quase de certeza vá visitar o outro.
- Filtragem na contagem mínima e máxima de antecedentes entre 1 e 999.

Ao aplicar esta configuração encontramos 3 Regras:

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.051	0.915	0.055	2.070	7.984	0.044	pumpseopentoes, botas	→	botins
0.052	0.905	0.058	1.770	8.861	0.046	lon_4004, lon_2125, lon_4508	→	lon_4504
0.052	0.918	0.057	2.110	7.640	0.045	lon_4004, lon_4504, lon_2125	→	lon_4508

Figura 4 - Combinação 1

1. A primeira regra mostra que a visualização do produto "pumpseopentoes" e "bota" está fortemente associada à visualização de "botins" com uma confiança notável de 91.5%. Esta informação sugere uma forte ligação entre estes produtos, indicando que os utilizadores que exploram "pumpseopentoes" e botas também têm uma elevada probabilidade de interessarem-se por botins.
2. A segunda regra aponta para uma associação entre diferentes produtos, especificamente "lon_4004," "lon_2125," e "lon_4508," com uma confiança de 90.5%. Este padrão sugere que os utilizadores que os produtos "lon_4004," "lon_2125," e "lon_4508," tendem a interessar-se pelo produto "lon_4504".
3. A última regra destaca uma diversidade de produtos associados, "lon_4004," "lon_4504," e "lon_2125," com uma confiança de 91.8%. Este padrão sugere que há uma alta probabilidade de os utilizadores que estes três produtos estarem também interessados no produto "lon_4508".

Para a segunda configuração, foram utilizados os seguintes parâmetros:

- Suporte de 1% para abranger mais produtos, no entanto produtos populares na mesma.
- Confiança de 100% para indicar uma relação sem qualquer falha entre produtos.
- Filtragem na contagem mínima de antecedentes entre 9 e 999 para encontrar um grande conjunto de produtos relacionados, para por exemplo coloca-los todos na mesma página no site.

Ao aplicar esta configuração encontramos 28 Regras:

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.010	1.000	0.010	5.385	18.314	0.010	estilo, rc, robots, seguranca, tecnologia, green21, spy, stockoff, display.category*homepage	→ divertimento
0.010	1.000	0.010	4.077	24.189	0.010	estilo, divertimento, rc, seguranca, tecnologia, green21, spy, stockoff, display.category*homepage	→ robots
0.010	1.000	0.010	3.846	25.640	0.010	estilo, divertimento, rc, robots, tecnologia, green21, spy, stockoff, display.category*homepage	→ seguranca
0.010	1.000	0.010	6.615	14.907	0.009	estilo, divertimento, rc, robots, seguranca, green21, spy, stockoff, display.category*homepage	→ tecnologia
0.010	1.000	0.010	3.923	25.137	0.010	divertimento, rc, robots, seguranca, tecnologia, green21, spy, stockoff, display.category*homepage	→ estilo
0.010	1.000	0.010	4.231	23.309	0.010	estilo, divertimento, rc, robots, seguranca, tecnologia, spy, stockoff, display.category*homepage	→ green21
0.012	1.000	0.012	4.667	18.314	0.011	estilo, jogos, rc, robots, seguranca, tecnologia, green21, spy, stockoff	→ divertimento
0.012	1.000	0.012	3.533	24.189	0.011	estilo, divertimento, jogos, rc, seguranca, tecnologia, green21, spy, stockoff	→ robots
0.012	1.000	0.012	3.333	25.640	0.011	estilo, divertimento, jogos, rc, robots, tecnologia, green21, spy, stockoff	→ seguranca
0.012	1.000	0.012	3.400	25.137	0.011	divertimento, jogos, rc, robots, seguranca, tecnologia, green21, spy, stockoff	→ estilo
0.012	1.000	0.012	5.733	14.907	0.011	estilo, divertimento, jogos, rc, robots, seguranca, green21, spy, stockoff	→ tecnologia
0.012	1.000	0.012	3.667	23.309	0.011	estilo, divertimento, jogos, rc, robots, seguranca, tecnologia, spy, stockoff	→ green21
0.010	1.000	0.010	5.385	18.314	0.010	estilo, jogos, rc, robots, seguranca, tecnologia, green21, stockoff, display.category*homepage	→ divertimento
0.010	1.000	0.010	4.077	24.189	0.010	estilo, divertimento, jogos, rc, seguranca, tecnologia, green21, stockoff, display.category*homepage	→ robots
0.010	1.000	0.010	3.846	25.640	0.010	estilo, divertimento, jogos, rc, robots, tecnologia, green21, stockoff, display.category*homepage	→ seguranca
0.010	1.000	0.010	3.923	25.137	0.010	divertimento, jogos, rc, robots, seguranca, tecnologia, green21, stockoff, display.category*homepage	→ estilo
0.010	1.000	0.010	6.615	14.907	0.009	estilo, divertimento, jogos, rc, robots, seguranca, green21, stockoff, display.category*homepage	→ tecnologia
0.010	1.000	0.010	3.385	29.136	0.010	estilo, divertimento, jogos, rc, robots, seguranca, tecnologia, green21, display.category*homepage	→ stockoff
0.012	1.000	0.012	4.667	18.314	0.011	estilo, jogos, robots, seguranca, tecnologia, green21, spy, stockoff, display.category*homepage	→ divertimento
0.012	1.000	0.012	3.533	24.189	0.011	estilo, divertimento, jogos, seguranca, tecnologia, green21, spy, stockoff, display.category*homepage	→ robots
0.012	1.000	0.012	3.333	25.640	0.011	estilo, divertimento, jogos, robots, tecnologia, green21, spy, stockoff, display.category*homepage	→ seguranca
0.012	1.000	0.012	5.733	14.907	0.011	estilo, divertimento, jogos, robots, seguranca, green21, spy, stockoff, display.category*homepage	→ tecnologia
0.012	1.000	0.012	3.400	25.137	0.011	divertimento, jogos, robots, seguranca, tecnologia, green21, spy, stockoff, display.category*homepage	→ estilo
0.012	1.000	0.012	3.667	23.309	0.011	estilo, divertimento, jogos, robots, seguranca, tecnologia, spy, stockoff, display.category*homepage	→ green21
0.012	1.000	0.012	2.933	29.136	0.011	estilo, divertimento, jogos, robots, seguranca, tecnologia, green21, spy, display.category*homepage	→ stockoff

Figura 5 - Combinação 2

Pelo que podemos retirar algumas observações destas mesmas, regras. Como por exemplo que utilizadores que visitam produtos de estilo, divertimento, rc, segurança, tecnologia, green21, spy e stockoff visitaram sempre também produtos “robots”.

Ao verificar as restantes regras, verificamos que os antecedentes acabam por ser sempre os mesmos produtos, mas vão comutando com o consequente, pois estes produtos acabam por estar todos interligados entre si. Assim, podemos concluir com esta configuração que estes produtos são muito procurados em conjunto, o que indica que devem ser apresentados e recomendados aos utilizadores ou na mesma página ou na página de cada um dos produtos.

Conclusão

Para concluir, o trabalho iniciou-se com numa análise abrangente do dataset "We-Commerce." A compreensão inicial da estrutura e dos atributos do dataset foi crucial para a criação de uma base de dados eficaz, na qual armazenámos informações essenciais. Para as quais foram criadas vistas específicas que permitiram a realização de cálculos estatísticos, proporcionando uma visão mais detalhada do comportamento dos utilizadores na plataforma. Essas métricas, foram também importantes para uma análise mais aprofundada.

Focámos também a nossa análise em utilizadores mais ativos, reduzindo o dataset para aqueles com um número significativo de sessões, compreendido entre 5 e 30 sessões. Essa abordagem visa concentrar-nos em utilizadores que nos conseguem fornecer uma perspetiva mais rica e significativa do comportamento dos utilizadores. Assim, o dataset resultante foi exportado para permitir análises posteriores.

A normalização e transformação do dataset para o formato .basket foram passos cruciais antes de aplicarmos técnicas de associação de produtos. Pois este formato simplificado e livre de informações desnecessárias tornou-se mais adequado para a análise de padrões de associação entre os produtos visualizados pelos utilizadores.

Por fim, empregámos a ferramenta Orange, com a extensão "associate", para conduzir análises de associação, explorando diversas configurações de suporte e confiança. As regras de associação identificadas proporcionam percepções valiosas sobre padrões de visualização de produtos, permitindo decisões estratégicas de marketing que visam melhorar a personalização da experiência do utilizador e direcionar campanhas promocionais de maneira mais eficaz.

Em resumo, este trabalho representa um passo significativo para a compreensão e otimização contínua da plataforma "We-Commerce." As conclusões derivadas desta análise não apenas oferecem uma visão aprofundada do comportamento dos utilizadores, mas também fornecem um alicerce sólido para a implementação de estratégias futuras, visando a maximização da satisfação do utilizador e a eficácia das iniciativas de marketing.