UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Solução de Recomendação de Conteúdos Personalizados

Catarina Sofia Esteves Leote

Mestrado em Informática

Versão Pública

Dissertação orientada por:
Prof. Doutor José Romana Baptista Coelho

2019

Agradecimentos

Em primeiro lugar quero agradecer aos meus orientadores, o Professor José Romana Baptista Coelho e o Filipe Rebelo (por parte da Accenture) por toda a orientacão, acompanhamento e disponibilidade ao longo destes meses ajudando-me sempre para que esta tese fosse possível.

Quero agradecer à minha família, pais, avós e irmão por todo o apoio, força e carinho demonstrados ao longo do meu percurso académico, principalmente nos momentos mais dificeis.

Ao meu namorado Diogo por todo o apoio e carinho para que eu não perdesse a força, obrigada por leres a minha tese das 50 mil vezes que te pedi.

Aos meus amigos, por me perdoarem sempre que não fui ter convosco porque tinha de estudar ou escrever a tese, o vosso apoio ao longo de todos estes anos motivou-me a chegar aqui.

Aos meus colegas da Accenture, por todos os dias me lembrarem do grande objectivo e não permitirem que eu desmotivasse até ao fim.

Obrigada a todos os que de qualquer forma contribuiram para o meu percurso académico, sem vocês nada disto seria possível.

Ao meu irmão, por me ensinar a não desistir.

Resumo

Todos os dispositivos mantêm registos das nossas decisões, escolhas e hábitos, aumentando assim a quantidade de dados associados ao indivíduo. Estes dados são um mundo de informação para o comércio e indústria, atribuindo uma lógica a dados que antes não tinham significado.

De acordo com este problema surgiu a temática para este projecto, uma Solução de Recomendação de Conteúdos Personalizados que permitirá, de acordo com os comportamentos do cliente registados, direccionar não só conteúdos disponíveis no serviço em questão, como também campanhas mais adequadas ao perfil comportamental do cliente. Esta Solução terá como base dados reais provenientes de um negócio de telecomunicações e a sua correspondência com os dados do IMDb.

Na solução apresentada, utilizámos dois algoritmos de Aprendizagem Automática com o intuíto de identificar grupos de clientes com os mesmos comportamentos. Estes grupos permitiram estabelecer três cenários distintos de recomendação com base nas *features* escolhidas. Para cada um dos cenários, foram recomendados conteúdos utilizando uma recomendação híbrida, neste caso um conjunto de três tipos diferentes de recomendação.

Com a Solução desenvolvida, conseguimos obter recomendações apropriadas para 90% dos clientes da amostra utilizada. Também foi possível identificar através dos comportamentos destes mesmos clientes, que tipo de campanhas seriam as mais adequadas.

Tudo isto influenciará a experiência do cliente para com o serviço podendo motivar à sua fidelização para com o negócio.

Palavras-chave: Aprendizagem Automática, Prospecção de Dados, Segmentação de Clientes, Sistemas de Recomendação

Abstract

All devices keep a record of our decisions, choices and habits, raising data volume associated with each person. Such data is extremely valuable to business and industries, providing meaning to information that before didn't have any meaning.

On a Solution for Personalized Content Recommendation, that will, from registered customer behavior, not only recommend pertinent visualization options but also provide valuable input for marketing campaigns based on the customer profile. The data used in this solution is originated from the real world deployment of a telecommunications company, and was matched with IMDb data entries.

For such achievement, in the presented solution, we used two Machine Learning algorithms in order to identify customers groups with the same behavior. This groups aim to establish three different recommendation scenarios based on the choosen features. For each one of the scenarios, the content was also recommended using an hybrid recommendation, in this case a set of three different recommendation types.

With the developed solution, we were able to obtain appropriate recommendations for 90% of the clients in the used sample. It was also possible to identify through these customer behaviors, which type of campaigns would be the most appropriate.

All of this will influence the customer experience in the service being able to motivate its loyalty to the business.

Keywords: Customer Segmentation, Data Mining, Machine Learning, Recommendation Systems

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este projecto foi realizado no âmbito da Dissertação do Mestrado em Informática da Faculdade de Ciências da Universidade de Lisboa, com vista à conclusão do Mestrado em Informática.

Este projecto decorreu na empresa *Accenture* num período de 9 meses com o fim de, construir uma Solução de Recomendação de Conteúdos Personalizados que permitirá, de acordo com os comportamentos do cliente registados, direccionar não só conteúdos disponíveis no serviço em questão, como também campanhas mais adequadas ao perfil comportamental do cliente.

Neste capítulo é apresentada a motivação para o desenvolvimento deste projecto, os seus objectivos, contribuições e a estrutura que o documento apresenta.

1.1 Motivação

"Se os dados são o petróleo, a Inteligência Artificial é o motor."

Young Sohn, presidente da Samsung *in* Público [6 de Novembro de 2018]

Encontramo-nos na geração orientada aos dados. Os dispositivos electrónicos que utilizamos nas actividades do dia-a-dia gravam as nossas decisões, escolhas e hábitos aumentando consideravelmente os dados associados ao indivíduo e consequentemente a quantidade de dados no mundo. [1]

Todos estes dados pessoais são um mundo de informação para o comércio e indústria. Padrões de comportamento dos clientes podem ser analisados, atribuindo uma lógica a dados que antes não tinham significado. Sabendo tais características, é mais fácil de identificar os grupos de risco de modo a que o tratamento ao cliente seja diferenciado e personalizado, reduzindo assim a taxa de abandono dos serviços.

Os dados estão presentes em tudo o que nos rodeia, e podem por isso ser de diferentes tipos: geográficos, culturais, científicos, financeiros, estatísticos, meteorológicos, naturais ou de transportes (Figura 1). Estes dados são importantes porque nos permitem:

- Tomar melhores decisões, como encontrar novos clientes, melhorar o serviço ou prever tendências de vendas.
- Resolver problemas, como encontrar as razões da quebra de performance da empresa, identificando que passos que são necessários optimizar.
- Visualizar o desempenho do departamento, empresa ou campanha publicitária.
- Melhorar os processos de negócio diminuindo o desperdício de dinheiro e de tempo.
- Compreender os clientes e o mercado.

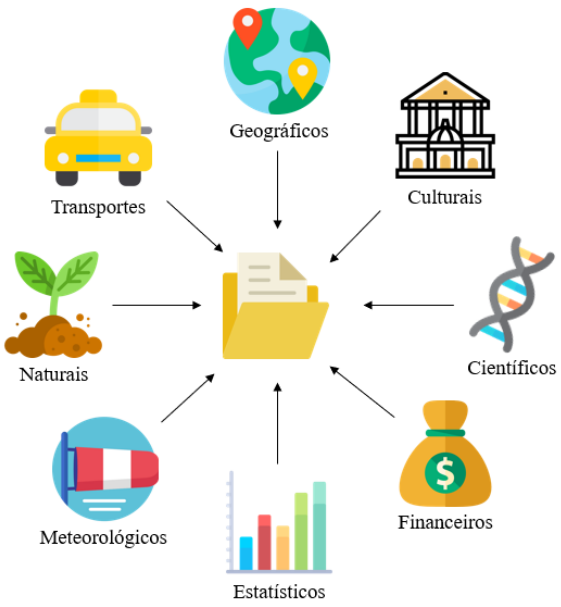

Figura 1: Diferentes tipos de dados.

No entanto, é muito mais rápido produzir estes dados do que extrair informações sobre eles. Como o volume de dados aumenta exponencialmente, a compreensão sobre estes dados diminui levando à falta de utilização de informações com potencial. Por isso, a necessidade de sistemas que permitam a análise e compreensão dos dados é cada vez maior. [1, 2]

Um desses sistemas é a perfilagem de clientes, indicada para perceber o potencial do negócio inexplorado, melhorar o direccionamento comercial e aumentar a taxa de resposta do serviço. Como por exemplo, neste projecto desenvolvido, através da análise comportamental dos clientes, é possível identificar quais as suas preferências. Isto permitirá a recomendação de conteúdos o mais personalizados possível e o direccionamento de campanhas de *upsell* com o objectivo de aumentar a sua satisfação para com o serviço e a empresa.

No ponto de vista do cliente, a existência dos dados e o seu tratamento e análise para a extracção de informações é importante porque permite uma maior personalização do serviço, tornando assim toda a experiência melhor. Com os conhecimentos que o negócio obtém sobre o cliente é possível ajudá-lo, poupando o seu tempo com campanhas não adequadas, simplificar a sua vida fornecendo os serviços que necessita e respeitá-lo, aumentando a sua lealdade para com o serviço que lhe é prestado.

Neste contexto surgiu o tema para este projecto, uma solução de recomendações de conteúdo personalizado que irá segmentar o cliente de acordo com o seu histórico com o objectivo de criar uma experiência o mais personalizada possível. Utilizando algoritmos de Aprendizagem Automática será possível construir o perfil dos clientes de modo a recomendar conteúdos mais direccionados, dirigir campanhas mais adequadas e auxiliar estratégias de *upsell* (quando o vendedor induz à compra de determinados itens, pacotes ou *upgrades*).

1.2 Objectivos

A solução proposta é uma solução de *data science* que permite segmentar o cliente de um serviço de telecomunicações utilizando os seus históricos de visualizações, com vista à criação de uma experiência personalizada.

Esta solução vai permitir a recomendação de conteúdos e campanhas para estes clientes com a utilização de algoritmos de Aprendizagem Automática. Para que tal seja possível, utilizou-se a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), onde estão definidos alguns passos fundamentais para a concretização deste projecto: [3]

• A Análise e compreensão dos dados, onde se identificam e adquirem os dados essenciais para a concretização do objectivo;

• A Construção do modelo de segmentação, onde se utiliza algoritmos de Aprendizagem Automática capazes de prever sobre os dados através dos erros aprendidos;

• E a Elaboração de um manual da solução apresentando relatórios que a justifiquem.

De modo a alcançar o principal propósito foram definidos os seguintes objectivos secundários:

1. Criação de um Perfil Virtual do Cliente e Identificação de Perfis Comportamentais:

De acordo com o seu histórico criar-se-á o Perfil Virtual do Cliente para identificar a que grupo este pertence. Também utilizando as suas visualizações irá ser possível identificar se existe mais do que um perfil comportamental a visualizar o mesmo dispositivo associado ao serviço de acordo com os gostos dos utilizadores.

2. Recomendação de Conteúdos Personalizados:

Agrupando os Clientes de acordo com as suas preferências registadas no serviço irá recomendar-se os conteúdos presentes na plataforma que sejam mais adequados a cada uma das situações.

3. Direccionamento de Campanhas *Upsell*:

Após identificação dos diferentes Perfis Comportamentais presentes no mesmo Número de Conta Cliente é mais fácil direccionar as Campanhas de *Upsell* para promover Paco-

tes de Conteúdos PPV (*Pay per View*), *Upgrades* de serviços, *Boxs*, pacotes de serviço *premium*, entre outros.

Apesar da Solução desenvolvida nesta tese ser um exemplo de uma recomendação estática (através dos históricos existentes obtém-se os conteúdos adequados), no futuro, mesmo não sendo um dos objectivos deste projecto, esta Solução poderá ser implementada num contexto real com base no que aqui foi desenvolvido, tornando-se numa recomendação dinâmica que receberá os dados à medida que os clientes utilizem o serviço.

1.3 Contribuições

A primeira contribuição deste projecto está associada com a origem dos dados utilizados. Até hoje, estes dados representavam um potencial por explorar, e com este projecto, tornou-se possível ter um real entendimento do tipo de informação bem como do tipo de exploração possível de executar por forma a definir padrões de clientes e aumentar o conhecimento sobre estes. Este passo está associado a um conjunto muito grande de vantagens directamente relacionadas com os serviços e os negócios da empresa de telecomunicações em questão e representa uma valiosa contribuição no sentido de criar novas valias de negócio. Indirectamente representa também uma valorização da *Accenture* junto de clientes deste tipo.

Em segundo lugar, o facto deste projecto trabalhar sobre um conjunto de dados reais, permite também uma mais concreta contextualização do tipo de problema em análise. Todos os resultados obtidos através deste projecto representam assim uma aplicação real no contexto empresarial em que se inserem.

Em terceiro lugar, a utilização de ferramentas de *Big Data*, como as englobadas no contexto desta tese, com dados reais é também ainda relativamente incomum num contexto universitário. Esta tese possibilita assim uma análise académica que enriquece substancialmente a temática académica em que se insere. Também pela escolha dessas ferramentas é possível então aplicar a solução criada a qualquer projecto semelhante que possua um conjunto de dados comportamentais dos seus clientes. Esta solução permite adquirir informações essenciais sobre os clientes tal como melhorará a qualidade dos seus serviços com a recomendação de conteúdos adequada.

Em quarto lugar, outro dos factores importantes que representa uma contribuição é a utilização de dois algoritmos distintos de Agrupamento. A solução desenvolvida, por permitir a aplicação de ambos os algoritmos num contexto real, com dados reais, possibilita o aumento do conhecimento sobre estes dois algoritmos, analisando assim o seu comportamento e desempenho.

Em quinto lugar, e em relação à Recomendação obtida, foi possível identificar 3 cenários distintos de acordo com a *feature* escolhida e o número de iterações aplicadas ao

conjunto de dados. O processo que demonstra a identificação destes cenários representa um caso de uso de grande utilidade para a área de negócio em que a tese se insere, bem como o seu produto representa várias oportunidades de adequação de serviços baseados em TV bem com a informação que os compõe.

Em sexto lugar, os perfis comportamentais identificados nesta tese, permitem uma melhoria no direccionamento de campanhas. Assim, estas podem ser adequadas aos clientes de acordo com o seu histórico no serviço.

Em último lugar, foi possível identificar um conjunto de testes e implementações extra a realizar numa perspetiva futura e para a solução construída. Esta identificação representa uma oportunidade cientifica e de exploração de dados virados para conteúdos multimédia que apenas poderá ser satisfeita na continuação do mesmo trabalho, criando por isso mais oportunidades de negócio tanto para a *Accenture* como para o Cliente em causa.

1.4 Estrutura do documento

O presente relatório está dividido em 8 capítulos. Os títulos foram definidos de acordo com a metodologia CRISP-DM utilizada para o projecto e estão organizados da seguinte forma:

Capítulo 1 - Introdução:

Neste capítulo introduz-se a temática descrevendo a motivação para a execução deste projecto e os objectivos propostos.

Capítulo 2 - Conceitos e Trabalho relacionado:

Explicação de conceitos teóricos e trabalhos relacionados com o projecto, nomeadamente sobre Aprendizagem Automática, Prospecção de Dados e Sistemas de Recomendação. Também é exposto o trabalho realizado pela *Accenture* relacionado com esta temática.

Capítulo 3 - Metodologias e Planeamento:

Descrição das metodologias e ferramentas a utilizar. Apresentação do planeamento do projecto.

Capítulo 4 - Conhecimento:

Especificação dos requisitos para o projecto de acordo com os objectivos estabelecidos anteriormente, identificando do problema e definindo o plano de abordagem. Explicação dos procedimentos realizados para a extracção e análise dos dados.

Capítulo 5 - Preparação dos Dados:

Transformação e limpeza dos dados necessária para a resolução dos problemas anteriormente estabelecidos. Preparação do conjunto de dados para o restante procedimento.

Capítulo 6 - Modelação:

Aplicação de técnicas de Prospecção de Dados com o objectivo de identificar padrões antes desconhecidos. Análise dos modelos construídos de acordo com os objectivos inicialmente propostos. Discussão dos resultados obtidos.

Capítulo 7 - Avaliação:

Proposta de um conjunto de testes a realizar capaz de avaliar os resultados obtidos.

Capítulo 8 - Conclusão:

No último capítulo são apresentadas as conclusões extraídas após a realização do projecto, também são apresentados os pontos com potencial para continuação futura de investigação.

Capítulo 2

Conceitos e Trabalho relacionado

2.1 Aprendizagem Automática e Prospecção de Dados

A **Aprendizagem Automática** surgiu em 1950. [4] Envolve vários campos de estudo como a inteligência artificial, prospecção de dados, estatística, entre outros, e equivale à construção de algoritmos que aprendem pela sua experiência de modo a conseguirem prever sobre os dados (Figura 2). Estes modelos são treinados nos dados existentes, que é necessário que sejam muito diversificados e detalhados, até que encontrem os padrões necessários para conseguirem tomar decisões precisas. [5]

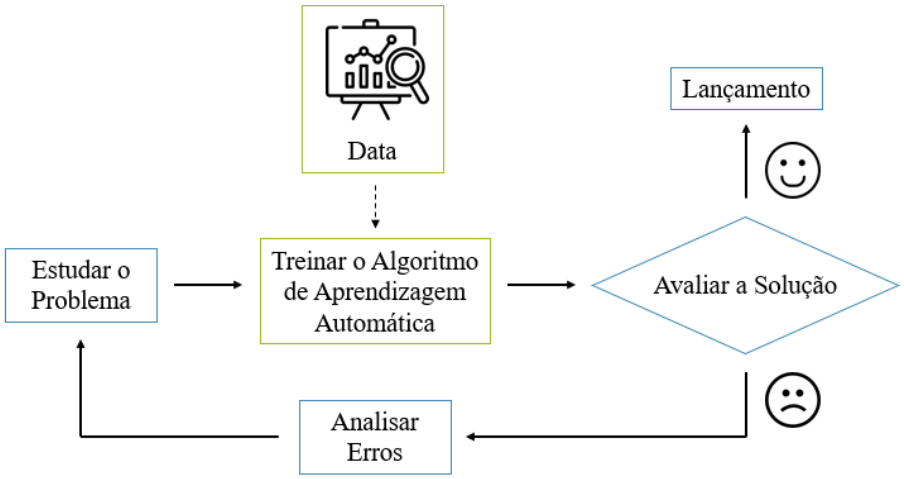

Figura 2: Abordagem da Aprendizagem Automática
Adaptado de *Géron [6]*

Há quatro tipos de Aprendizagem Automática: [6, 7]

• **Aprendizagem Supervisionada:** Quando há *input* (X) e *output* (Y) identificados e o algoritmo aprende a prever o *output* a partir do *input* através da função de mapeamento (f).

$$Y = f(X)$$

Neste tipo de aprendizagem o objectivo é conseguir melhorar a função de mapeamento (f) de tal forma para que, aquando de um novo *input* (X) seja possível prever o *output* (Y) para aqueles dados. Os problemas de aprendizagem supervisionada podem ser agrupados em problemas de regressão e de classificação de acordo com o tipo de variável, isto é, variáveis contínuas serão analisadas por regressão e variáveis descontínuas por classificação.

Alguns exemplos deste tipo de aprendizagem são a Regressão Linear e as Árvores de Decisão.

• **Aprendizagem Não Supervisionada:** Quando os dados não estão identificados e o algoritmo aprende a sua própria estrutura através dos dados de *input* (X).
Os problemas de aprendizagem não supervisionada podem ser agrupados em problemas de agrupamento (*clustering*) e de associação de acordo com o seu objectivo de análise, ou seja, se o objectivo for agrupar o conjunto de dados conforme a sua semelhança utilizam-se algoritmos de agrupamento, por outro lado, se o objectivo for encontrar associações entre os dados do conjunto utilizam-se algoritmos de associação.

Alguns exemplos deste tipo de aprendizagem são o K-Means e as Regras de Associação.

• **Aprendizagem Semi-supervisionada:** Quando os algoritmos conseguem actuar em dados de treino parcialmente identificados. Normalmente a maioria está não identificada e existem apenas alguns identificados. A maioria dos algoritmos semi-supervisionados é uma combinação dos supervisionados com os não supervisionados.

• **Aprendizagem por Reforço:** Quando o sistema de aprendizagem (o agente) pode observar o contexto, seleccionar e agir recebendo recompensas em retorno (ou penalidades se as recompensas forem negativas). Este algoritmo deve de aprender por si qual é a melhor estratégia para receber mais recompensas ao longo do tempo.

A **Prospecção de Dados** surgiu como o processo de aplicar metodologias informáticas para recolher conhecimento dos dados. É um processo iterativo, ou seja, que se repete várias vezes até chegar a um resultado, e na sua maioria é aplicado a análises exploratórias com o objectivo de obter resultados interessantes. É a procura de informações novas, valiosas e não triviais em grandes volumes de dados utilizando técnicas de Aprendizagem Automática, Análise Estatística e Visualização de Dados. [8]

Na prática, os dois principais objectivos da Prospecção de Dados são a previsão e a descrição e por isso é possível dividir a Prospecção de Dados em duas categorias: [9]

• **Prospecção de Dados Predictiva:** Em que a previsão implica a utilização de variáveis do conjunto de dados para prever características desconhecidas ou outras variáveis de interesse, produzindo o modelo do sistema descrito pelo conjunto de dados fornecido.

• **Prospecção de Dados Descritiva:** Que se foca em encontrar padrões nos dados que possam ser interpretados por humanos, produzindo informações novas e não triviais

baseadas na avaliação do conjunto de dados disponível.

A Prospecção de Dados é utilizada para estudar um problema particular com um objectivo de negócio. Assim sendo, utiliza os dados armazenados e tecnologias de manipulação de dados de modo a que os consiga preparar para a análise. Os dados brutos raramente vêm prontos para os algoritmos de aprendizagem e normalmente são necessárias ferramentas de transformação, limpeza e redução de dimensionalidade (resolução de problemas associados ao excesso de dados desorganizados) para conseguir que estes fiquem aptos a treinar o modelo. [5, 7]

A utilização da Prospecção de Dados nos negócios pode ser feita de várias maneiras:

• *Perfilagem do cliente*, identificando os conjuntos de clientes mais vantajosos para o negócio;

• *Direccionamento*, determinando as características dos clientes vantajosos que foram conquistados pelos competidores;

• *Análise baseada no mercado*, estabelecendo as aquisições de produtos por clientes, que podem ser utilizados para disposição e venda cruzada. [10]

É importante para uma empresa prever o conjunto de clientes com maior probabilidade em aceitar certas ofertas e campanhas de acordo com as suas características pessoais ou comportamentais durante o período de fidelização. De tal modo, as empresas começaram a perfilar os seus clientes utilizando técnicas de Aprendizagem Automática que se focam em obter informações até ao momento desconhecidas pela análise de padrões comportamentais dos seus clientes.

Assim, para a construção da Solução iremos aplicar uma perspectiva híbrida de Prospecção de Dados com a utilização das duas vertentes. A Descritiva que nos vai permitir sumariar os acontecimentos passados através da identificação de padrões nos históricos dos clientes. E a Preditiva que nos permitirá descrever o que pode acontecer no futuro prevendo assim os conteúdos que o cliente irá gostar e que lhe devem ser recomendados.

Estes métodos de Prospecção de Dados terão como base os algoritmos de Aprendizagem Automática que levarão à construção de modelos, que após serem treinados, identifiquem as tais informações não triviais presentes nos dados.

2.2 Personalização de Conteúdos

A personalização é a capacidade de fornecer conteúdos e serviços adequados ao indivíduo com base no conhecimento sobre as suas preferências e comportamentos, esta pode ser apropriada para um grupo ou para um indivíduo em específico.

É o modo utilizado para providenciar recomendações pessoais com base no conhecimento de quem o utilizador é, como se comporta e o quão semelhante ele é aos outros utilizadores, permitindo extrair conhecimento dos dados disponíveis no perfil do utilizador. Os perfis dos clientes são a base para os diferentes tipos de personalização fazendo uso de técnicas como a Prospecção de Dados e Interfaces Personalizadas.

Os dados utilizados para a identificação das regras são de dois tipos, reais (quem o utilizador é) e transaccionais (o que é que o utilizador faz). Por consequência, o perfil do utilizador tem duas partes, uma real e uma comportamental. O perfil real tem informações como o nome, género, data de nascimento, entre outros e o perfil comportamental modela as acções do utilizador e normalmente é derivado dos dados transaccionais.

A personalização de conteudos permite-nos então encontrar o item mais adequado para o cliente em questão mesmo em grandes negócios, fazendo-nos lembrar do dono da pequena mercearia que sabe sempre o que procuramos. Este aspecto melhora a nossa relação com o serviço pois faz com que este saiba sempre do que precisamos até mesmo antes de nós o realmente precisarmos. [11, 12]

Na perspectiva deste projecto, a personalização de conteúdos é a chave fundamental para obter uma boa recomendação através da solução construída, dado esse facto, serão criados os perfis virtuais dos clientes de acordo com as informações reais disponíveis no conjunto de dados (o número da conta de cliente). A partir das informações transaccionais (históricos de visualizações) vai então ser possível identificar o perfil comportamental de cada cliente.

A identificação destes perfis comportamentais é essencial para tornar a recomendação obtida o mais personalizada possível.

2.3 Sistemas de Recomendação

Dada a enorme oferta de conteúdos digitais e, numa perspectiva de progresso e melhoria de acções de marketing, os Sistemas de Recomendação tornaram-se uma área importante de pesquisa. [13]

Os Sistemas de Recomendação combinam várias técnicas computacionais, para seleccionar itens personalizados de acordo com os interesses de outros utilizadores ou do contexto onde estão inseridos. Estes são normalmente classificados em 3 categorias, segundo a forma como a recomendação é realizada: [13]

2.3.1 Recomendações com base no conteúdo

A utilidade $u(c,s)$ de um item s para um utilizador c é estimada com base nas utilidades $u(c,s_i)$ atribuídas pelo utilizador c aos itens $s_i \in S$, semelhantes ao item s. Neste caso, para

recomendar filmes ao utilizador, o sistema de recomendação com base no conteúdo percebe as semelhanças entre os filmes que o utilizador visualizou ou deu boa classificação anteriormente (de acordo com os géneros, actores, directores, ou outras características do conjunto de dados), e então, recomenda os que têm elevado grau de semelhança de acordo com as suas preferências.

Neste caso, é importante a construção do perfil do utilizador, de modo a possuir informações como os seus gostos, preferências e necessidades. Esta informação pode ser adquirida explicitamente, por exemplo através de questionários, ou implicitamente através dos comportamentos transaccionais ao longo do tempo.

Esta categoria apresenta algumas dificuldades:

• Análise de conteúdo limitado: Onde a recomendação é limitada pelas características explicitamente associadas aos itens recomendados pelo sistema. Isto é, se existirem dois itens diferentes representados pelas mesmas características, estes tornam-se indistinguíveis.

• Super-especialização: Quando o sistema apenas consegue recomendar itens que foram muito vistos ou têm uma alta classificação de acordo com o seu perfil. Nestes casos o utilizador fica limitado a itens semelhantes aos que já demonstrou interesse e assim não são sugeridos novos itens fora do esperado, de acordo com o perfil do utilizador.

• Problema do novo utilizador: O utilizador tem de visualizar ou classificar um número suficiente de itens para que o sistema de recomendação consiga entender as suas preferências e sugira recomendações adequadas. Neste caso, um novo utilizador, visto que tem poucas visualizações ou classificações, não conseguirá ter boas recomendações. [13]

2.3.2 Recomendações colaborativas

Os sistemas de recomendação colaborativa tentam prever a utilidade dos itens para um determinado utilizador de acordo com os itens previamente classificados por outros utilizadores, isto é, a utilidade $u(c,s)$ de um item s para o utilizador c é estimada de acordo com as utilidades $u(c_j,s)$ atribuídas para o item pelos utilizadores $c_j \in C$, semelhantes ao utilizador c. Por exemplo, para recomendar filmes a um utilizador, o sistema de recomendação colaborativa procura outros utilizadores que tenham interesse nos mesmos filmes, e então, apenas estes filmes vão ser recomendados.

Há duas classes de algoritmos para recomendações colaborativas:

• **Algoritmos com base na memória**, que fazem as previsões de acordo com a colecção de itens previamente visualizados ou classificados pelos utilizadores, ou seja, o valor desconhecido $r_{c,s}$ para o utilizador c e o item s é normalmente computado como o total de outros utilizadores (normalmente, os N mais semelhantes) para o mesmo item s.

• **Algoritmos com base no modelo**, que utilizam a colecção das visualizações ou classificações para aprender o modelo, para que seja utilizado para fazer previsões.

Esta categoria apresenta algumas dificuldades:

• Problema do novo utilizador: Sendo exactamente a mesma dificuldade que nos sistemas de recomendação com base no conteúdo. O sistema tem de aprender primeiro as preferências do utilizador antes de conseguir recomendar algum item.

• Problema do novo item: Itens novos são adicionados regularmente ao sistema mas visto que ainda não possuem nenhuma classificação, não são recomendados.

• Dispersão: Este sistema de recomendação necessita de um determinado número de utilizadores que visualize ou classifique os diferentes itens. Se houverem poucos utilizadores as previsões realizadas não corresponderão a uma boa recomendação. [13]

2.3.3 Abordagens híbridas

É a combinação de métodos de recomendação com base no conteúdo e métodos de recomendação colaborativa. Esta combinação pode ser classificada de diferentes maneiras:

• **Implementando separadamente métodos colaborativos e com base no conteúdo, combinando as suas previsões:** Havendo dois cenários, pode-se combinar os interesses obtidos individualmente numa recomendação final. Por outro lado, pode-se utilizar sempre os sistemas de recomendação individualmente escolhendo a qualquer altura o melhor, de acordo com a qualidade da recomendação.

• **Incorporando algumas características com base no conteúdo numa abordagem colaborativa:** Permite ultrapassar alguns problemas de dispersão porque a semelhança é sempre calculada entre dois utilizadores. Por outro lado, torna possível recomendar um item mesmo que este não esteja altamente classificado ou visualizado por utilizadores com perfis semelhantes.

• **Incorporando algumas características colaborativas numa abordagem baseada no conteúdo:** Usam-se técnicas de redução de dimensionalidade num grupo de perfis com base no conteúdo de modo a diminuir a complexidade do processamento dos algoritmos de Aprendizagem Automática provocada pelo aumento do número de atributos.

• **Construindo um modelo geral único que incorpora tanto características com base no conteúdo como colaborativas:** Utilizando técnicas com base no conhecimento aumenta-se a precisão das recomendações para dar resposta a algumas das limitações. [13]

No contexto deste projecto, será utilizada a abordagem Híbrida de Recomendação por se tratar do tipo de recomendação mais vantajoso para o cliente. Este tipo de recomendação combina as vantagens dos dois tipos existentes e evita os seus problemas. Assim, será construído um Sistema de Recomendação tendo em conta as características tanto da Recomendação com Base no Conteúdo como da Recomendação Colaborativa de modo

a que esta seja o mais completa possível.

2.4 Accenture - Princípios Relacionados

Como o projecto foi desenvolvido num contexto empresarial, integrado num projecto real da *Accenture*, este foi desenvolvido segundo os valores praticados neste tipo de análise. A *Accenture* tem uma parceria com o MIT em *Business Analytics*, onde é aplicada uma investigação colaborativa focada no desenvolvimento de novas soluções analíticas para o negócio, auxiliando as empresas dos dias de hoje na resolução de alguns dos seus desafios mais críticos. [14]

Com este trabalho relacionado por parte da *Accenture*, foi possível enquadrar os conhecimentos académicos com a perspectiva de negócio também importante na Solução desenvolvida.

2.4.1 *Accenture Customer Insight*

A abordagem da Accenture ao cliente, *Accenture Customer Insight* (ACI), permite uma melhor compreensão dos dados, tecnologias e analíticas avançadas resultando numa evolução das funções para os clientes devido ao conhecimento obtido sobre eles.

Figura 3: Analítica do cliente de acordo com a abordagem *Accenture Customer Insight*.
Adaptado de *Accenture Costumer Insight [15]*

A Accenture é lider na estrutura analítica do cliente, e a abordagem ACI inclui muitas das ferramentas e capacidades que têm sido utilizadas para auxiliar organizações em vários sectores. Os serviços de analítica, Figura 3, incluem ferramentas e métodos para

perceber tendências e conhecimentos dos dados de modo a suportar melhores tomadas de decisão. Estes serviços podem ser realizados como estratégias a curto prazo ou serviços de gestão a longo prazo **detectando** as oportunidades nos dados, **modelando** analíticas preditivas e testando e refinando o modelo construído para a **execução**.

Integrando analíticas multifucionais, as equipas adquirem uma visibilidade ampla e consistente da performance da empresa e dos seus clientes, e podem trabalhar de uma maneira mais coordenada para criar experiências excepcionais e contínuas que optimizem variáveis como o portefólio de produtos, preços e promoções. [15]

2.4.2 *Accenture Video Analytics*

Neste mundo de dados, as analíticas são fundamentais, o aumento de novos conteúdos leva ao aumento da personalização destes e da sua análise. Em relação aos sistemas de recomendação, estes têm o necessário para prever as necessidades dos consumidores de acordo com os históricos de compras anteriores, comportamento online, classificações, *reviews* e outros atributos personalizados. A Analítica de Video da Accenture, *Accenture Video Analytics* (AVA), é uma plataforma que converte os dados em percepções que ajudam a entender o que os espectadores querem, permitindo assim corresponder rapidamente às suas necessidades com uma tomada de decisões estratégica.

A plataforma AVA permite uma estratégia de conteúdos, publicidade digital e operações para melhorar a experiência do utilizador, desenvolver um compromisso de consumo, aumentar a lealdade, descobrir novas fontes de rentabilização e conhecer os desafios da era digital. Suporta algoritmos de Aprendizagem Automática que são aplicados em diferentes casos, como por exemplo, para as visualizações, retenções, experiências de consumo, qualidade de experiência, anúncios, tarifários, direccionamento de publicidade e optimização de conteúdos. [16]

Capítulo 3

Metodologias e Planeamento

Este projecto foi desenvolvido na *Accenture*, no âmbito de um projecto realizado para uma empresa de telecomunicações com uma vasta quantidade de dados. Por esse motivo, as ferramentas utilizadas neste projecto são metodologias para *Big Data* usualmente implementadas pela empresa nas diversas plataformas.

3.1 CRISP-DM

A metodologia CRISP-DM, *Cross Industry Standard Process for Data Mining*, utilizada neste projecto é uma ferramenta essencial para estruturar um projecto de Prospecção de Dados. Esta tecnologia surgiu em 1996, numa época em que o interesse pela Prospecção de Dados começou a surgir. Contém as fases de um projecto, as suas respectivas tarefas e as relações entre estas. O ciclo de vida de um projecto, de acordo com esta metodologia, consiste em 6 fases (Figura 4): [3]

1. Conhecimento do Negócio

Esta fase inicial foca-se em perceber os objectivos do projecto e os seus requerimentos na perspectiva do negócio, para então converter o conhecimento num problema de Prospecção de Dados e definir um plano que consiga atingir os objectivos propostos.

2. Conhecimento dos Dados

Esta fase começa com uma colecção de dados inicial que através de operações informáticas permitem a familiarização com os dados, identificação de problemas de qualidade e descoberta de subconjuntos que auxiliam na formulação de hipóteses do negócio.

3. Preparação dos Dados

A fase de preparação dos dados abrange todas as operações necessárias para construir o *dataset* final. Entre estas operações encontra-se a transformação e limpeza dos dados para as ferramentas de modelação.

4. Modelação

Nesta fase, várias técnicas de modelação são seleccionadas e aplicadas ao *dataset* escolhido, os seus parâmetros são calibrados para valores óptimos. Normalmente existem

diversas técnicas para o mesmo tipo de problema de Prospecção de Dados, muitas vezes é necessário voltar à fase da Preparação dos Dados (Fase 3) devido aos requerimentos específicos de certas técnicas.

5. Avaliação

Nesta etapa do projecto, o modelo(ou modelos) já foi construído e apresenta uma elevada qualidade na perspectiva da análise de dados. Antes da implementação do modelo, é importante avaliá-lo e rever as fases para a sua criação de maneira a assegurar que o modelo concretiza os objectivos propostos inicialmente.

6. Implementação

A criação do modelo normalmente não é o fim do projecto. Mesmo que o propósito do modelo construído seja aumentar o conhecimento sobre os dados, esse conhecimento precisa de ser organizado e apresentado ao cliente de modo a que este o possa utilizar.

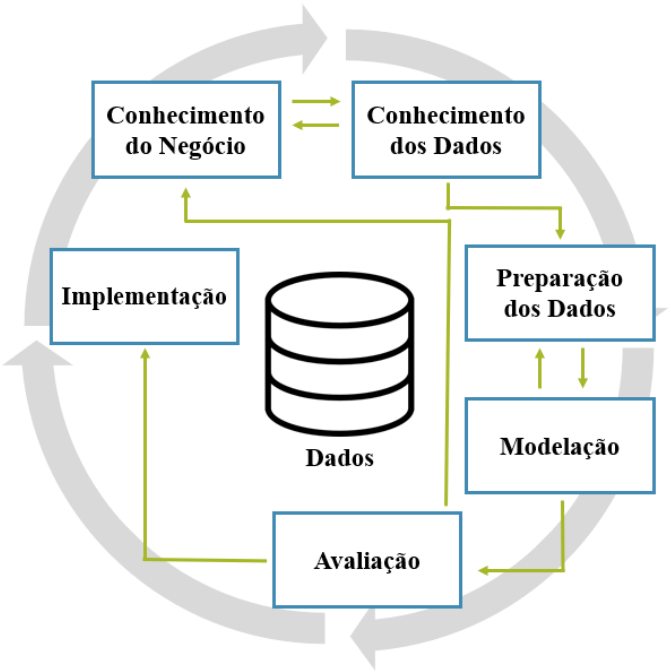

Figura 4: Fases do modelo de referência CRISP-DM.
Adaptado de *Chapman et al. [3]*

Esta metodologia tem como finalidade fazer grandes projectos de Prospecção de Dados, com menores custos, mais fiáveis, reproduzíveis, rápidos e viáveis. O modelo vai servir como um ponto de referência aumentando o conhecimento sobre dados cruciais. [17]

3.2 Ferramentas Informáticas

De acordo com o Relatório de 2017 de *Data Scientists* [18], mais de 50% do tempo destes é gasto na recolha, identificação, limpeza e organização dos dados. Por isso, a escolha das ferramentas correctas é extremamente importante para garantir a eficiência do processo.

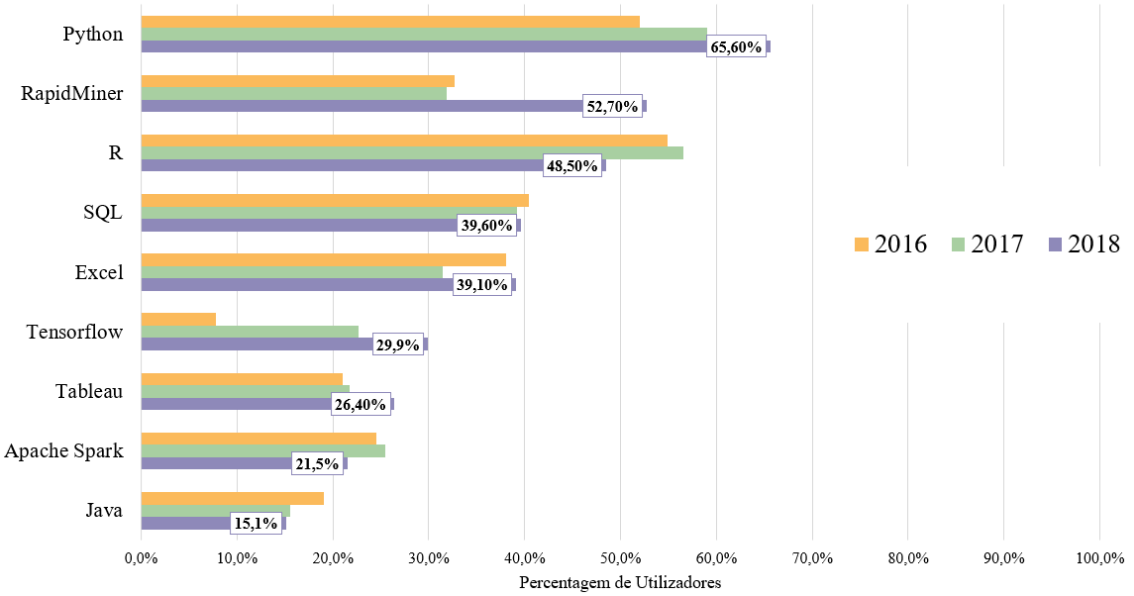

Figura 5: Sondagem sobre ferramentas de Aprendizagem Automática.
Adaptado de *KDnuggets [19]*

Analisando sondagens dos últimos 3 anos sobre as ferramentas mais utilizadas para Aprendizagem Automática, presentes na Figura 5, pode-se concluir que o Python, RapidMiner, e R são as 3 principais. No entanto nenhuma destas ferramentas é adequada para *Big Data*, para tal a ferramenta mais utilizada é o Apache Spark, que ocupa o 8º lugar na sondagem apresentada. Esta é a ferramenta utilizada durante este projecto.

3.2.1 Apache Spark

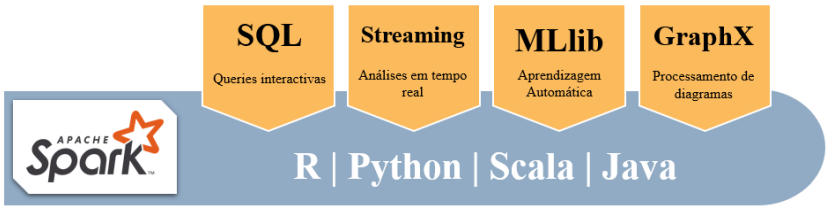

Figura 6: Núcleo do Apache Spark.
Adaptado de *Hien Luu [20]*

O Apache Spark (Figura 6) é uma ferramenta livre de computação em *cluster* para processamento em tempo real, isto é, o processo ocorre em computadores independentes que

estão combinados num único sistema. Esta ferramenta tem como objectivo o processamento de grandes conjuntos de dados de forma paralela e distribuída.

É uma ferramenta poliglota (podendo ser utilizada em R, Python, Scala e Java), rápida, admite múltiplos formatos de dados, com processamento em tempo real, integrada no *Hadoop* (outra ferramenta de processamento distribuído) e permite aplicar Aprendizagem Automática através do seu componente MLlib (*Machine Learning Library*).

Combinando o Apache Spark como uma framework de MPP (*Massively Parallel Processing*) e a linguagem Python, foi escolhida a API PySpark que permite a utilização das duas frameworks conciliando a simplicidade da linguagem com o poder de processamento da framework. Outra das ferramentas utilizada neste projecto foi a biblioteca MLlib, um dos núcleos de Spark dedicado à análise de dados utilizando algoritmos de Aprendizagem Automática. Esta biblioteca apresenta muitos dos algoritmos utilizados para classificação, regressão, agrupamento e sistemas de recomendação. [21]

3.3 Planeamento

O plano adoptado no projecto, com as tarefas e respectivas durações ao longo dos 9 meses de trabalho, está descrito na Figura 7.

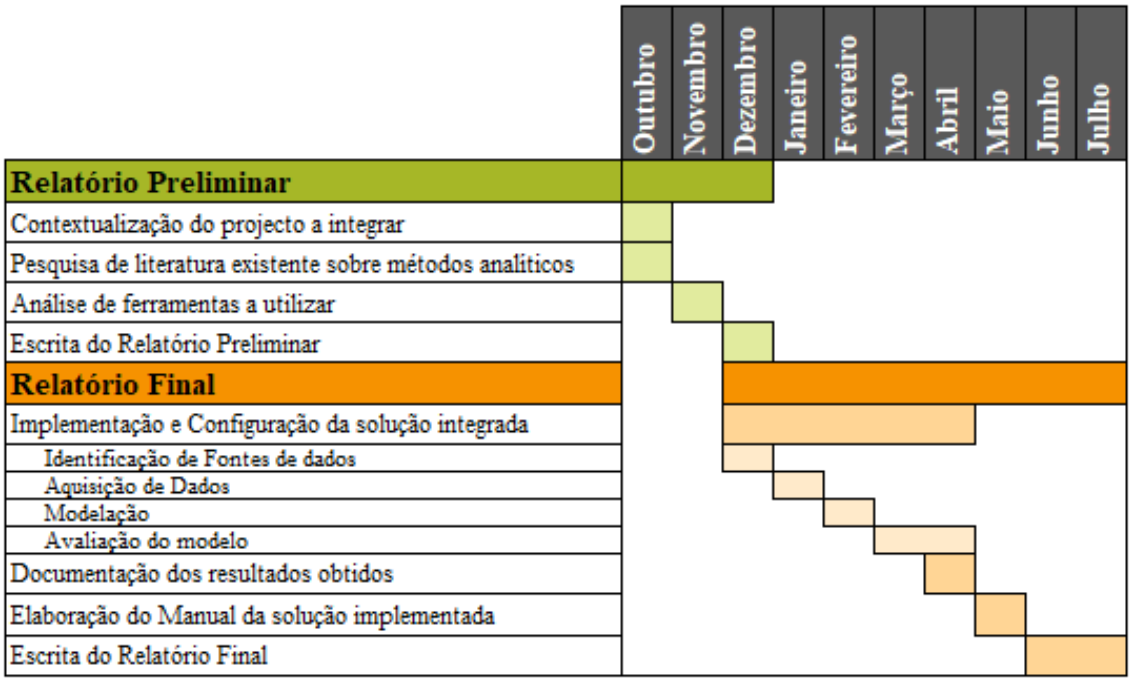

Figura 7: Planeamento inicial do projecto.

Este plano engloba as diferentes fases necessárias para a construção da Solução de Recomendação de Conteúdos Personalizados de acordo com a metodologia adoptada e tendo em conta as duas datas principais que correspondem às duas fases de entrega do Relatório.

Numa primeira fase foi necessário identificar os objectivos e planos para o decorrer do projecto e foi importante contextualizar a solução a desenvolver de acordo com os objectivos da *Accenture*. Para tal, foi necessário pesquisar literatura relevante para a construção do estado de arte e analisar as ferramentas disponíveis por parte do cliente, tendo em conta as necessidades dos requisitos levantados.

Numa segunda fase, com vista ao desenvolvimento da Solução, foram identificadas as fontes de dados e a sua respectiva acquisição. Estes mesmos dados foram então transformados para que fosse possível aplicar modelos de Aprendizagem Automática.

De acordo com o plano inicialmente traçado, Figura 7, existiram algumas alterações no decorrer do projecto. A fase correspondente à Implementação e Configuração da solução integrada envolveu alguns atrasos devido a restruturações da proposta inicial por parte da empresa. De tal forma, a fase da Avaliação do Modelo, inicialmente planeada, não foi realizada sendo apresentada no entanto uma proposta de testes a realizar no futuro da solução construída.

Capítulo 4

Conhecimento

4.1 Conhecimento do Negócio

O projecto foi desenvolvido com os dados disponibilizados pelos serviços de uma empresa de telecomunicações, conteúdos, média, entretenimento e publicidade. [redacted]

O objectivo do projecto desenvolvido centrou-se na segmentação do cliente de um serviço em específico com o fim de melhorar a sua experiência, tornando-a o mais personalizada possível através da recomendação mais apropriada. Por segmentação entende-se a divisão dos clientes em grupos distintos de acordo com as suas características. Este processo actua sobre o facto de que todo o cliente é diferente e de que acções de *marketing* ou ofertas de produtos personalizadas levam a um aumento das taxas de satisfação, conduzindo assim a um crescimento da fidelização.

Considerando-se a realização de tal fim, propuseram-se outros três objectivos secundários que visam à criação de um perfil virtual do cliente permitindo a identificação dos perfis comportamentais presentes no mesmo através das informações recolhidas dos históricos das suas visualizações e *downloads*.

A importância da identificação dos perfis comportamentais está relacionada com a utilização da mesma conta pelos diferentes constituintes do agregado. [redacted] Uma experiência pouco personalizada conduzirá a um menor índice de satisfação em relação ao produto, que por sua vez influenciará as taxas de abandono do serviço.

Comparativamente ao que existe hoje em dia, esta solução é bastante inovadora pois para além da identificação dos perfis comportamentais tem em consideração dados ainda não explorados, relativos à utilização de um serviço que inclui todas as visualizações e

downloads realizados via *Internet* dos conteúdos programáticos existentes na *Set Top Box*, dispositivo utilizado para visualizar o conteúdo televisivo.

A partir da identificação de padrões nestes dados, esta solução contribuirá para o aumento do número de atributos recolhidos sobre o cliente, especialmente os que são observados através do seu comportamento enriquecendo o domínio de informação sobre este. Estes atributos irão complementar outras soluções já utilizadas no negócio, auxiliando-o nas suas estratégias de analítica avançada permitindo adquirir informações sobre o comportamento futuro dos clientes. Outro aspecto fundamental na solução desenvolvida é que esta está integrada na plataforma *Big Data*, ou seja, não necessitará de reengenharia pois o seu desenvolvimento já foi efectuado tendo em conta as *frameworks* disponibilizadas pela plataforma do cliente (neste caso, PySpark versão 2.2.0 com a versão 2.7.16 de Python). Assim, dado que as fontes de dados utilizadas são as já existentes na plataforma, será facilitada a integração e produção da solução desenvolvida.

Como tal foram identificados três desafios para este projecto, a construção do perfil virtual do cliente a partir dos seus históricos, a recomendação de conteúdos personalizados para cada cliente e a identificação da constituição dos diferentes perfis de utilização associados ao mesmo número de conta.

4.2 Conhecimento dos Dados

Para o desenvolvimento deste projecto foram utilizadas diversas fontes de dados. [redacted].

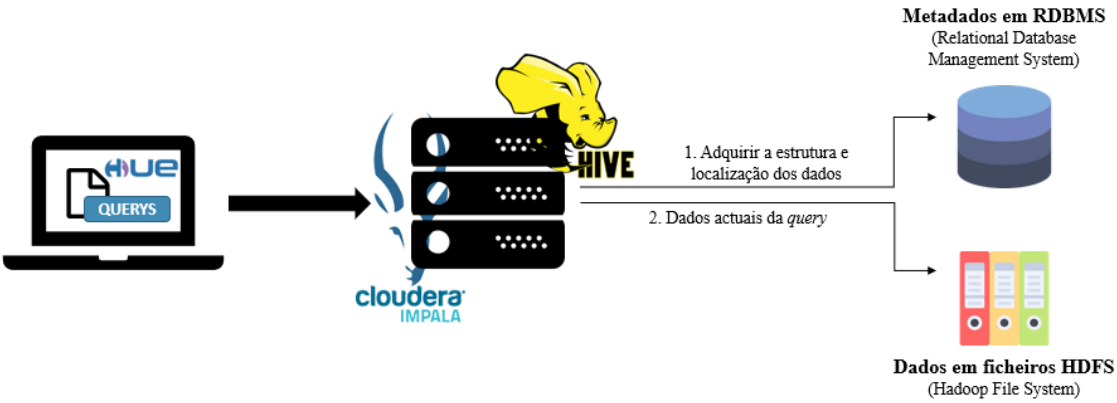

Figura 8: Processo decorrido para a consulta de dados utilizando o HDFS.

Sabendo que a Solução produzida está integrada na plataforma de *Big Data*, os dados foram exportados das fontes de dados para o HDFS (*Hadoop File System*), sistema

utilizado pela empresa. Este sistema de armazenamento de ficheiros, tal como descrito na Figura 8, foi acedido através da interface *Hue* (*Hadoop User Environment*) permitindo assim utilizar tanto o *Impala* como o *Hive* para consultar os dados armazenados. A exportação dos dados necessários realizou-se no dia 12 de Março de 2019.

Outra das fontes de dados utilizada neste projecto foram os dados provenientes do IMDb (*Internet Movie Database*) disponíveis *online* (*datasets.imdbws.com*) e extraídos no dia 15 de Novembro de 2018. Estes dados irão ser essenciais para o sistema de recomendação permitindo recolher maior número de informações sobre os títulos disponíveis no catálogo VOD.

4.2.1 Dados do Negócio

A fonte de dados principal, correspondente aos dados fornecidos pela empresa de telecomunicações, foi adquirida através dos históricos dos eventos de visualizações e de *downloads* dos clientes para um serviço específico como indicado anteriormente. Para tal foi necessário utilizar três tabelas distintas, os Eventos de Visualização, os Eventos de *Download* e o Catálogo de *Video on Demand* (Tabela 1).

Tabela 1: Descrição das variáveis dos dados disponibilizados pelo negócio utilizados neste projecto.

Origem	**Variável**	**Tipo**	**Exemplo**
Eventos de Visualização	Data do evento	datetime	$2014-09-19\ 08:35:03.197$
	Número de conta cliente	nvarchar	123456789
	Código do produto	nvarchar	AA-1234567
	Data actualização	datetime	$2014-09-26\ 08:34:37.447$
Eventos de *Download*	Data do evento	datetime	$2014-09-19\ 08:35:03.197$
	Número de conta cliente	nvarchar	123456789
	Código do produto	nvarchar	AA-1234567
	Data actualização	datetime	$2014-09-26\ 08:34:37.447$
Catálogo de *Video on Demand*	Código do produto	varchar	AA-1234567
	Descrição da oferta	varchar	O Sr. Perfeito
	Data do início da licença	datetime	$2018-01-08\ 00:00:00$
	Data do fim da licença	datetime	$2024-12-30\ 23:59:59$
	Data do início da oferta	datetime	$2018-01-08\ 00:00:00$
	Data do fim da oferta	datetime	$2024-12-28\ 23:59:59$
	Descrição das categorias	varchar	Géneros
	Código do ano	int	2015
	Código do estado	varchar	Active
	Descrição do género	varchar	Acção
	Código da classificação	varchar	-1
	Preço da oferta	float	$2,440000057220459$
	Data actualização	datetime	$2019-03-12\ 08:57:26.46000$

Os Eventos de Visualização têm aproximadamente [redacted] de registos, cada um

destes registos corresponde à visualização de um conteúdo disponível no serviço, por exemplo, no dia 19 de Setembro de 2014, o cliente 123456789 viu no serviço em questão, o conteúdo AA-1234567.

Desta forma, os Eventos de *Download* correspondem ao *download* físico para um dispositivo (*tablet*, *smartphone*, entre outros) para poder aceder a este mesmo conteúdo *offline*, como por exemplo, no dia 19 de Setembro de 2014, o cliente 123456789 descarregou para o seu dispositivo o conteúdo AA-1234567. Estes eventos contêm sensivelmente [redacted] registos.

No total, estes eventos em conjunto correspondem a [redacted] anos e meio de históricos compreendendo os registos entre [redacted]

O Catálogo de *Video on Demand* apresenta [redacted] entradas, estas correspondem a todo o tipo de conteúdo disponível na plataforma, compreendendo desde filmes do clube de vídeo, séries e até mesmo algum do conteúdo televisivo disponivel na *Set Top Box*.

4.2.2 Dados IMDb

Os dados do *IMDb* estão organizados em seis tabelas distintas, das quais apenas três foram utilizadas neste projecto (Anexo A). Estas são relativas às informações básicas sobre os títulos, às classificações que os títulos têm de acordo com a votação dos utilizadores da plataforma, e a uma tabela que corresponde a todas as traduções e adaptações dos nomes dos títulos (Tabela 2).

Tabela 2: Descrição das variáveis dos dados do IMDb utilizados neste projecto.

Origem	**Variável**	**Tipo**	**Exemplo**
Informações básicas	Identificador do título	string	$tt0000001$
	Tipo do título	string	short
	Título principal	string	Carmencita
	Título original	string	Carmencita
	Ano de início	int	1894
	Ano final	int	1894
	Duração (em minutos)	int	1
	Géneros	array string	Documentary, Short
Adaptações	Identificador do título	string	$tt0000001$
	Título	string	carmencita - spanyol tánc
	Região	string	HU
Classificações	Identificador do título	string	$tt0000001$
	Classificação média	float	5.8
	Número de votos	int	1436

Estes dados contêm informações sobre mais de 5 milhões de títulos, incluindo filmes, séries ou até mesmo jogos. Estas informações são importantes pois aumentam o número

de características a utilizar na identificação dos perfis comportamentais dos clientes aumentando o universo de informações disponíveis.

Capítulo 5

Preparação dos Dados

Com o objectivo de formar o conjunto de dados final que será a base da Solução construída, é necessário preparar os dados com o propósito de uniformizar a amostra recolhida de modo a que esta nos permita retirar informações essenciais ainda não conhecidas.

Através da análise dos dados é possível concluirmos em relação à sua qualidade e que desempenho poderemos esperar na Solução.

5.1 Catálogo de *Video on Demand*

[redacted]

Seleccionaram-se apenas os produtos disponíveis e mais actualizados escolhendo as entradas cuja data de actualização era a mais recente, [redacted]

Um dos objectivos da Solução é a identificação dos perfis comportamentais e portanto foi necessário normalizar a amostra para as características a utilizar nessa fase, nomeadamente a Descrição das categorias e a Descrição do género.

[redacted]

Tabela 3: Número de Registos ao longo das fases de transformação dos dados do Catálogo de *Video on Demand*.

[redacted]

Em relação à Descrição do género [redacted] O Catálogo apresenta [redacted] géneros distintos dos quais se pode observar, na Figura 9, que o predominante é o de [redacted] seguindo-se o de [redacted] e o de [redacted]. [redacted]

Figura 9: Top 5 de Géneros de acordo com o número de títulos disponíveis no Catálogo VOD.

[redacted]

5.2 Eventos de Visualização e de Download

Quando extraídos da base de dados da empresa em questão, os Eventos de Visualização tinham [redacted] de registos e os Eventos de *Download* aproximadamente [redacted]. [redacted]

Destes registos por fim foram removidos os que não apresentavam correspondência directa com os conteúdos disponíveis no Catálogo de *Video on Demand*, pois não seriam importantes para a Solução construída visto que não seria possível retirar qualquer tipo de informação em relação aos seus históricos de visualizações e *downloads* [d] (Tabela 4).

Tabela 4: Número de Registos ao longo das fases de transformação dos dados dos Eventos de Visualização e de *Download*.

[redacted]

Para o resto do projecto considerámos então [redacted] registos de visualizações e [redacted] de *downloads*, num total de [redacted] registos de eventos que foram combinados num mesmo dataset adicionando uma coluna com a identificação do tipo de evento (podendo ser de visualização ou de *download*). Estes [redacted] eventos foram os históricos considerados para a construção da solução apresentada nesta tese.

Em relação à distribuição do número de eventos (visualização e *download*) ao longo do período da amostra (Figura 10), entre [redacted], observa-se que o número de visualizações aumentou a partir de [redacted]. Mais tarde observou-se um pico em [redacted] que corresponde a uma actualização do serviço em análise. Nos últimos 2 anos o número médio de visualizações é de aproximadamente [redacted] por mês não apresentando grandes variações.

[redacted]

Figura 10: Número de eventos de Visualização e *Download* por mês.

Os dados utilizados correspondem a [redacted] clientes, sendo que os clientes apresentam desde [redacted] a [redacted] registos, mas apenas cerca de [redacted] clientes têm mais de [redacted] eventos registados no serviço. Analisando os históricos dos clientes sabe-se que apenas [redacted] títulos dos [redacted] existentes no Catálogo de *Video on Demand* foram requisitados (quer para visualização como para *download*).

5.3 Correspondência com a *Internet Movie Database*

O IMDb armazena aproximadamente 5.3 milhões de títulos (incluindo programas de televisão, filmes, jogos, documentários, entre outros). Esta base de dados é essencial para o sistema de recomendação por permitir adicionar novas características ainda desconhecidas e fundamentais. Para que a correspondência entre os dados provenientes do negócio e as características provenientes do IMDb fosse possível alguns passos foram necessários:

· Os títulos das ofertas na base de dados da empresa possuíam informações adicionais como "(em h.d.)", "(v.o.)", entre outros, que fariam com que o título não tivesse correspondência directa. Também foram encontrados caracteres especiais que não existiam na base de dados do IMDb e, por esse motivo, estas informações foram retiradas e os caracteres substituídos.

· De seguida, dispondo dos títulos das ofertas transformados, foi possível procurar qual das entradas do IMDb tinha correspondência directa tanto para os títulos presentes nas Informações básicas como para as traduções existentes na tabela das Adaptações. Neste passo foram também consideradas as datas das ofertas, pois o seu ano (informação registada nos dados da empresa como Código do ano) tinha de se encontrar entre o ano de início e o ano final do título do IMDb para que a relação fosse a correcta.

· Por fim, algumas das ofertas obtinham mais do que uma correspondência com o IMDb, algumas por se tratarem de episódios de séries outras por existirem diversas versões do mesmo filme. Para seleccionar apenas uma correspondência para cada oferta escolheu-se a oferta cujo Identificador do título (id) do IMDb fosse o menor pois após analisar diversas entradas do IMDb, verificou-se que os id's eram atribuídos de acordo com a sua importância, ou seja, o menor id era o título mais relevante.

Através deste método foi possível obter a correspondência para [redacted] ofertas das [redacted] entradas do Catálogo de *Video on Demand* da empresa, ou seja, apenas 30% dos conteúdos disponibilizados. Para todos os conteúdos que não foram possíveis de cruzar com a base de dados do IMDb foi-lhes atribuido o género *Outro* possibilitando assim a identificação dos clientes que visualizam conteúdos sem correspondência, este facto irá influenciar as conclusões obtidas com a análise de dados.

Capítulo 6

Modelação

A Modelação, de acordo com a metodologia utilizada neste projecto é a fase onde seleccionamos a técnica a utilizar para alcançar o objectivo final, construindo assim o modelo de modo a obtermos os resultados pretendidos. De acordo com esta metodologia será realizada a discussão de resultados à medida que os vamos enunciando.

Como descrito anteriormente, os três principais desafios para o desenvolvimento desta solução foram os seguintes:

1. Construção do Perfil Virtual do Cliente de acordo com os seus comportamentos;

2. Recomendação de Conteúdos Personalizados para cada cliente;

3. Identificação dos diferentes perfis de utilização associados ao mesmo número de conta.

A primeira abordagem para a construção da Solução foi a identificação das *features* que podiam ser utilizadas nesta fase do projecto, tanto dos dados do negócio como dos dados obtidos através da correspondência com o IMDb. Uma *feature* em Aprendizagem Automática é uma propriedade possível de estimar, ou seja, uma coluna do conjunto de dados. [9]

Depois de obtermos apenas 30% de sucesso para a correspondência com a base de dados do IMDb, percebemos que analisar os dados apenas com base nestas informações não iria ser conclusivo, pois uma grande percentagem de eventos não iria ter correspondência e portanto, não iria ser possível retirar qualquer tipo de informação sobre o seu comportamento.

De tal forma, começámos a análise utilizando as *features* disponibilizadas pelos dados do negócio, estas *features* abrangem todos os eventos utilizados tornando possível analisar para todos os clientes os padrões identificados.

6.1 Construção do Perfil Virtual do Cliente

De acordo com os desafios anteriormente identificados, o primeiro com que nos deparámos foi a construção do Perfil Virtual do Cliente utilizando os seus comportamentos no serviço em análise. Este ponto corresponde também a uma parte do primeiro objectivo estabelecido para este projecto, construindo assim o Perfil Virtual do Cliente através dos seus históricos para que seja possível identificar o grupo a que este pertence.

A análise dos comportamentos do cliente permite obter conhecimento sobre o seu *feedback* em relação aos conteúdos disponibilizados pelo serviço. Neste caso em especial, é um feedback implícito porque é obtido através das escolhas anteriormente realizadas pelo cliente, ou seja, são registadas as interacções utilizador-item e através destas é construído o Perfil Virtual do Cliente. [22] Nos dados disponibilizados pelo negócio, apesar de possuirmos uma característica que corresponde ao Código da classificação, esta não apresentava valores para todos os eventos, nem tínhamos conhecimento se essa classificação correspondia à realizada pelo cliente no serviço em questão. Por estas razões apresentadas, não foi considerado o *feedback* explícito dos utilizadores para o restante projecto.

A construção do Perfil Virtual do Cliente vai depender directamente da *feature* escolhida. Isto significa que sempre que a alterármos, teremos de recalcular o Perfil Virtual do Cliente.

O Perfil Virtual do Cliente é um vector construído através da média do número de eventos de visualização e *download* de um cliente. Esta métrica foi a escolhida por ter em conta todos os eventos, obtendo assim um valor normalizado, através do qual é possível identificar que conteúdos têm maior importância para o cliente em questão. O cálculo do Perfil Virtual do Cliente é explicado no seguinte exemplo (Tabela 5) onde se atribui o valor ”1”caso a interacção utilizador-item seja observada e o valor ”0”caso contrário. Os valores atribuídos nada estão relacionados com o facto de o cliente gostar ou não do conteúdo em análise, mas sim com o seu interesse naquele tipo de conteúdos. [22]

Tabela 5: Exemplo de um Perfil Virtual do Cliente construído a partir dos géneros dos dados do negócio.

Cliente	**Conteúdo Visualizado**	**Acção**	**Animação**	**Comédia**	**Thriller**	**...**
54321	O Grande Ditador	0	0	1	0	...
	Os Goonies	1	0	0	0	...
	American Outlaws	1	0	0	0	...
	A Janela Secreta	0	0	0	1	...
	Eragon	1	0	0	0	...
	Take Me Home Tonight	0	0	1	0	...
Perfil Virtual do Cliente		0.50	0.00	0.33	0.17	...

Neste exemplo, ao construirmos o Perfil Virtual do Cliente ”54321”, utilizando a *feature* dos géneros disponibilizados pelo negócio, identificamos que este cliente visualizou um total de 6 conteúdos dos quais 50% foram do género Acção, 33% do género Comédia

e 17% do género Thriller. Estas percentagens permitem identificar que géneros (neste caso) têm maior impacto nas decisões do cliente, permitindo assim uma análise que tem em conta o seu comportamento registado neste serviço.

Outra das *features* considerada para este projecto foram os géneros do IMDb que podem ser apresentados de três formas: como género no singular (Acção), como uma combinação de dois géneros (Acção, Comédia) ou como uma combinação de três géneros (Acção, Comédia, Drama). Todos estes factores foram tidos em consideração para o cálculo do Perfil Virtual do Cliente, ou seja, se um cliente viu um conteúdo do género "Acção, Comédia, Drama", os géneros Acção, Comédia e Drama serão assinalados com o valor "1"representando a sua interacção.

A construção do Perfil Virtual do Cliente é independente do número de eventos registados para cada cliente, tendo sido aplicada a todos os clientes que apresentam pelo menos um evento registado neste serviço.

A construção do Perfil Virtual do Cliente é um passo essencial para a Solução desenvolvida pois possibilita a transformação de dados categóricos em métricas possíveis de agrupar. Ao decompormos todas as possibilidades de combinação de forma binária (com os valores "1"e "0"), permitimos a análise de atributos complexos de uma forma simplificada, não perdendo no entanto o detalhe que a *feature* nos dá.

6.2 Identificação de Perfis Comportamentais através de algoritmos de Agrupamento

Os restantes desafios estão relacionados com os outros objectivos estabelecidos para o projecto. A identificação de perfis comportamentais é a base, quer para a recomendação de conteúdos personalizados como para o direccionamento de campanhas *upsell*.

Os perfis comportamentais permitem aumentar o detalhe da análise realizada, ou seja, ao conseguirmos especificar os diferentes comportamentos para um mesmo cliente (igual Número de Conta Cliente), conseguimos aumentar o leque de informações que conhecemos sobre este e assim melhorar tanto o conteúdo que lhe recomendamos nos serviços que costuma utilizar, como as campanhas que lhe são sugeridas passam a ser as mais aprovadas.

Para a identificação dos perfis comportamentais foram então utilizados algoritmos de Aprendizagem Automática para Agrupamento (*clustering*). Este tipo de algoritmos tem como objectivo a identificação de padrões nos dados, e de acordo com a semelhança detectada permite o seu agrupamento, ou seja, permite a identificação de k grupos, que neste caso irão partilhar interesses pelo mesmo tipo de conteúdos de acordo com a *feature*

em análise.

De acordo com as ferramentas disponíveis escolheram-se dois algoritmos distintos para serem aplicados aos dados em análise:

6.2.1 K-Means

Um dos algoritmos utilizados, baseia-se no reagrupamento dos dados em *k* grupos, chamados de *clusters* e consiste em 3 fases principais:

1. Sendo *k* o número total de clusters, escolhe-se *k* pontos aleatórios e iniciam-se assim os *centróides* (centros dos *clusters*).

2. Para cada ponto dos dados em análise, encontra-se o *centróide* mais próximo, utilizando a distância Euclideana, e atribui-se ao respectivo *centróide* este ponto.

3. No fim, percorre-se todos os *k centróides* e para cada um volta-se a computar a sua posição de acordo com todos os pontos que a si estão atribuídos. Neste passo é utilizado o cálculo da média para obter os valores finais do *centróide*.

O algoritmo termina quando após re-computação dos *centróides* não se apresentam alterações. [23]

6.2.2 Latent Dirichlet Allocation (LDA)

O segundo algoritmo utilizado consiste na escolha do número de tópicos *t* que melhor descreve o conjunto de clientes (neste caso reflectindo todos os seus eventos neste serviço através do Perfil Virtual do Cliente). Estes tópicos têm um número fixo e pré-estabelecido de termos que por sua vez são atribuídos a cada tópico de acordo com a sua distribuição multinomial, denominada como o peso do termo. O algoritmo consiste em 2 fases principais:

1. Visto que se trata de um algoritmo com base na distribuição multinomial dos termos, sendo *t* o número total de tópicos, escolhem-se de acordo com a probabilidade de ocorrência cada termo que constituí o tópico em análise.

2. Por fim, depois de atribuídos os pesos para cada termo que constituí os diferentes *t* tópicos, é conferida também uma distribuição multinomial para cada cliente de acordo com a probabilidade de estar mais ou menos relacionado com cada um dos tópicos antes definidos. [24]

6.2.3 Determinação do Número de Grupos através do Método *Elbow*

Para ambos os casos foi necessário determinar o número de grupos a utilizar (de *clusters* ou de tópicos). Para tal foi escolhido o Método "*Elbow*", o método mais antigo para determinar o número de *clusters*.

Este método começa com o número de *clusters* igual a 2 e a partir daí vai-se aumentando este número, um de cada vez, até atingir um valor limite pré-definido. Para cada iteração, sempre que se aumenta o número de *clusters* será calculado o seu custo.

Neste caso a função de custo utilizada foi o somatório do quadrado das distâncias, uma medida que avalia a dispersão. Juntando todos os resultados obtidos ao longo das iterações é construído um gráfico como o da Figura 11.

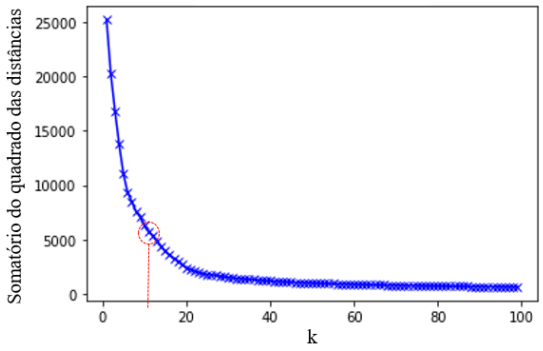

Figura 11: Método *Elbow* para a determinação do número de *clusters* para os Géneros dos dados do negócio.

Por se tratar de um método visual, o seu resultado é inferido através do gráfico resultante (Figura 11). O número de *clusters* óptimo, como o nome do método indica, corresponde ao ”cotovelo”da curva representada. Este ”cotovelo”é assim considerado por ser o ponto onde é possível distinguir os grupos uns dos outros mas, ao mesmo tempo, não se observar elevados valores de dispersão entre os pontos que constituem cada um dos grupos. O número óptimo de *clusters* inferido neste caso é 10 como assinalado na Figura 11 a vermelho. [25]

A ferramenta utilizada para este projecto, o *PySpark* dispõe de quatro algoritmos de Aprendizagem Automática para Agrupamento. Destes quatro, foram escolhidos para analisar o K-Means e o LDA. O K-Means foi escolhido por se tratar do algoritmo mais utilizado pela comunidade de *data scientist* para agrupamento, permitindo agrupar os clientes de acordo com as suas escolhas no serviço em questão. O LDA, normalmente utilizado para identificar documentos semelhantes de acordo com a sua constituição, foi também escolhido para este projecto. [26, 27, 28]

Para ambos os algoritmos anteriormente escolhidos era necessário definir um número de grupos (*clusters* ou tópicos) e portanto utilizámos o algoritmo mais antigo para o fazer, o Método *Elbow*. Através deste método inferiu-se que o número óptimo de grupos seria 10. Este método foi aplicado para todas as variações de dados a utilizar nos métodos de agrupamento obtendo sempre resultados muito semelhantes (Anexo B). De tal forma, para todos os algoritmos aplicados durante o projecto e para todas as suas variações foi sempre considerado o mesmo número de *clusters*.

6.3 Recomendação de Conteúdos

A ideia básica dos Sistemas de Recomendação é que alguns utilizadores partilham os mesmos interesses, isto é, se dois utilizadores no passado tiveram interesse pelo mesmo artigo, a probabilidade de ambos terem interesse num outro artigo no futuro é grande. [29]

Esta foi uma das abordagens da Solução de Conteúdos Personalizados desenvolvida, transformar todas as informações recolhidas pelo ser viço que estamos a analisar em recomendações de conteúdos apropriados, permitindo assim a identificação de grupos de clientes que partilhem as mesmas escolhas.

A identificação dos diferentes grupos existentes para esta amostra de dados permite-nos analisar as escolhas do universo em análise. Com este processo iremos também aumentar o conhecimento sobre o cliente conduzindo assim ao aumento da personalização dos conteúdos recomendados.

6.3.1 Identificação dos conteúdos a recomendar através do K-Means

Começámos por utilizar o K-Means e através dos resultados obtidos foi possível definir três cenários de recomendação possíveis:

Utilizando apenas os Géneros do negócio

Figura 12: Resultados de 9 *clusters* do K-Means realizado com os Géneros dos dados da empresa.
As barras verticais representam o tamanho do universo ao qual corresponde o *cluster* identificado.

Com o objectivo de agrupar os diferentes clientes de acordo com as suas escolhas, começámos por analisar a sua distribuição em relação aos géneros provenientes dos dados do negócio. Escolheu-se esta *feature* porque esta era a que melhor representava o conteúdo

visualizado visto que as do IMDb, por apenas 30% do conteúdo ter correspondência, não iriam permitir a identificação de padrões conclusivos.

Na Figura 12 é possível observar 9 dos 10 *clusters* resultantes do K-Means utilizando como *feature* o Perfil Virtual do Cliente calculado através dos géneros do negócio.

Todos estes *clusters* representam bons resultados porque foi possível especificar para 80% dos clientes, qual é o género predominante do grupo em que se enquadram de acordo com os seus comportamentos.

Quando todos os grupos de clientes se apresentam bem definidos, como os da imagem, sabemos que a recomendação vai ser específica para os géneros que têm maior percentagem, não fazendo sentido recomendar um filme com o género Animação para o grupo de clientes associado ao género Acção.

No entanto falta referir o outro *cluster* que resultou deste processo, o da Figura 13.

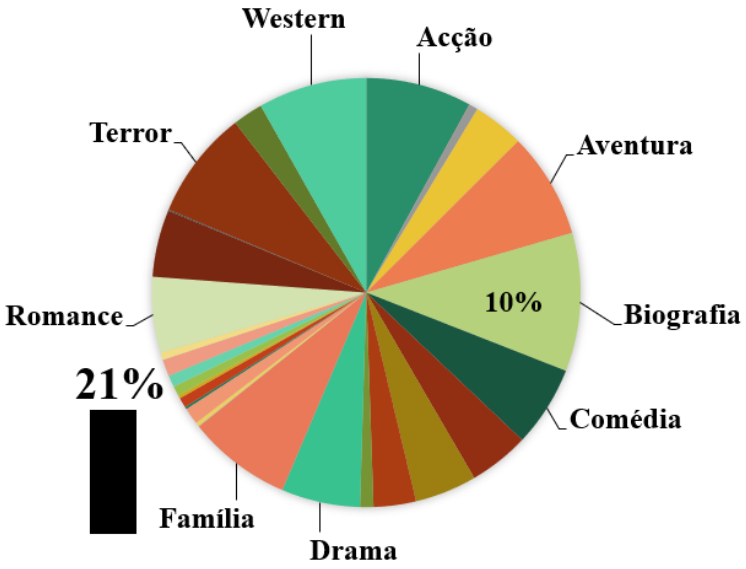

Figura 13: 10º *Cluster* resultado do K-Means realizado com os Géneros dos dados da empresa.
A barra representa o tamanho do universo ao qual corresponde o *cluster* identificado.

De acordo com este resultado, 21% dos clientes analisados não apresentavam distinção suficiente tendo sido agrupados no mesmo *cluster*. Este caso representa uma amostra muito heterogénea, em que não é possível identificar as preferências específicas de todos estes clientes apenas com o processo realizado.

Utilizando apenas esta abordagem, a recomendação não seria personalizada para estes quase [redacted] clientes e, no pior dos casos, teria em conta as preferências de géneros aqui identificadas (os géneros com probabilidade igual ou superior a 10%, neste caso a Biografia). No entanto, de maneira a melhorar a personalização da recomendação, pensámos então em complementar esta análise obtida fazendo correspondência com os dados do IMDb.

Utilizando a Correspondência com os Géneros do IMDb

Analisámos então apenas os clientes que ficaram agrupados no *cluster* cuja recomendação não era específica (Figura 13), utilizando os géneros da correspondência com o IMDb, obtendo os resultados presentes na Figura 14.

Figura 14: Resultados do K-Means realizado com os Géneros do IMDb para os clientes que com os Géneros do negócio não tinham distinção suficiente.
As barras verticais representam o tamanho do universo ao qual corresponde o *cluster* identificado.

Através destes resultados conseguimos aumentar a percentagem de clientes com recomendação personalizada para mais 10% do que tinhamos conseguido anteriormente, tendo assim obtido um total de aproximadamente 90%.

Neste caso, a correspondência com os dados do IMDb tornou-se benéfica, porque apesar de não ter resultado em *clusters* tão bem definidos como os obtidos para os géneros do negócio (Figura 12), conseguimos restringir as opções para a recomendação tendo em conta os géneros que apresentam maiores percentagens de preferência dentro do *cluster*.

No entanto, ainda identificámos dois *clusters* que não apresentaram resultados conclusivos em relação à recomendação. Um deles, o primeiro da Figura 14, apresenta novamente uma mistura muito heterogénea, correspondendo à percentagem de clientes para os quais ainda não foi possível atribuir, de acordo com as suas preferências, a um dos outros *clusters*. No pior dos casos, para estes clientes, a recomendação teria em conta os géneros com maiores percentagens do *cluster*. Para solucionar este caso, o que poderíamos fazer seria, continuar o número de iterações, utilizando os dados do IMDb, de modo a obtermos o máximo de informação possível sobre esta amostra de clientes.

No outro *cluster*, o segundo da Figura 14, encontramos todos os clientes que visualizaram na sua maioria conteúdos para os quais não foi possível fazer correspondência com os dados do IMDb, identificados com o género "Outros". Para estes clientes, a recomendação ou seria efectuada tendo em conta os géneros do negócio anteriormente

identificados com maiores percentagens, ou, visto que o número de iterações com os dados do IMDb nunca dariam mais informações do que as já conhecidas, poderia continuar a iteração com os géneros do negócio, para obtermos maior distinção dos géneros sobre esta amostra de clientes.

Utilizando a Correspondência com as Combinações dos Géneros do IMDb

Até agora já conseguimos identificar recomendações personalizadas para 90% dos clientes da amostra. No entanto, sabemos que a recomendação é mais adequada quanto maior for o seu nível de detalhe e por isso voltámos a utilizar os dados do IMDb para completarmos algumas das conclusões anteriormente identificadas.

Utilizando o *cluster* da Figura 12 que apresenta maior percentagem de clientes, podemos identificar que estes preferem os conteúdos de Acção mas, sabendo que os géneros do IMDb contêm informações sobre as combinações de géneros possíveis de fazer, como podemos concluir que um cliente identificado neste *cluster* gosta do género Acção (no singular) e não da combinação de Acção com outros géneros?

Figura 15: Resultados do K-Means realizado com as Combinações de Géneros do IMDb para os clientes identificados no *cluster* do Género de Acção.
As barras verticais representam o tamanho do universo ao qual corresponde o *cluster* identificado.

Ao aplicar o K-Means utilizando todas as 360 Combinações de Géneros identificadas na base de dados do IMDb (Figura 15), conseguimos aumentar o nível de especificidade na recomendação realizada. Com esta análise, podemos identificar que os clientes que se encontravam no *cluster* do género Acção, de acordo com os clusters identificados, gostam da combinação de Acção com outros géneros e não de conteúdos de Acção em específico.

Tal como previsto, um cliente que goste de Acção pode gostar em específico da combinação ”Acção, Aventura, História”e não da combinação de géneros ”Acção, Comédia”, mas sem este passo, iriamos recomendar a todos os clientes conteúdos do género Acção que podiam não ser os adequados.

Este tipo de análise poderá ser aplicada a qualquer um dos *clusters* anteriormente identificados, permitindo adquirir maior detalhe sobre a recomendação mais apropriada para o cliente em questão.

• Exemplos de Recomendações Obtidas

De acordo com os três cenários anteriormente descritos foi possível recomendar conteúdos personalizados com base no comportamento do cliente registado neste serviço.

Como tal, a abordagem de recomendação sugerida é a Híbrida. Este tipo de recomendação tem em consideração tanto a Recomendação Colaborativa (utilizadores semelhantes preferem os mesmos conteúdos) como a Recomendação com base no Conteúdo (conteúdos semelhantes são preferidos pelo mesmo cliente).

De tal forma, foram desenvolvidos três tipos distintos de recomendação que foram aplicados aos três cenários em análise:

• A Recomendação Colaborativa, que terá em consideração os conteúdos mais visualizados pelos clientes que se enquadram no mesmo *cluster*;

• A Recomendação com base no Conteúdo dos artigos mais visualizados dentro dos géneros preferidos pelo *cluster*, ou seja, identificando os géneros que apresentam maior percentagem seleccionam-se os artigos mais visualizados;

• E a Recomendação com base no Conteúdo de artigos aleatórios dentro dos géneros preferidos, isto é, identificando os géneros que apresentam maior percentagem no *cluster* seleccionam-se aleatóriamente artigos para que, mesmo que um artigo seja novo no catálogo tenha possibilidade de ser recomendado.

1. Recomendação utilizando apenas uma iteração com os géneros do negócio:

Considerando um cliente com o seguinte histórico:

Data do Evento	Conteúdo Visualizado	Género do Negócio
2014	Turbo (v.p.)	Animação
2015	Nutri Ventures - O Reino Verde	Animação
2015	Happy Feet 2 (v.p.)	Animação
2017	Cegonhas	Animação
2017	Transformers: O Último Cavaleiro	Acção
2018	Abelha Maia: Os Jogos do Mal (em hd)	Animação

De acordo com os resultados obtidos na Figura 12, este cliente estaria enquadrado no *cluster* do género Animação. De tal forma, os conteúdos recomendados seriam:

> Para a Recomendação Colaborativa:

Conteúdo Recomendado	**Género do Negócio**
Vaiana	Animação
Cantar!	Animação
Gru - O Maldisposto 3	Animação

> Para a Recomendação com base no Conteúdo dos artigos mais visualizados:

Conteúdo Recomendado	**Género do Negócio**
Vaiana	Animação
Gru - O Maldisposto 3	Animação
Cantar!	Animação

> Para a Recomendação com base no Conteúdo de artigos aleatórios:

Conteúdo Recomendado	**Género do Negócio**
Jake e os Piratas da Terra do Nunca	Animação
Horton e o Mundo dos Quem	Animação
O Caminho para a Fama	Animação

Este caso, representa um cliente que foi facilmente identificado de acordo com os seus históricos no *cluster* da Animação, sendo de tal forma direccionadas todas as recomendações sugeridas.

2. Recomendação utilizando duas iterações para analisar a mistura heterogénea obtida dos géneros do negócio:

Considerando um cliente com o seguinte histórico:

Data do Evento	**Conteúdo Visualizado**	**Género do Negócio**	**Género do IMDb**
2017	Pesadelo	Terror	Documentário, Horror
2017	A Possessão	Terror	Horror, Mistério, Thriller
2018	Contaminação	Thriller	Horror, Sci-Fi, Thriller

De acordo com os resultados obtidos na primeira iteração do K-Means com os géneros do negócio, este cliente estaria enquadrado no *cluster* que apresentava uma mistura heterogénea de clientes com preferências não específicas (Figura 13). De tal forma, os conteúdos recomendados seriam:

> Para a Recomendação Colaborativa:

Conteúdo Recomendado	Género do Negócio
The Revenant: O Renascido	Western
Ronaldo	Desporto
A Bela e o Monstro	Musical

> Para a Recomendação com base no Conteúdo dos artigos mais visualizados:

Conteúdo Recomendado	Género do Negócio
O Lobo de Wall Street	Biografia
Barry Seal, Traficante Americano	Biografia
12 Anos Escravo	Biografia

> Para a Recomendação com base no Conteúdo de artigos aleatórios:

Conteúdo Recomendado	Género do Negócio
Yves Saint Laurent	Biografia
127 Horas	Biografia
Nascido a 4 de Julho	Biografia

Como podemos analisar, este tipo de recomendação não seria de todo a apropriada para o cliente em questão porque, apesar da Recomendação Colaborativa representar as preferências dos clientes do *cluster* em que este se insere, como o género com maior percentagem desse *cluster* é a Biografia, na Recomendação com base no Conteúdo todos os artigos recomendados são desse género, não estando em nada relacionados com o conteúdo visualizado pelo cliente em análise.

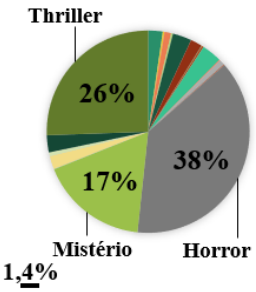

Figura 16: *Cluster* resultante do K-Means aplicado aos Géneros do IMDb.

Através da segunda iteração efectuada com os dados do IMDb (Figura 14) este cliente encontra-se no *cluster* da Figura 16 que corresponde ao grupo de clientes que tem como preferências os géneros Horror, Thriller e Mistério. Como tal, as recomendações sugeridas mudaram para as seguintes:

> Para a Recomendação Colaborativa:

Conteúdo Recomendado	Género do IMDb
Feliz Dia para Morrer	Horror, Mistério, Thriller
Foge	Horror, Mistério, Thriller
Anabelle	Horror, Mistério, Thriller

> Para a Recomendação com base no Conteúdo dos artigos mais visualizados:

Conteúdo Recomendado	Género do IMDb
Anabelle	Horror, Mistério, Thriller
Feliz Dia para Morrer	Horror, Mistério, Thriller
Anabelle 2: A Criação do Mal	Horror, Mistério, Thriller

> Para a Recomendação com base no Conteúdo de artigos aleatórios:

Conteúdo Recomendado	Género do IMDb
Abrigo	Horror, Mistério, Thriller
The Nun: A Freira Maldita	Horror, Mistério, Thriller
Jigsaw: O Legado de Saw	Horror, Mistério, Thriller

Estas recomendações sugeridas já se adequam ao comportamento registado do cliente, pois todos os conteúdos recomendados são representativos do *cluster* onde o cliente está inserido. Estes resultados comprovam que, a realização da segunda iteração permitiu aumentar o número de clientes com uma recomendação mais apropriada de acordo com o seu histórico no serviço.

3. Recomendação utilizando duas iterações para aumentar o nível de detalhe sobre um cliente já classificado:

Considerando um cliente com o seguinte histórico:

Data do Evento	Conteúdo Visualizado	Género Negócio	Género IMDb
2016	Entre Rivais	Acção	Acção, Comédia
2017	Olha que Duas	Acção	Acção, Aventura, Comédia

E sabendo que este cliente se enquadra no *cluster* da Acção da Figura 12, a recomendação sugerida seria:

> Para a Recomendação Colaborativa:

Conteúdo Recomendado	Género do Negócio
Blood Father - O Protector	Acção
Transformers - O Último Cavaleiro	Acção
Velocidade Furiosa 8	Acção

> Para a Recomendação com base no Conteúdo dos artigos mais visualizados:

Conteúdo Recomendado	Género do Negócio
Lucy	Acção
Velocidade Furiosa 8	Acção
Blood Father - O Protector	Acção

> Para a Recomendação com base no Conteúdo de artigos aleatórios:

Conteúdo Recomendado	Género do Negócio
Administrator	Acção
O Homem dos Punhos de Ferro 2	Acção
Olhos de Dragão	Acção

Neste caso, a recomendação seria a correcta porque está direccionada para o género Acção. No entanto, ao analisarmos com os dados do IMDb é possível identificar que este cliente pertence ao *cluster* da combinação de géneros "Acção, Comédia"(Figura 17).

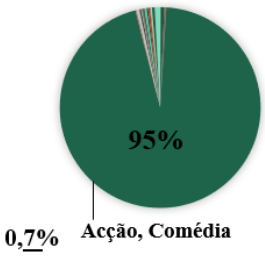

Figura 17: *Cluster* resultante do K-Means aplicado aos Géneros do IMDb para obter maior detalhe sobre o *cluster* de Acção.

Como tal, a recomendação de conteúdo do género Acção seria demasiado generalista para este cliente, pois a sua preferência é realmente a combinação dos géneros Acção e Comédia. Como tal, tendo isso em consideração, as recomendações sugeridas seriam:

> Para a Recomendação Colaborativa:

Conteúdo Recomendado	Género do IMDb
Entre Rivais	Acção, Comédia
Polícias em Grandes Apuros	Acção, Comédia
The Revenant: O Renascido	Acção, Aventura, Biografia

> Para a Recomendação com base no Conteúdo dos artigos mais visualizados:

Conteúdo Recomendado	Género do IMDb
Polícias em Grandes Apuros	Acção, Comédia
A Toda a Velocidade	Acção, Comédia
Blood Father - O Protector	Acção, Comédia

> Para a Recomendação com base no Conteúdo de artigos aleatórios:

Conteúdo Recomendado	Género do IMDb
Thai Fighter	Acção, Comédia
Anacleto: Agente Secreto	Acção, Comédia
Seguranças de Alto Risco	Acção, Comédia

Estas recomendações seriam muito mais apropriadas ao cliente porque são específicas tendo em conta as suas preferências. Não teria lógica recomendar conteúdos do *cluster* "Acção, Aventura, História"para este cliente porque o que ele realmente prefere, de

acordo com o seu histórico, é a combinação ”Acção, Comédia”.

Dos três cenários anteriormente descritos podemos concluir que com o aumento do número de iterações melhoramos o nível de especificidade da recomendação, tal como foi comprovado com os exemplos de recomendação analisados.

Num primeiro cenário, apenas utilizando as *features* do negócio obtivémos aproximadamente 80% de sucesso, ou seja, foi possível identificar todos estes clientes num grupo específico de acordo com os seus comportamentos no serviço em análise, o que permitiu recomendar conteúdos personalizados.

Como demonstrado no segundo cenário, para os clientes em que não conseguimos sequer obter uma recomendação direccionada apenas pela primeira iteração, é essencial fazer a segunda iteração com os dados obtidos pela correspondência com o IMDb. Neste caso, a segunda iteração permitiu aumentar a percentagem de clientes para os quais foi possível recomendar conteúdos personalizados para aproximadamente 90%, no entanto, estas iterações poderão continuar a ser realizadas até conseguirmos identificar para todos os clientes uma recomendação direccionada.

No terceiro cenário apresentado, conseguimos aumentar o nível de especificidade da recomendação, ou seja, aumentando o nível de detalhe das *features* fomos capazes de agrupar os clientes de acordo com as suas preferências mais exactas, com o maior nível de detalhe possível. Este tipo de análise poderá ser efectuada para qualquer outro cluster obtido nos outros dois cenários anteriores, aumentando o número de informações conhecidas sobre o clinte.

6.3.2 Identificação dos conteúdos a recomendar através do LDA

O segundo algoritmo utilizado nos dados da amostra para a Recomendação de Conteúdos foi o *Latent Dirichlet Allocation* (LDA). Este algoritmo normalmente é utilizado para identificar os principais tópicos presentes em diversos documentos. No nosso caso em estudo, pensámos neste algoritmo como uma abordagem para agrupar os clientes de acordo com os géneros presentes no seu Perfil Virtual do Cliente.

Este algoritmo permite a escolha do número de termos que constituem um tópico, e portanto numa perspectiva de analisar com maior detalhe as recomendações sugeridas começámos por escolher o número de termos igual a 3. Os resultados obtidos quando aplicado à *feature* dos géneros do negócio, foram os seguintes:

Tabela 6: Tópicos definidos a partir do LDA utilizando os géneros do negócio.

Tabela 7: Exemplo dos resultados obtidos a partir do LDA utilizando os géneros do negócio..

Cliente	PVC	Tópico com Maior Peso
1	1 Drama	8 - [Drama, Comédia, Acção]
2	0.75 Acção +0.25 Comédia	7 - [Acção, Comédia, Drama]
3	1 Animação	3 - [Animação, Romance, Comédia]
4	1 Biografia	5 - [Adultos, Biografia, Animação]

Como é possível observar na Tabela 6, foram definidos 10 tópicos com 3 termos cada um. Cada um desses termos que constituem o tópico tem um determinado peso associado à sua probabilidade de ocorrência nos dados da amostra. Ou seja, o género mais provável de ocorrer é o que tem o maior peso, sendo neste caso apresentado o género Acção.

Na Tabela 7 podemos analisar alguns exemplos dos resultados obtidos para este algoritmo. Identificámos que, para clientes com comportamentos evidentes, ou seja, clientes como o exemplo do cliente "1", o seu Perfil Virtual do Cliente demonstra que prefere conteúdos do género Drama. Neste caso e utilizando o algoritmo em análise, serão recomendados conteúdos que englobam outros dois géneros presentes no tópico, Comédia e Acção, que nada estão relacionados com as preferências identificadas no Perfil Virtual do Cliente. Outro destes casos e até mais evidente é o cliente "4"que, por apenas se adequar ao tópico 5, a recomendação conterá também conteúdos dos géneros Adulto e Animação. Esta combinação de géneros não apresenta qualquer lógica no contexto apresentado, demonstrando que este algoritmo não será o apropriado para realizar a Recomendação de Conteúdos desejada.

O LDA não nos permite perceber as relações existentes entre géneros através dos comportamentos dos clientes. Esta análise demonstra que a selecção dos termos que

constituem os diferentes tópicos é exclusivamente probabilística, ou seja, são seleccionados os 10 géneros com maior probabilidade entre os clientes não tendo em consideração as interacções entre eles. Este tipo de análise não é apropriada para o problema em análise porque, por não nos permitir identificar as interacções entre os diferentes géneros, a selecção de tópicos é completamente indiferente às relações entre os termos.

6.4 Direccionamento de Campanhas através do K-Means

Para cumprir o 3º objectivo proposto neste projecto, o direccionamento de campanhas, é essencial obter mais informações sobre o perfil dos clientes. Isto é, a única informação disponível pela qual construímos o Perfil Virtual do Cliente é o Número de Conta Cliente, mas este não nos permite conhecer realmente o cliente e as suas preferências. Um perfil pode consistir em mais do que uma pessoa a utilizar o mesmo serviço, e se realmente for mais do que uma pessoa, estas vão preferir todas o mesmo tipo de conteúdos ou conteúdos distintos?

[redacted] Até aos dias de hoje, todas as informações conhecidas sobre o perfil do cliente são-lhe questionadas via *call-center*, mas o direccionamento de campanhas não vai ser personalizado o suficiente se não houver uma análise com base no comportamento do cliente.

Podemos pensar que aquele cliente, por se tratar de um jovem de 30 anos que vive sozinho, apenas está interessado em conteúdos relacionados com Filmes, por ser o tipo de conteúdos que normalmente um cliente do sexo masculino nesta faixa etária gostaria. Mas e se o nosso cliente apenas visualizar conteúdos Infantis por ser a sua preferência? O direccionamento da campanha neste caso iria falhar porque não traçámos o perfil deste cliente de acordo com o seu comportamento.

Utilizando outras *features* disponíveis na amostra de dados é-nos possível analisar o perfil comportamental do cliente. A *feature* que escolhemos analisar foram as Descrições das Categorias pois dá-nos informação sobre a categoria em que o conteúdo visualizado se adequa, ou seja, de uma maneira geral, pois só existem [redacted] categorias distintas, é possível identificar diferentes tipos de utilizadores presentes no mesmo Perfil Virtual do Cliente. Os resultados obtidos encontram-se na Figura 18.

Figura 18: Resultados do K-Means realizado com as Categorias dos dados da empresa.
As barras verticais representam o tamanho do universo ao qual corresponde o *cluster* identificado.

Através dos *clusters* obtidos pelo K-Means ao utilizar a *feature* da Descrição das Categorias dos conteúdos escolhidos pelos clientes (Figura 18), podemos identificar 10 tipos diferentes de clientes de acordo com os seus comportamentos no serviço. Cada um destes perfis identificado, permite o direccionamento de campanhas muito mais apropriadas porque através destes resultados conseguimos perceber que tipo de conteúdos correspondem a cada cliente.

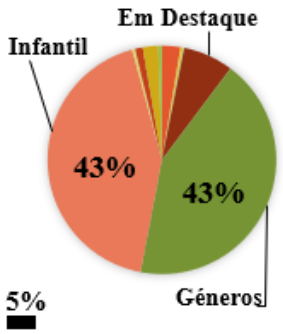

Figura 19: Um dos *clusters* resultante do K-Means aplicado às Categorias do negócio.

Para um cliente que se enquadra no *cluster* que tem 94% de preferência pelo conteúdo *premium* (Mais Packs) terá todo o sentido oferecer-lhe tanto promoções de outros canais *premium* como recomendar-lhe conteúdos *Pay Per View*. Já para um cliente enquadrado no *cluster* da Figura 19, visto que o tipo de conteúdos que este prefere é bastante variado, poderá ou não corresponder a um caso em que um Número de Conta Cliente é utilizado pelos diferentes constituintes de uma família. Dado isto, poderia ser-lhe recomendado por exemplo adquirir mais uma *box* ou um *upgrade* de serviço adequado para famílias. Neste último caso poderiamos estar a direccionar erradamente a campanha porque um cliente enquadrado neste *cluster* pode também ser um cliente que apenas tem preferências muito variadas.

Através da identificação do *cluster* onde o cliente se enquadra estamos a confirmar realmente se o comportamento apresentado está ou não de acordo com as informações conhecidas. Mesmo não tendo certezas sobre o número de pessoas reais que utilizam o mesmo acesso conseguimos ter certezas sobre o tipo de pessoas, permitindo assim direccionar as campanhas mais adequadas de acordo com o seu perfil comportamental neste serviço.

Este tipo de análise contribuirá para o aumento do número de atributos recolhidos sobre o cliente, auxiliando o negócio nas suas estratégias de analítica avançada.

Capítulo 7

Avaliação

A Avaliação de um Sistema de Recomendações real implicaria a sua Implementação num contexto também real e isso não é fácil nem barato para uma empresa de realizar. Este processo incluiria a implementação dos diferentes cenários com os três tipos de recomendação diferentes e, só depois de testados com um grupo de clientes significativo, é que poderíamos retirar algum tipo de conclusões sobre a eficiência dos diferentes cenários.

Como explicado anteriormente, não foi possível realizar a Avaliação da solução construída. No entanto, foram elaboradas duas possíveis abordagem que podem ser utilizadas para testar a solução, uma antes da sua implementação e outra durante:

Antes da Implementação

Numa primeira abordagem, antes da implementação da Solução construída, poderíamos dividir o conjunto de dados disponíveis em dois grupos. Um dos grupos teria todos os dados até aos últimos 6 meses antes do último evento registado, e o outro grupo de dados apenas teria os correspondentes aos últimos 6 meses.

A ideia passa por testar a Solução construída implementando-a ao primeiro conjunto de dados (com os dados até aos últimos 6 meses) e depois testaríamos com os dados dos últimos 6 meses. Desta forma, poderíamos retirar conclusões como 80% das recomendações realizadas corresponderam à realidade dos eventos que foram visualizados, percebendo se o conteúdo registado nos últimos 6 meses correspondia ao comportamento previsto.

Eis uma representação, com dados meramente fictícios criados para este exemplo:

O cliente X tem o seguinte histórico:

Data	Conteúdo Visualizado	Género Negócio	Género IMDb
2017/06	Entre Rivais	Acção	Acção, Comédia
2017/12	O Irmão Secreto	Acção	Acção, Comédia
2018/02	Thai Fighter	Acção	Acção, Comédia
2018/05	O Outro Lado da Fronteira	Acção	Crime, Drama
2018/12	Polícias em Grandes Apuros	Acção	Acção Comédia
2019/03	Vingança ao Anoitecer	Acção	Drama, Thriller

Este cliente para o 1º cenário (apenas utilizando os géneros do negócio) estaria enquadrado no *cluster* da Acção e seriam-lhe recomendados os seguintes conteúdos:

Recomendação	Conteúdo Recomendado	Género do Negócio
Colaborativa	Blood Father - O Protector	Acção
	Transformers - O Último Cavaleiro	Acção
	Velocidade Furiosa 8	Acção
Conteúdo - Artigos mais visualizados	Lucy	Acção
	Velocidade Furiosa 8	Acção
	Blood Father - O Protector	Acção
Conteúdo - Artigos aleatórios	Administrator	Acção
	O Homem dos Punhos de Ferro 2	Acção
	Olhos de Dragão	Acção

O que significa que, analisando os eventos dos últimos 6 meses esta recomendação não seria a adequada, pois nenhum dos conteúdos visualizado pelo cliente nos últimos 6 meses faz parte da lista de conteúdos recomendados pela solução. Já para o 2º cenário (utilizando as combinações de géneros possíveis do IMDb) o cliente estaria enquadrado no *cluster* da Acção e Comédia e seriam-lhe recomendados os seguintes conteúdos:

Recomendação	Conteúdo Recomendado	Género do IMDb
Colaborativa	Entre Rivais	Acção, Comédia
	Polícias em Grandes Apuros	Acção, Comédia
	The Revenant: O Renascido	Acção, Aventura, Biografia
Conteúdo - Artigos mais visualizados	Polícias em Grandes Apuros	Acção, Comédia
	A Toda a Velocidade	Acção, Comédia
	Blood Father - O Protector	Acção, Comédia
Conteúdo - Artigos aleatórios	Thai Fighter	Acção, Comédia
	Anacleto: Agente Secreto	Acção, Comédia
	Seguranças de Alto Risco	Acção, Comédia

Podemos observar que o evento realizado a 2018/12 corresponde a um dos conteúdos recomendados pela solução construída, obtendo assim uma taxa de 50% de sucesso.

Na Implementação

Quando a solução for implementada, como descrito anteriormente, será necessário testar os diferentes cenários apresentados e os diferentes tipos de recomendações possíveis.

Para tal, uma das abordagens possíveis seria introduzir um *feedback* explícito que nos permitisse realmente perceber:

1. Se o cliente está interessado no conteúdo que lhe é recomendado,

2. E se o cliente ao estar interessado no conteúdo que lhe é recomendado, gostou realmente da escolha sugerida.

Estes dois testes são muito diferentes um do outro, na realidade, a recomendação à primeira impressão (quando os títulos nos são apenas apresentados) pode parecer a correcta para o cliente mas depois, o conteúdo visualizado pode não ir de encontro às suas expectativas. Este tipo de testes também nos irá fornecer informações sobre a validação dos dados utilizados para construir o perfil do cliente de acordo com o seu comportamento.

Eis uma representação, com dados meramente fictícios criados para este exemplo:

O cliente Y tem o seguinte histórico:

Data	Conteúdo Visualizado	Género Negócio
2017/06	Hércules	Animação
2017/12	Inspector Gadget 2	Animação
2018/02	Monstros: a Universidade	Animação
2018/05	O Pequeno Vampiro	Animação

Segundo o 1º cenário descrito (apenas utilizando os géneros do negócio) este cliente estaria enquadrado no *cluster* da Animação e portanto os conteúdos recomendados seriam os seguintes:

Conteúdo Recomendado	Género do Negócio	Interesse	Apreciação
Vaiana	Animação		
Cantar!	Animação		
Gru - O Maldisposto 3	Animação	x	+1
Jake e os Piratas da Terra do Nunca	Animação		
Horton e o Mundo dos Quem	Animação		
O Caminho para a Fama	Animação		

Neste caso aqui demonstrado, para os 6 títulos recomendados pela solução o cliente Y demonstrou interesse por um dos títulos, para o qual também demonstrou a sua apreciação depois da sua visualização, demonstrando assim que a recomendação foi bem sucedida.

Estes conjuntos de testes terão de análisar os diferentes casos de uso possíveis para a solução, como por exemplo:

• Quando um cliente tem preferência, de igual forma, por dois géneros completamente opostos;

• Quando o cliente não está enquadrado em nenhum dos *clusters* para os quais é possível personalizar a recomendação;

• Quando os conteúdos no histórico do cliente, na sua maioria, fazem parte dos 70% de conteúdos para os quais não foi possível obter correspondência com o IMDb, e portanto a sua recomendação apenas tem como base os géneros do negócio;

Apenas com uma abordagem de avaliação completa que cubra todos os casos possíveis para os diferentes cenários, poderemos retirar conclusões em relação à eficácia e precisão da solução construída neste projecto. Com esta avaliação será possível concluir se esta solução é ou não melhor que outras alternativas e comprovar que cenário (com mais ou menos detalhe, utilizando uma ou mais *features*) é o mais adequado para que a recomendação sugerida seja o mais personalizada possível.

Capítulo 8

Conclusão

Este projecto realizou-se num contexto empresarial e como tal, a primeira abordagem foi um levantamento das necessidades do negócio e de que forma os meus conhecimentos académicos podiam melhorar o que hoje era aplicado nos seus serviços.

Durante o projecto surgiram algumas limitações para as quais foi necessário repensar o plano inicialmente proposto. A disponibilização deste tipo de dados, principalmente de dados com informações pessoais, é um processo bastante demoroso associado a várias políticas de protecção de dados, em especial nos dias de hoje com todo o mediatismo envolvido nesta temática. Desta forma, para além da dificuldade que foi obter acessos a este tipo de dados, também foi limitada a quantidade de características disponíveis para a análise.

Outra das limitações está relacionada com o facto de que todo o projecto foi realizado em *black box*, ou seja, para a construção da solução apresentada, não foi considerada nenhuma informação sobre as características do sistema de recomendações hoje utilizado pelo negócio, pois essas não eram conhecidas.

A primeira fase do projecto consistiu na identificação das fontes de dados a utilizar. A empresa de telecomunicações através da qual foi possível obter os dados utilizados neste projecto, tem uma Base de Dados com inúmeros clientes, armazenando assim informações preciosas capazes de revelar padrões ainda desconhecidos mas essenciais para o seu negócio. Os dados do negócio inicialmente estavam divididos em eventos de visualização e de *download*, como tal, na preparação de dados foi necessário convergir estes dois conjuntos de dados num só, para que a solução construída tivesse como base todos os comportamentos registados do cliente.

Para além dos dados do negócio, também quisémos complementar esta solução com os dados do IMDb, aumentando assim o universo de características disponíveis para a análise. Ao fazer a correspondência dos conteúdos disponibilizados pelo negócio com

os títulos existentes no IMDb, apenas conseguímos 30% de sucesso pois a maioria dos conteúdos na base de dados do negócio tinham caracteres especiais nos nomes das ofertas, ou por exemplo, o código do ano errado dificultando assim a correspondência entre estas duas bases de dados.

Como o nosso principal objectivo era a segmentação do cliente, decidimos que iriamos utilizar algoritmos de Agrupamento de modo a perfilar os diferentes clientes de acordo com as suas escolhas registadas. Para tal necessitávamos de duas coisas:

1. Uma representação normalizada do cliente que nos permitisse aplicar os algoritmos de Aprendizagem Automática,

2. Definir o número de *clusters* a utilizar nestes mesmos algoritmos.

Para o primeiro ponto, construímos o Perfil Virtual do Cliente, onde, através da análise dos históricos do cliente é calculado um vector que represente as suas escolhas no serviço. Para o segundo ponto, utilizámos o Método *Elbow* que nos permitiu inferir através do gráfico resultante que o número de *clusters* a utilizar ao longo do projecto seria 10.

De acordo com a ferramenta escolhida para a solução, tinhamos 4 algoritmos distintos de Agrupamento. O K-Means, o Latent Dirichlet Allocation (LDA), o Bisecting K-Means e o Gaussian Mixture Model. Numa primeira investigação, qualquer um dos 4 algoritmos pareceu-nos apropriado para o problema em questão, no entanto tivémos de começar a análise dos dados escolhendo o algoritmo, que dento dos diponíveis, era o mais popular para este tipo de abordagem, o K-Means.

Para a construção da Solução de Recomendação de Conteúdos Personalizados, começámos por analisar a *feature* disponível por parte do negócio mais apropriada para perfilar os clientes, o género dos conteúdos presentes nos seus históricos. Esta análise permitiu-nos identificar 3 cenários de recomendação distintos:

1. Recomendação de Conteúdos com apenas uma iteração sobre as *features* dos dados do negócio.

2. Recomendação de Conteúdos utilizando uma segunda iteração (com base nos géneros do IMDb) sobre os dados para os quais não se obteve uma recomendação conclusiva no cenário anterior.

3. E uma Recomendação de Conteúdos onde se aplica uma segunda iteração, utilizando os dados obtidos do IMDb, para aumentar o nível de detalhe da recomendação conseguida no primeiro cenário.

Para além destes três cenários acima descritos também foram utilizados três tipos diferentes de Recomendação:

1. Recomendação Colaborativa, que tem em conta as preferências dos clientes que estão enquadrados no mesmo *cluster* que o cliente em análise.

2. Recomendação com base no Conteúdo dos títulos, através da identificação dos géneros com maiores probabilidades no *cluster* onde se enquadra o cliente.

2.1. Recomendação dos conteúdos mais escolhidos dentro destes géneros.

2.2. Recomendação de conteúdos aleatórios dentro destes géneros.

O segundo algoritmo utilizado foi o LDA, no entanto, por se tratar de um algoritmo probabilistico, não foi possível obter uma segmentação dos clientes de acordo com o conteúdo presente nos seus históricos. Obteve-se assim, uma recomendação com base nos géneros com maiores probabilidades dentro das escolhas registadas dos clientes. Concluiu-se que, este tipo de recomendação não seria a apropriada por não ter em consideração, no seu processo de agrupamento, as relações existentes entre os diferentes géneros.

Por fim, para alcançarmos o terceiro objectivo voltámos a utilizar o K-Means para segmentar os clientes, de acordo com a categoria dos conteúdos presentes nos seus registos. Esta análise foi importante para o direccionamento de campanhas pois permitiu retirar informações mais relevantes para o negócio, como o tipo de conteúdos em que o cliente provavelmente estaria interessado.

Todos estes resultados permitiram cumprir os três objectivos definidos para este projecto. Foi possível construir o Perfil Virtual do Cliente e através deste vector criado, identificar as diferentes preferências do cliente. Em relação ao segundo e terceiro objectivos, foi possível obter Recomendações de Conteúdos Personalizados para 90% da amostra utilizada e através da *feature* das categorias foi possível obter um maior conhecimento sobre o cliente, direccionando-lhe assim as campanhas mais apropriadas.

8.1 Trabalho Futuro

A Solução desenvolvida tem como finalidade a sua implementação no contexto real e como tal, existem sempre implementações extra que, podem ou devem, ser realizados no futuro.

8.1.1 Implementações Extra

Correspondência com o IMDb

É possível melhorar a percentagem de correspondência entre os dados do negócio e os dados do IMDb. Neste momento apenas 30% dos conteúdos disponibilizados pelo negócio são possíveis de relacionar com as informações disponibilizadas pelo IMDb. O aumento desta percentagem conduzirá ao aumento da percentagem de clientes para os quais é possível recomendar conteúdos personalizados.

Aumentar o número de variáveis em análise

Outra implementação possível é o aumento do leque de variáveis que irão caracterizar os perfis construídos, como a utilização por exemplo da Data do Evento, Classificação

ou até o aumento de atributos recolhidos do IMDb, como os dados relacionados com os Directores, Realizadores e Actores, por exemplo.

Todas estas variáveis irão aumentar o número de possibilidades capazes de influenciar a personalização do conteúdo recomendado. Por exemplo, a variável Data permitirá retirar conclusões relacionadas com a importância do horário da visualização ou até mesmo se as diferentes alturas do ano alteram o comportamento dos clientes.

Recomendação de títulos para o negócio

Outra das sugestões que seria possível implementar nesta Solução seria a utilização dos dados adquiridos através do IMDb, direccionando conteúdos para o negócio disponibilizar no seu serviço. Isto é, com o aumento do universo de dados disponível, através da correspondência dos conteúdos do IMDb, podemos prever que conteúdos seriam mais vantajosos para o negócio disponiblizar no serviço em análise, desta forma, o negócio direccionaria as suas ofertas de acordo com o comportamento previsto dos clientes que utilizam o serviço em questão.

Outras abordagens de análise

Neste momento, a Solução construída necessita de várias iterações para conseguir cruzar os resultados obtidos, mas e se conseguissemos atribuir ponderações diferentes que representassem as combinações de géneros disponibilizadas pelo IMDb? Desta forma, iriamos melhorar os tempos de processamento para obter os mesmos resultados porque conseguiriamos analisar todas as opções numa única iteração.

Por exemplo, se um cliente viu um filme do género ”Acção, Comédia, Drama”, podíamos atribuir na construção do seu Perfil Virtual do Cliente para o género ”Acção, Comédia, Drama”o valor ”1”, depois nas possibilidades ”Acção, Comédia”, ”Acção, Drama”e ”Comédia, Drama”atribuíamos o valor ”0.33”(correspondente a um terço) e aos géneros individuais ”Acção”, ”Comédia”e ”Drama”atribuiríamos o valor ”0.17”(correspondente a um sexto). Esta abordagem possibilitaria a implementação de todos as possibilidades numa só iteração.

Apêndice A

Diagrama organização dados IMDb

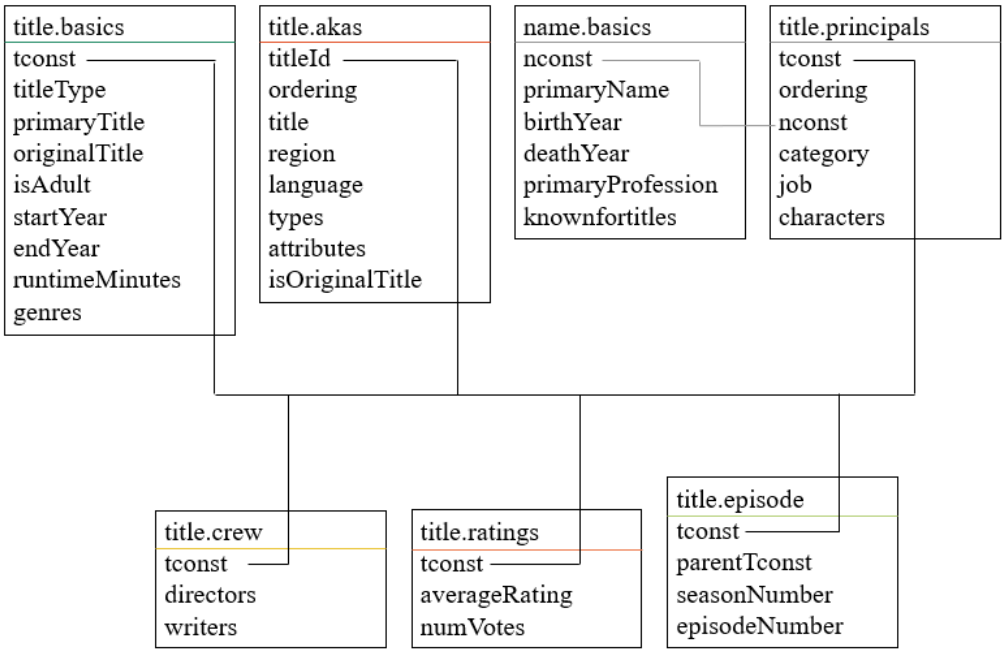

Figura 20: Diagrama com a organização original dos dados do IMDb.

Apêndice B

Resultados Método *Elbow* para todas as *features* analisadas

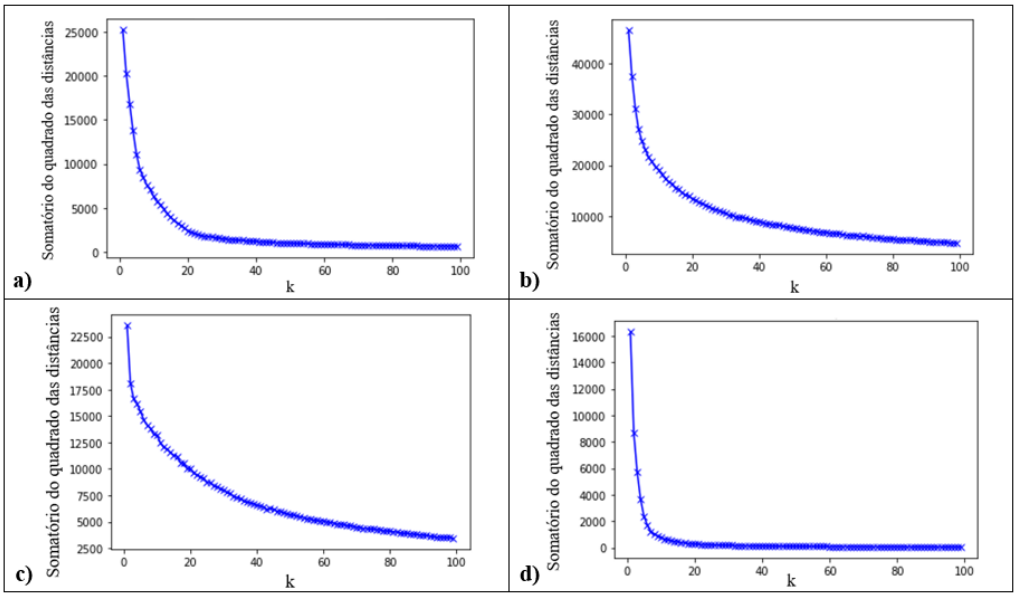

Figura 21: Método *Elbow* para todas as *features* analisadas neste projecto.
a) para os géneros do negócio; **b)** para os géneros do IMDb (individuais); **c)** para os géneros do IMDb (combinações); **d)** para as categórias do negócio.

Bibliografia

[1] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi, "The State of the Art in Integrating Machine Learning into Visual Analytics," *Computer Graphics Forum*, 2017.

[2] I. H.Witten, E. Frank, M. A. Hall, and C. J.Pal, "*Data Mining Pratical Machine Learning Tools and Techniques*," 2011. Morgan Kaufmann Publishers.

[3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "*CRISP-DM 1.0 - Step-by-step data mining guide*," 2000. CRISP-DM Consortium.

[4] L. Bottou, F. E. Curtis, and J. Nocedal, "*Optimization Methods for Large-Scale Machine Learning*," 2018. Society for Industrial and Applied Mathematics.

[5] P. Hall, W. Phan, and K. Whitson, "*The Evolution of Analytics (Opportunities and Challenges for Machine Learning in Business)*," 2016. O'REILLY.

[6] A. Géron, "*Hands-On Machine Learning with Scikit-Learn and TensorFlow*," 2017. O'REILLY.

[7] S. Raschka and V. Mirjalili, "*Python Machine Learning, Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*," 2017. Packt.

[8] I. Bose and R. K.Mahapatra, "*Business data mining - a machine learning perspective*," 2001. Elsevier Science.

[9] M. Kantardzic, "*Data Mining - Concepts, Models, Methods, and Algorithms*," 2011. Institute of Electrical and Electronics Engineers.

[10] D. L.Olson and D. Delen, "*Advanced Data Mining Techniques*," 2008. Springer.

[11] G. Adomavicius and A. Tuzhilin, "*Using Data Mining Methods to Build Customer Profiles*," 2001. Institute of Electrical and Electronics Engineers.

[12] P. Schubert and M. Koch, "*The power of Personalization: Customer Collaboration and Virtual Communities*," 2002. Association for Information Systems.

[13] G. Adomavicius and A. Tuzhilin, "*Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*," 2005. Institute of Electrical and Electronics Engineers.

[14] D. Simchi-Levi, N. Mulani, A. Fano, B. McCarthy, T. Villatoro, and L. Sheppard, "*Annual Research Update: The Accenture and MIT Alliance in Business Analytics*," 2014. Accenture.

[15] J. Zoghby and J. Schneider, "*Accenture Customer Insight, Creating Value through Actionable Customer Intelligence*," 2013. Accenture.

[16] G. Galassi and C. Castiglioni, "*Accenture Video Analytics*," 2017. Accenture.

[17] R. Wirth, "*CRISP-DM: Towards a Standard Process Model for Data Mining*," 2000. Proceedings Of The 4th International Conference On The Practical Applications Of Knowledge Discovery And Data Mining.

[18] CrowdFlower, "*Data Scientist Report*," 2017.

[19] G. Piatestsky, "*Python eats away at R: Top Software for Analutics, Data Science, Machine Learning in 2018: Trends and Analysis*," [Online]. Disponível em: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html. [Consultado em 23 Novembro 2018].

[20] H. Luu, "*Beginning Apache Spark 2: With Resilient Distributed Datasets, Spark SQL, Structured Streaming and Spark Machine Learning Library*," 2018. Apress.

[21] P. Singh, "*Machine Learning with PySpark: With Natural Language Processing and Recommender Systems*," 2018. Apress.

[22] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han, "*Personalized Entity Recommendation: A Heterogeneous Information Network Approach*," 2014. ACM.

[23] M. Kumar, D. Yadav, A. Singh, and V. K. Gupta, "*A Movie Recommender System: MOVREC*," 2015. International Journal of Computer Applications.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "*Latent Dirichlet Allocation*," 2003. Journal of Machine Learning Research.

[25] T. M. Kodinariya and P. R. Makwana, "*Review on determining number of Cluster in K-Means Clustering*," 2013. International Journal of Advance Research in Computer Science and Management Studies.

[26] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "*Top 10 algorithms in data mining*," 2008. Springer.

[27] B. Chao and A. Sirmorya, "*Automated Movie Genre Classification with LDA-based Topic Modeling*," 2016. International Journal of Computer Applications.

[28] R. Katarya, "*Movie Recommender System with Metaheuristic Artificial Bee*," 2018. The Natural Computing Applications Forum.

[29] S. Rajarajeswari, S. Naik, S. Srikant, M. K. S. Prakash, and P. Uday, "*Movie Recommendation System*," 2019. Springer Nature.

[30] N. Bhatia and P. Patnaik, "*Netflix Recommendation based on IMDB*," 2008.

[31] Y. Zhang, M. Chen, D. Huang, D. Wu, and Y. Li, "*iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization*," 2015. Future Generation Computer Systems.

[32] Z. Wang, J. Liao, Q. Cao, H. Qi, and Z. Wang, "*Friendbook: A Semantic-based Friend Recommendation System for Social Networks*," 2013. Institute of Electrical and Electronics Engineers.

[33] B. Smith and G. Linden, "*Two Decades of Recommender Systems at Amazon.com*," 2017. Institute of Electrical and Electronics Engineers.