INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Impacto de caraterísticas de consumo na perceção dos utilizadores: Conteúdo televisivo e cinematográfico

Mestrado:

Business Analytics

Mestrando:

Diogo Alberto Pereira Figueiredo

Orientação:

Professora Graça Trindade, Professora Auxiliar, Departamento de Métodos Quantitativos para Gestão e Economia

Professora Maria da Conceição Santos, Professora Associada, Departamento de Marketing, Operações e Gestão Geral

Outubro, 2022

BUSINESS SCHOOL

Impacto de caraterísticas de consumo na perceção dos utilizadores: Conteúdo televisivo e cinematográfico

Mestrado:

Business Analytics

Mestrando:

Diogo Alberto Pereira Figueiredo

Orientação:

Professora Graça Trindade, Professora Auxiliar, Departamento de Métodos Quantitativos para Gestão e Economia

Professora Maria da Conceição Santos, Professora Associada, Departamento de Marketing, Operações e Gestão Geral

Outubro, 2022

Agradecimentos

Um agradecimento para quem contribuiu, direta ou indiretamente, para o desenvolvimento e conclusão do presente estudo.

À orientação recebida e disponibilidade demonstrada pelas docentes que supervisionaram o progresso do estudo.

À minha família, pelo apoio e acompanhamento do progresso efetuado ao longo da duração da realização da investigação.

À amizade dos dois "*focus testers*" que preencheram o questionário realizado antes da sua disseminação, oferecendo dicas e *insight* para melhoramento do mesmo.

À *account manager* da minha entidade patronal, que se predispôs a assistir na distribuição do questionário pelos colaboradores da empresa, desbloqueando um dos obstáculos apresentados na realização do estudo.

Estou grato,

Diogo

Sumário

O presente trabalho de investigação visa analisar caraterísticas de utilização de consumidores de conteúdos de entretenimento, televisivos e cinematográficos. Consequentemente, pretende determinar o impacto que estes hábitos ou características de consumo têm na satisfação e qualidade percecionada dos consumidores de conteúdo de entretenimento, qual quer que seja a modalidade de consumo escolhida.

A revisão da literatura procurou verificar que áreas científicas são invocadas no âmbito desta temática, que tipologias de estudo, métricas e dados são empregados, de forma a perceber o grau de adequação da direção assumida para este estudo, servindo também como um exercício preliminar à formulação das intenções de estudo. Por fim, procura referenciar que outras conclusões já foram retiradas para as questões formuladas no contexto desta investigação.

O questionário produzido foi disseminado, principalmente, com recurso à rede interna de colaboradores da entidade patronal do mestrando, tirando também partido da sua presença nas diversas redes sociais.

Os dados produzidos foram carregados e tratados via *Jupyter Notebook*, uma plataforma computacional/editor de texto, onde foram realizadas e demonstradas as diversas consultas. As bibliotecas linguísticas *Numpy*, *Pandas* e *Scikit-Learn* foram a fundação da formulação destas consultas, que têm como base a linguagem programática *Python*. O desenvolvimento deste trabalho, precedido pelo tratamento e processamento dos dados gerados, culminou em previsões que relacionaram as variáveis-chave identificadas, através de modelação preditiva.

Palavras-chave: Consumo, Televisivo, Cinematográfico, Modelação, Preditiva, *Python*.
Sistema de classificação JEL: M20, L82

Abstract

The present investigative work looks to analyze consumption patterns among entertainment content consumers, focused on the television and film industries. The goal was to determine the correlation between these patterns or characteristics and perceived levels of satisfaction or quality of the content consumed, regardless of outlet or platform used.

The literature review sought validation regarding which scientific areas are employed in the context of the investigation carried out that explore the same subject matter, to verify which types of studies, metrics and data are utilized. It will also consist of an important brainstorming exercise for the formulation of the research hypotheses as well as taking note of what conclusions were drawn from previous investigative works.

The questionnaire will be distributed mainly resorting to the collaborators' internal network of the student's current employment entity. Additionally, social media presence was leveraged, even if minor, to reach the student's network of people.

The data produced was loaded into Jupyter Notebook, a computing platform/text editor where the various queries were written, and results gathered. The Numpy, Pandas and Scikit-learn libraries will be the foundation for the queries, based on the object-oriented, Python programming language. This project's main development portion will consist of data normalization and processing, followed by an explorative analysis and extrapolation, culminating in predictions based on the relationships identified between the key variables, achieved through predictive modelling.

Keywords: Consumption, Television, Film, Predictive, Modelling, Python.
JEL classification system: M20, L82

Índice Geral

Sumário vii

Abstract ix

Índice Geral xi

Índice de Figuras xiii

Índice de Tabelas xiv

Anexos xv

Glossário e Acrónimos xvii

1 Introdução 1

1.1 Descrição da proposta de investigação 1

1.2 Questões de investigação e motivação 2

1.3 Estrutura e proposta metodológica 3

1.3.1 Compreensão da indústria e contextualização da investigação 3

1.3.2 Formulação dos dados 3

1.3.3 Preparação dos dados 3

1.3.4 Análise exploratória e modelação 3

1.3.5 Avaliação 3

1.3.6 Implementação/*Deployment* 3

1.4 Projeções e potenciais ilações a retirar 4

2 Revisão da literatura 5

2.1 Relevância 5

2.2 Definição da *query* e critérios de pesquisa 5

2.3 Revisão do título e *abstract* pós-filtragem 7

2.4 Contextualização 8

2.5 Principais factos apurados 9

2.6 As motivações dos consumidores 10

2.7 O panorama de consumo atual 11

2.8 O impacto e disrupção dos *Over-The-Top* 13

2.9 Relevância dos *Over-The-Top* e implicações de investigação 14

2.10 Métodos empíricos e modelação analítica dos artigos revistos 16
2.11 Avaliação da qualidade dos artigos revistos 16
3 Metodologia 19
3.1 Limitações, lapsos e *input* recebido 20
3.2 Carregamento e tratamento dos dados 20
3.2.1 Normalização da base de dados 21
3.2.2 Complementos à base de dados 22
3.3 Processamento e apresentação de resultados 23
4 Resultados e ilações 25
4.1 Sumário estatístico 25
4.2 Caracterização do perfil da amostra inquirida 26
4.3 Análise comparativa das modalidades de consumo de conteúdo de entretenimento 27
4.4 Distribuição dos valores para *features* com múltipla seleção 29
4.5 Amplitude e distribuição de valores para os principais índices de análise 32
4.6 Modelação preditiva 34
4.6.1 Regressão linear 37
4.6.2 Diagnóstico da multicolinearidade e homogeneidade 40
4.6.3 Árvores de decisão 42
4.6.3.1 Modelo 1: Satisfação com a acessibilidade e/ou preço pago 43
4.6.3.2 Modelo 2: Qualidade percecionada 43
4.6.3.3 Modelo 3: Satisfação global em função dos preditores selecionados 44
4.6.3.4 Modelo 4: Satisfação global em função das outras variáveis-alvo 45
4.6.4 *Random forest* 46
4.6.5 *Support vector machine* 46
4.7 Resultados relevantes dos artigos revistos e validação das intenções de estudo 47
5 Conclusões 49
5.1 Contributos 50
5.2 Limitações e futuras pistas de investigação 51
Referências bibliográficas 53
Anexos 55

Índice de Figuras

Figura 1: Palavras-chave a utilizar na definição da *query* 6
Figura 2: Diagrama da filtragem realizada e respectivos resultados 6
Figura 3: Representatividade de género.....26
Figuras 4: Características da amostra inquirida.....27
Figura 5: Modalidade de consumo predominante 27
Figura 6: Frequência dos diferentes critérios 30
Figura 7: Frequência dos diferentes géneros.....30
Figura 8: Frequência dos serviços subscritos.....31
Figura 9: Frequência das razões motivadoras para subscrever serviços de subscrição 31
Figura 10: Frequência dos modos de descoberta.....31
Figura 11: Satisfação com a acessibilidade e/ou preço pago 32
Figura 12: Qualidade percecionada 33
Figura 13: Satisfação global.....33
Figura 14: *Heatmap* para os rácios de correlação.....34
Figuras 15: Satisfação global 35
Figuras 16: Satisfação com a acessibilidade/preço pago e qualidade percecionada.....36
Figuras 17: Satisfação global em função dos dois outros índices-alvo da análise 36

Índice de Tabelas

Tabela 1: Principais questões de investigação formuladas....2
Tabela 2: Objectivos e proposta de concretização....2
Tabela 3: Questões que pautam a revisão da literatura....5
Tabela 4: Critérios que ditam os artigos a incluir e excluir na revisão da literatura....6
Tabela 5: Critérios de qualidade....8
Tabela 6: Renomeamento e resumo funcional das colunas da base de dados....22
Tabela 7: Mapeamento das novas *features* numéricas a adicionar....23
Tabela 8: Principais métricas das *features* numéricas....25
Tabela 9: Apresentação das métricas calculadas – confronto das modalidades de consumo....28
Tabela 10: Mapeamento das relações lógicas entre as variáveis....37
Tabela 11: Resultados da aplicação dos modelos de regressão linear simples....38
Tabela 12: Resultados da aplicação dos modelos de regressão linear múltipla....39
Tabela 13: Apresentação dos resultados do diagnóstico....41
Tabela 14: Resultado dos cálculos de regressão para modelos de árvores de decisão....42
Tabela 15: Principais métricas para o modelo 1....43
Tabela 16: Principais métricas para o modelo 2....44
Tabela 17: Principais métricas para o modelo 3....44
Tabela 18: Principais métricas para o modelo 4....45
Tabela 19: Principais métricas para os diferentes modelos através de *random forest*....46
Tabela 20: Principais métricas para os diferentes modelos através de *support vector machine*....47

Anexos

A - Comandos introduzidos no *jupyter notebook* 55

Comando 1: Importação das bibliotecas basilares necessárias para a prossecução das tarefas 55

Comando 2: Adição de opções para expandir a apresentação dos dados nas células de *output* 55

Comando 3: Carregamento do *dataset*-objecto de estudo 55

Comando 4: Listagem das colunas originais para efeitos de contextualização 55

Comando 5: Contagem do número de registos 55

Comando 6: Eliminação de colunas excendentárias que não acrescentam valor à análise 55

Comando 7: Renomeamento das colunas consoante a listagem original para facilitar a leitura 55

Comando 8: Listagem do número de registos nulos e identificação da tipologia de dados 55

Comando 9: Frequência da opção "Nenhum/Terminar *Survey*" selecionada na Q4 55

Comando 10: Eliminação das respectivas entradas identificadas no comando 9 55

Comando 11: Frequência das opções "Televisão por cabo" ou "Cinema" selecionadas na Q12 55

Comando 12: Substituição dos valores nulos 56

Comando 13: Conversão da tipologia das últimas 3 colunas 56

Comando 14: Adição de variáveis dicotómicas 56

Comando 15: Adição de novas colunas e conversão da tipologia de dados 56

Comando 16: Reordenação lógica das colunas 56

Comando 17: Sumário estatístico das variáveis numéricas 56

Comando 18: Importação das bibliotecas necessárias para a análise de dados exploratórios 56

Comando 19: Construção de *pie-charts* para caracterização da amostra 56

Comando 20: Cálculo das métricas a apresentar na análise de dados exploratórios 56

Comando 21: Representação gráfica em barras para *features* de múltipla seleção e *boxplot* 56

Comando 22: Construção do mapa de verificação de relações entre as *features* numéricas 57

Comando 23: Construção dos gráficos *lineplot* 57

Comando 24: Teste de adequação e diagnósticos 57

Comando 25: Importação das bibliotecas afetas à regressão linear 57

Comando 26: Construção dos modelos de regressão linear e respectiva apresentação 57

Comando 27: Cálculo das principais métricas dos modelos preditivos 57

Comando 28: Separação da amostra em treino e teste para os modelos preditivos 57

Comando 29: Instalação das ferramentas que possibilitam a construção de árvores de decisão 58

Comando 30: Importação das bibliotecas afetas aos algoritmos de *machine learning* 58

Comando 31: Definição das variáveis a utilizar e separação nas amostras treino e teste 58

Comando 32: Definição da profundidade das árvore de decisão e chamada das variáveis 58

Comando 33: Construção da matrix de confusão, relatório de valores e representação gráfica 58

Comando 34: Construção da árvore 58

B - Fórmulas 59

Fórmula 1: Precisão 59

Fórmula 5: Sensibilidade 59

Fórmula 3: *F1 score* 59

Fórmula 4: PECC 59

Fórmula 6: Especificidade 59

C – Artigos científicos revistos 60

D – Apresentação do questionário 62

E – Apresentação das árvores de decisão 64

Árvore de decisão para o modelo 1 – Profundidade 5 64

Árvore de decisão para o modelo 2 – Profundidade 4 65

Árvore de decisão para o modelo 3 – Profundidade 6 66

Árvore de decisão para o modelo 4 – Profundidade 4 67

F – Cotações atribuídas aos artigos revistos em função dos critérios e questões 68

G – Outros suportes e consultas 69

Matriz de confusão – *Scikit-learn* 69

Reparos e melhorias – *Feedback* questinário 69

Nível de interesse – Participantes do questionário 69

Glossário e Acrónimos

AD – Árvores de Decisão

ADE – Análise de dados exploratórios

API – Application Programming Interface – Meio de comunicação entre componentes computacionais

CART – Classification and Regression Tree – Árvore de Classificação e Regressão

Checklist – Lista que possibilita múltipla seleção

CRISP-DM – Cross-Industry Standard Process for Data Mining

Decision tree pruning – Técnica de compressão que remove secções não críticas e/ou redundantes

DF – Dataframe: Estrutura de dados organizada bidimensionalmente

DMM – Digital Media Marketing

DTO – Download-To-Own – Descarregamento do conteúdo usufruído

Erro absoluto médio – Média dos erros absolutos entre o previsto e o observado

Erro quadrático médio – Média dos erros absolutos ao quadrado entre o previsto e o observado

KNN – K-Nearest Neighbor

LR – Regressão Linear

OTT – Over-The-Top – Serviço que disponibiliza o produto diretamente ao consumidor via Internet

PECC – Percentagem de exemplos corretamente classificados

PPV – Pay-Per-View – Pagamento por um tipo de conteúdo singular

RF – Random Forest – Baseado na construção de uma pluralidade de árvores

RL – Revisão da Literatura

Rule of thumb – Regra convencionada e encorajada

$\mathbf{R^2}$ – Coeficiente de Determinação – Mede o ajuste do modelo à variável dependente

Statement – Unidade sintática de uma linguagem programática imperativa

Streaming – Método de transmissão através de uma rede computacional

SVM – Support Vector Machine

Tastemaker – Curador de conteúdo

TS – Topic Searches For: Pesquisa agrupada por tópico nos diretórios utilizados na RL

VI – Variáveis independentes

VIF – Variance Inflation Factor – Medida de deteção da multicolinearidade entre as Vis

VOD – Video-on-Demand

1 Introdução

O panorama do consumo de entretenimento televisivo e cinematográfico é cada vez mais dominado por serviços de subscrição que distribuem e oferecem conteúdo da mesma natureza, sendo atualmente, substitutos prolíficos ao consumo através de métodos de distribuição mais convencionais e com maior maturidade. No capítulo introdutório, é descrito o propósito da investigação, o modus operandis para alcançar os objectivos propostos e as potenciais ilações a retirar após determinação e apresentação de resultados.

1.1 Descrição da proposta de investigação

A presente investigação visa analisar os padrões de consumo de conteúdo televisivo e cinematográfico, com destaque para serviços de subscrição das plataformas que disponibilizam conteúdo desta natureza, confrontando com outros métodos, digitais ou físicos, em que o utilizador geralmente paga mediante o que pretende consumir, no caso de uma sessão de cinema ou através de pacotes, no caso da televisão por cabo.

Pretende-se perceber a adequação e impacto destes serviços e outros fatores adjacentes, de modo a determinar o grau de satisfação e qualidade percecionada para os consumidores, com base nas suas características e hábitos de consumo. Tendo em consideração a discrepância de maturidade entre as diversas modalidades de consumo, pretende-se determinar se as ofertas existentes vão de facto ao encontro das expectativas dos utilizadores.

A partir de uma amostra de consumidores, com interesses e preferências específicos, serão apontadas as particularidades dos seus hábitos de consumo e as razões inerentes das escolhas que fazem para interagir com as diferentes plataformas de entretenimento.

Para o efeito, efetuar-se-á uma análise exploratória destes padrões de consumo, procurando obter atributos explicativos destes hábitos de forma a inferir sobre a modalidade de consumo que será consensualmente mais vantajosa para os consumidores quer em termos monetários quer a nível da qualidade dos conteúdos oferecidos. Para além das relações que se vão estabelecer entre as diversas variáveis, foram construídos modelos preditivos que estimam a qualidade percecionada e a satisfação dos utilizadores.

A informação das duas diferentes indústrias de entretenimento e das duas modalidades de consumo, a identificar seguidamente, será confrontada de modo a determinar as métricas mais relevantes para cada uma. Aliado a isto, pretende-se perceber o nível de preferência que os utilizadores ainda têm por ver filmes no cinema físico ou por assistir a séries televisivas através de pacotes de televisão por cabo, contornado assim os serviços de subscrição vigentes.

Tirando partido da biblioteca de artigos científicos que exploraram este tema, a investigação tentou alicerçar-se nos vários estudos já levados a cabos, técnicas utilizadas e se possível, tipos de dados e

métricas empregues, que possibilitem a identificação de relações causais significativas entre as variáveis a definir posteriormente.

1.2 Questões de investigação e motivação

O principal elemento indutor desta exploração parte de um entendimento de que, apesar de a oferta ser vasta, é por vezes difícil reconhecer qual é a melhor relação preço-valor entre as diversas ofertas existentes ao nível do entretenimento e qual é a mais apropriada para um tipo de utilizador. Adicionalmente, se algum dos seus hábitos pode ter mais ou menos influência no sua perceção do conteúdo distribuído atualmente. Este é o *gap* de investigação que identifica presentemente, depois de efetuada uma pesquisa preliminar com incidência neste tema.

O conhecimento potencialmente extraído da investigação terá a máxima utilidade para produtores/distribuidores/criadores de conteúdo na medida em que ajudará a perceber as tendências dos consumidores, quais os aspectos a que atribuem mais valor quando utilizam as respectivas plataformas e que modalidade de distribuição permite melhor disseminar os conteúdos disponbilizados da forma pretendida. Para os consumidores, pode ser ilustrativo da influência dos seus hábitos de consumo na sua satisfação e de que modo podem mudar a forma como interagem com o conteúdo do qual usufruem.

Tabela 1: Principais questões de investigação formuladas

ID	Questão
A	Que tipo de consumidores preferem os novos serviços de subscrição, em detrimento das modalidades de consumo mais convencionais?
B	Que projecção é possível fazer através das caracterísicas de consumo dos utilizadores?
C	Como tiram os consumidores partido do conteúdo com o qual interagem?
D	Qual o grau de adequação do conteúdo oferecido consoante satisfação e qualidade percecionada?

Tabela 2: Objectivos e proposta de concretização

Objectivo	**Forma de validação**
Contextualização relativamente a trabalhos de investigação que exploram esta temática, no contexto da área científica do presente estudo	Revisão da literatura
Caracterização dos consumidores de conteúdo de entretenimento	Dados sociodemográficos resultantes do questionário produzido
Determinação da modalidade de consumo mais prevalente e desconstrução dos hábitos de consumo que se interpretam como mais relevantes	Dados de consumo/preferências resultantes do questionário produzido
Previsão da satisfação e da qualidade percecionada de forma a retirar ilações relativamente às caraterísticas de consumo e ao efeito disruptivo dos novos serviços de subscrição nas indústrias televisiva e cinematográfica	Dados resultantes da aplicação de modelação preditiva sobre os dados compilados e consultados.

1.3 Estrutura e proposta metodológica

Esta dissertação tem a máxima proximidade com a doutrina *CRISP-DM* na medida em que engloba o mesmo número de fases, sendo estas enquadradas na intenção de investigação, ou seja, ajustadas de forma a melhor refletir as etapas que compõem o trabalho desenvolvido e descrito nas secções subsequentes.

1.3.1 Compreensão da indústria e contextualização da investigação

Trata-se do enquadramento do tema em estudo, através da definição dos conceitos mais importantes assim como a contextualização de ambas as indústrias-objeto de estudo. Aliado a isto, identificar que tipo de investigações são realizadas neste âmbito, aquando da realização da revisão da literatura.

1.3.2 Formulação dos dados

Realização e distribuição do questionário, formulado com vista a gerar dados da forma pretendida, simultaneamente tornando a experiência de preenchimento o mais ágil possível. Os dados extraídos serão compilados e posteriormente carregados na ferramenta de edição onde a análise será realizada.

1.3.3 Preparação dos dados

Após carregamento, os dados serão tratados de forma a servirem o propósito de investigação e de maneira a que possam ser integrados, tanto na análise exploratória como na estimação de modelos preditivos. Os comandos utilizados são apresentados neste documento, tendo sido remetidos para o seu anexo.

1.3.4 Análise exploratória e modelação

O tratamento dos dados irá permitir normalizar a base de dados para que posteriormente se possam estabelecer relações entre as diferentes variáveis e realizar outras análises como, por exemplo, caraterizar o perfil dos utilizadores inquiridos. Esta fase precede a construção de modelos preditivos, consistindo em análises preliminares e respectivas conclusões consoante os sub-tópicos de exploração.

1.3.5 Avaliação

A aplicação dos modelos preditivos irá ser sucedida pelo confronto dos seus resultados, determinando se a relação existente entre as variáveis é significativa e se os diferentes tipos de modelação produziram resultados com materialidade e quais as discrepâncias verificadas.

1.3.6 Implementação/*Deployment*

Fase suplementar que sucede o produto tangível, a investigação, composta pelas anteriores fases. Corresponde à entrega/submissão do trabalho após a sua conclusão.

1.4 Projeções e potenciais ilações a retirar

A investigação propõe-se a perceber se os consumidores estão satisfeitos com a proposta de valor oferecida e qualidade do conteúdo do qual usufruem e quais as razões motivadoras inerentes para optar por uma modalidade de consumo em detrimento de outra.

Adicionalmente, a investigação propõe-se a enumerar as características de consumo dos utilizadores que podem influenciar diretamente a sua perceção. Pretende-se ainda averiguar de que forma os utilizadores estão a capitalizar os conteúdos oferecidos, nomeadamente, a nível da quantidade e diversidade consumida. Consequentemente, procura-se determinar a forma como os novos serviços disponíveis estão a influenciar estes hábitos de consumo e se o consumo de entretenimento através de métodos mais convencionais é ainda conservado.

Em suma, pretende-se concluir se é possível estabelecer a relação entre as características de consumo e a perceção dos consumidores, e se alguma destas é um forte preditor que permita projectar categoricamente um dos índices principais de análise, a saber, a satisfação e a qualidade percecionada dos consumidores de conteúdo televisivo e cinematográfico. Esta intenção é o mais próximo de um *gap* ao nível da bibliografia consultada, que esta investigação apresenta e pretende estudar.

2 Revisão da literatura

A presente revisão sistemática da literatura incidiu sobre a temática do conteúdo de entretenimento, televisivo e cinematográfico, com particular foco para a disrupção dos serviços *OTT* e especial ênfase em análises orientadas para a aplicação de métodos analíticos com base em algoritmos de *machine learning* e especificação de modelos preditivos que permitam estabelecer relações entre as variáveis, determinar potenciais preditores e tirar conclusões. Pretende-se extrair os dados apurados mais relevantes sobre a indústria e serviços, teorias aplicadas que suportam este tipo de análise e métodos utilizados no contexto da componente analítica e ou de modelação das investigações.

2.1 Relevância

Nesta revisão da literatura irá elaborar-se a compilação de informação proveniente de artigos científicos relacionados com o tema descrito. O principal conceito de exploração procurado é o panorama de entretenimento atual com particular incidência nas novas plataformas que são agora ubíquas nas diferentes indústrias e, em última instância, perceber qual o respectivo impacto nos hábitos de consumo dos utilizadores. A intenção primordial foi também a de identificar os tipos de estudo e áreas científicas invocadas neste tipo de investigação, com a expectativa de influenciar a forma como o questionário do presente estudo será formulado, em termos dos sub-tópicos abordados. Por fim, pretende-se perceber se existe algum aspecto que não tenha ainda sido explorado nos artigos consultados, e que possa ser destacado no presente estudo. A *web of science* e a *scopus* foram as fontes de dados ou diretórios utilizados para pesquisa e compilação da informação agregada de seguida.

Tabela 3: Questões que pautam a revisão da literatura

ID	Questão
Q1	Quais os tipos de investigações mais frequentemente conduzidas no âmbito desta temática?
Q2	Que áreas científicas são invocadas neste tipo de investigações?
Q3	Qual a tipologia de dados/métricas/análises empregues?
Q4	Qual o grau de adequação da presente proposta de investigação no contexto dos artigos consultados?

2.2 Definição da *query* e critérios de pesquisa

Esta secção detalha o modo como foi formulada e introduzida a principal *query* de consulta, que pautou a presente RL. São também aqui apresentados os diferentes critérios pelos quais a pesquisa se rege e respectivas implicações para o número de resultados encontrados.

Figura 1: Palavras-chave a utilizar na definição da *query*

Web of science:

(TS=(television) OR TS=(tv) OR TS=(series) OR TS=(shows)) AND (TS=(film) OR TS=(movie*) OR TS=(cinema)) AND (TS=(subscri*) OR TS=(streaming) OR TS=(ott) OR TS=(over-the-top) OR TS=(service*)) AND (TS=(analy*) OR TS=(behavi*) OR TS=(consum*)) AND (TS=(predict*) OR TS=(model*) OR TS=(machine learning) OR TS=(data mining))*

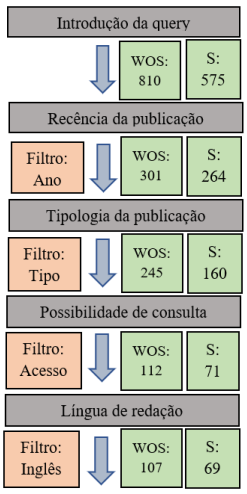
Scopus:

(TITLE-ABS-KEY("television") OR TITLE-ABS-KEY("tv") OR TITLE-ABS-KEY("series") OR TITLE-ABS-KEY("shows")) AND (TITLE-ABS-KEY("film") OR TITLE-ABS-KEY("movie*") OR TITLE-ABS-KEY("cinema")) AND (TITLE-ABS-KEY("subscri*") OR TITLE-ABS-KEY("streaming") OR TITLE-ABS-KEY("ott") OR TITLE-ABS-KEY("over-the-top") OR TITLE-ABS-KEY("service*")) AND (TITLE-ABS-KEY("analy*") OR TITLE-ABS-KEY("behavi*") OR TITLE-ABS-KEY("consum*")) AND (TITLE-ABS-KEY("predict*") OR TITLE-ABS-KEY("model*") OR TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("data mining"))*

Tabela 4: Critérios que ditam os artigos a incluir e excluir na revisão da literatura

Inclusão	Correspondência com a temática
	Realização de uma análise e caraterização das indústrias de entretenimento
Exclusão	Publicação redigida noutra língua que não a Inglesa
	Não passível de ser descarregado e consultado na totalidade
	Antiguidade superior a 5 anos
	Publicação não redigida em forma de artigo
	Título e *abstract* demonstram discordância com a temática proposta

Figura 2: Diagrama da filtragem realizada e respectivos resultados



Os resultados foram ordenados por número de referências/citações noutras publicações como métrica de relevância e qualidade para a presente investigação. Dado a pesquisa indicada não ter produzido o número de resultados pretendidos, foram feitas alguns ajustes e pesquisas paralelas de modo a acrescentar robustez ao portfólio de publicações consultadas. Posteriormente à aplicação da query indicada, e uma vez que esta se encontra organizada por encapsulamentos de palavras que partilham o mesmo tema, separados pelo operador "*AND*", foram feitas pesquisas com um ou mais destes retirados, tendo-se determinado que certos encapsulamentos poderiam ser eliminados sem significativo impacto para o âmbito/temática em estudo. Por conseguinte, foram abrangidas publicações que não dão particular destaque à indústria cinematográfica ou à emergência dos novos serviços de subscrição/*streaming* assim como outras que não se regem por princípios de modelação ou que apliquem algoritmos de machine learning. Adicionalmente, a antiguidade estipulada para a seleção de artigos prende-se com a natureza da temática e com a interpretação de que não será possível extrair informação materialmente relevante ou representativa da atualidade do entretenimento através de publicações com mais de 5 anos.

2.3 Revisão do título e *abstract* pós-filtragem

Para além dos critérios de inclusão/exclusão enumerados, analisou-se individualmente cada publicação, procurando a correspondência com a temática através da leitura/compreensão do respectivo título/*abstract*. Procurou-se, entre outros sub-tópicos, análises que tentam explicar o posicionamento das marcas/serviços nos mercados em que operam assim como os fatores socio-económicos que pautam os utilizadores de uma determinada região. Uma vez que o presente estudo visa decompor o impacto dos hábitos de consumo numa amostra específica e bastante restrita, irão salvaguardar-se os potenciais choques culturais dos mercados analisados nos artigos, no decorrer da RL. O efeito pandémico e a pirataria, não sendo foco do estudo, são tópicos também abordados/invocados ao longo da análise quando assim se justificar. O primeiro, em particular, constitui um fator que até hoje, contribui para a mais acentuada proliferação de novos serviços de entretenimento, confirmando a tendência que já se vinha a verificar mesmo em contexto pré-pandémico. Por último, procurou-se encontrar artigos que realizassem análises no âmbito da presente investigação, de forma a verificar se já existe uma metodologia preferencial para atingir o que é proposto no capítulo introdutório deste trabalho.

Tabela 5: Critérios de qualidade[1]

ID	Critério
C1	Contextualização da indústria televisiva e cinematográfica
C2	Pertinência da investigação e potencial colmatação de lacunas de outros estudos
C3	Explicação da metodologia utilizada no contexto de outras investigações
C4	Justificação dos modelos empregados
C5	Avaliação dos modelos através das métricas conhecidas
C6	Apresentação dos resultados em modo explicativo e não apenas descritivo
C7	Confronto com outras investigações e respectivas contribuições
C8	Apontamento das limitações e outros aspectos de melhoria

2.4 Contextualização

Os serviços de subscrição que dominam o panorama de entretenimento televisivo e cinematográfico constituem uma evolução natural na forma de consumir bens e serviços na medida em que permitem agrupar benefícios de natureza similar e oferecê-los a um preço reduzido, sendo o seu efeito disruptivo inegável. A comodidade conferida por estes, através do nível de personalização e facilidade de acesso, é mais condizente com um estilo de vida tendencialmente mais sedentário, muito promovido pelo efeito pandémico dos últimos anos. No que diz respeito ao conteúdo em si, trata-se de uma inevitabilidade dada a volatilidade das preferências dos utilizadores. Permite o acesso a uma maior diversidade de conteúdos, tendo como único *trade-off* o facto de o consumidor não adquirir uma licença daquilo de que está a usufruir, apenas a aluga pela duração da sua subscrição, perdendo este acesso aquando do seu término. Porém, o utilizador liberta-se das transmissões calendarizadas da programação via televisão por cabo, tendo sempre ao seu dispor, o catálogo de conteúdo oferecido, auferindo um mais elevado grau de flexibilidade. São, atualmente, substitutos viáveis ao conteúdo disseminado através de métodos mais tradicionais e obrigam as respectivas distribuidoras a adaptar as suas estratégias e modelos de negócio.

Entre outras plataformas igualmente prevalentes a nível da sua notoriedade, destacam-se alguns dos serviços mais populares:

- Netflix: Serviço de *streaming* lançado em 2007, dispõe de uma biblioteca de séries televisivas e filmes para "*streaming*", muitos dos quais, criados e produzidos originalmente pela plataforma ou adquiridos "*after the fact*". A grande maioria do conteúdo disponbilizado é resultante de parcerias estabelecidas entre a plataforma e as diversas produtoras que cedem os direitos de transmissão do seu conteúdo.

[1] Valores serão atribuídos a cada critério numa escala de 3: 0/0.5/1.

- Paramount +: Previamente lançado como CBS All Access em 2014, disponibiliza conteúdo de natureza similar ao da Netflix, também produzindo e oferecendo conteúdo original, daquela que é uma das empresas distribuidoras de cinema mais proeminentes de *Hollywood*.

Os dois serviços acima referidos foram destacados pela diferença de contexto em que foram instauradas. A Netflix foi criada com o pretexto de adquirir e distribuir conteúdo televisivo "*third-party*"[2], expandindo posteriormente o seu portfólio através de criação de programas originais. Com a proliferação do serviço, a plataforma expandiu a sua operação e agora produz também filmes, que contam com a participação dos intervenientes mais bem sucedidos da indústria.

A Paramount, por outro lado, percorreu um caminho praticamente inverso, tratando-se de uma das produtoras de conteúdo cinematográfico mais prolíficas, produzindo agora séries televisivas originais que auferem robustez ao conteúdo que disponibiliza. Apesar das disparidades de circunstâncias do seu surgimento, ambos os serviços competem agora pela mesma audiência, mercado e segmento de utilizadores.

A possibilidade de transmissão de conteúdo através de "*streaming*" é o elo em comum entre estas plataformas e quaisquer outras concorrentes emergentes dentro desta modalidade de consumo. Esta tecnologia possibilita a transferência e partilha de conteúdo audiovisual digitalmente, sem necessidade de presença local do lado do consumidor, sendo a experiência apenas afetada pela qualidade da respectiva ligação à Internet. Contudo, não se pretende inferir àcerca da satisfação do consumidor com a tecnologia em si mas sim relativamente à modalidade de serviços que agrupam e disponibilizam o conteúdo desta forma em detrimento de pagar consoante o que pretende usufruir, de forma singular, e.g. um jogo de futebol, um combate de boxe, uma corrida de fórmula 1 ou um filme no cinema.

2.5 Principais factos apurados

As organizações *media* encontram-se num processo de metamorfose para estabelecer plataformas digitais de forma a distribuir o seu conteúdo (Doyle, 2015, citado por Alcolea-Díaz et al., 2021). A televisão está agora mais afastada dos conceitos de canal e linearidade de transmissão como elementos que definem o conteúdo (Askwith, 2007, citado por Alcolea-Díaz et al., 2021). Este processo é alavancado pela disponibilidade de TVs *Smart*, que enfatizam a utilização de serviços *OTT*, colocando-os numa posição de destaque na sua interface, seja qual for o sistema operativo (Comscore, 2021, citado por Alcolea-Díaz et al., 2021). Em circunstâncias em que este tipo de competição emerge, os serviços locais de televisão por cabo têm de ter um portfólio abrangente de conteúdo, que implica maior investimento em produção e *marketing* (Alcolea-Díaz et al., 2021).

O entretenimento distribuído de forma *VOD*, globalmente, tem uma receita projetada de 87.1 biliões de dólares em 2024 (Markets and Markets, 2020, citado por Jang et al., 2021) e este valor pode ser

[2] Disponibilizado originalmente por uma distribuidora externa mas por intermédio da celebração de uma parceria entre esta e a plataforma que oferece o serviço *OTT*, a respectiva licença é cedida por um tempo limitado

desconstruído em várias sub-modalidades: gratuita, com anúncios embutidos, através de subscrições e transacional (Abreu et al., 2017, citado por Jang et al., 2021). A transacional segrega-se em *PPV* e *DTO*, não constando no âmbito da presente investigação, uma vez que os serviços *OTT*, foco do estudo, não tiram partido destas modalidades de consumo/negócio. Contudo, é importante destacar a proeminência do *PPV* no contexto de eventos desportivos e o *DTO* para a compra e *download* de filmes que o consumidor pretende deter (Jang et al., 2021). Do total da receita proveniente de serviços *VOD*, estima-se que 40% da *market share* seja detida apenas por 4 das plataformas vigentes: Netflix, Hulu (Grupo Disney), Youtube e Amazon Prime (Lee et al., 2021).

Com a normalização das tecnologias digitais entre a população, os serviços *OTT* tornam-se uma inevitabilidade (Prokopenko et al., 2019, citado por Alforova et al., 2021). A sua missão passa por oferecer conteúdo de elevada qualidade utilizando tecnologia acessível à maioria dos consumidores (Alforova et al., 2021). Porém, existe um *trade-off* politicamente paradoxal que advém de objectivos discordantes destes serviços, entre o que é mais vantajoso economicamente ou irá gerar mais receita e em termos de poupança para o consumidor (Gaustad, 2019, citado por Alforova et al., 2021). Ainda assim, o impacto registado traduz-se numa predisposição para pagar um "*premium*", sendo apontado em alguns estudos que 30% dos utilizadores escolhem serviços consoante a sua infraestrutura/rede 5G, capaz de providenciar melhor qualidade de vídeo e menos instâncias de "*buffering*"[3] (Choi & Kim, 2021).

A Federal Communications Comission, a entidade reguladora americana define um serviço *OTT* como um fornecedor *online* de conteúdo audiovisual através da *Internet* (Bury & Li, 2013, citado por Chen, 2019). A proliferação destes serviços pode ser decomposta em três fases desde 2000 (Steinkamp, 2010, citado por Chen, 2019):

1. Inclusão da *Internet* em campanhas promocionais para o aliciamento de utilizadores da *Internet* que vêem programas televisivos;
2. Programas curtos de baixo orçamento são produzidos mas os serviços *OTT* continuam a ter um papel de suporte;
3. *Internet* estabelece-se como um hub de conteúdos com o mesmo nível de produção e conquistando a sua própria audência oferecendo um maior nível de customização (Chen, 2019). Enquanto que a sua utilização era predominante nas audiências mais jovens dado o maior grau de aceitação para os elementos tecnológicos que caracterizam estes serviços, a abrangência de utilização está atualmente mais generalizada (Nijhawan & Dahiya, 2020).

2.6 As motivações dos consumidores

São formulados dois pontos de vista para medir a forma como uma modalidade de consumo se pode sobrepôr à outra: deslocamento simétrico, pela comparação do tempo que os utilizadores passam a

[3] Falha de rede que provoca interrupções aquando do consumo de conteúdo

consumir conteúdo de determinado formato, modalidade ou serviço; deslocamento funcional, pela quantificação da satisfação dos consumidores com objectivos específicos, obtidos através do conteúdo que selecionam (Greer & Ferguson, 2015, citado por Chen, 2019). Existe o argumento de que um tipo de *media* nunca é totalmente substituído mas apenas o é condicionalmente, ou seja, um utilizador escolhe determinado serviço em detrimento de outro de forma a satisfazer uma necessidade específica. Porém, raramente é o caso em que um consumidor tem uma necessidade singular e por isso, é forçado a recorrer a um conjunto de *media*/fontes de entretenimento (Newell et al., 2008, citado por Chen, 2019). Baseado na teoria de nicho (Dimmick & Rothenbuhler, 1984, citado por Chen, 2019), são enumeradas sete dimensões de gratificação dos utilizadores: informação, relaxamento, apreciação, interação social, benefício financeiro, facilidade de utilização e conveniência (Steinkamp, 2010, citado por Chen, 2019).

A motivação do espectador, juntamente com o fator da recência e imprevisibilidade, constituem os fatores mais importantes na produção de conteúdo cinematográfico e televisivo, nomeadamente na distribuição e alocação de recursos cognitivos (Pisarek & Zabielska-Mendyk, 2021). No que diz respeito a estes tipos de *media*, foram formuladas classificações para as necessidades dos utilizadores quando interagem com este tipo de conteúdo: 1) Necessidade de informação; 2) Confirmação do seu sistema de valores; 3) Integração e promoção de interação social; 4) Fonte de entretenimento (McQuail, 2008, citado por Pisarek & Zabielska-Mendyk, 2021).

2.7 O panorama de consumo atual

Para as entidades distribuidoras de conteúdos de entretenimento, existe uma evidente necessidade de acompanhar as modalidades de consumo mais populares, atualmente oferecidas por plataformas *OTT* como o Netflix, que se tornou o serviço *OTT* dominante em 2018, com 125 milhões de subscritores a nivel mundial (Spangler, 2017, citado por Herbert et al., 2019). É também importante salientar que a receita associada a estes serviços de subscrição ultrapassou, em 2016, as vendas físicas de conteúdo de entretenimento, por 0.8 biliões de dólares. Tanto os lançamentos em cinemas físicos para filmes como as transmissões "lineares"/por cabo para séries televisivas mantêm a sua proeminência, porém, é possível verificar o impacto disruptivo que os *OTT*s tiveram nas indústrias, quer a nível de produção como de distribuição. Por exemplo, a visualização de filmes no cinema físico preserva a componente de interação social, comprometendo a agilidade auferida por um serviço *OTT*, que cria um ambiente personalizado em que o utilizador detém todo o controlo e possibilita avançar, pausar ou repetir deteminada cena (Suwarto et al., 2021). Da necessidade apontada anteriormente, surgiram algumas celebrações de parcerias como é o caso da Netflix-Comcast ou a Netflix-Europe's Sky, mas também se verificou a redução do preço dos serviços oferecidos pelas distribuidoras de entretenimento de forma mais convencional como a Foxtel. Adicionalmente, é possível verificar a emergência de outros serviços concorrentes pertencentes a entidades com um nível de maturidade superior, como é o exemplo da HBO Max. No que diz respeito a tendências de consumo, diferentes fontes oficiosas registam uma diminuição da pirataria aquando da entrada dos serviços de subscrição *OTT*. Esta informação vai ao encontro de

uma afirmação feita pelo *Chief Content Officer* da Netflix relativamente à diminuição do tráfego do *BitTorrent*[4] à medida que o da Netflix cresce num determinado mercado/território. Esta constatação é também corroborada por estudos conduzidos pelo *Intellectual Property Office* no Reino Unido e o *Intellectual Property Awareness Foundation* na Austrália, ainda que o nível de proteção destes dados comerciais seja, na maior parte das instâncias de investigação, significativo, o que nem sempre possibilita a respectiva agregação de dados (McKenzie et al., 2019). Existem ainda bastantes perspectivas discordantes e alguns autores alegam, que no contexto global, a pirataria terá aumentado.

Em várias instâncias, já foi determinado o impacto negativo da pirataria nas interações legítimas por intermédio de venda ou subscrição e inerente predisposição a pagar por tais serviços quando o conteúdo está disponível de forma gratuita por meios ilegítimos (Bai & Waldfogel, 2012, citado por McKenzie et al., 2019). Da mesma forma, também já foi possível determinar que a proliferação da *Internet* resulta numa diminuição do tempo dispendido a consumir formas mais tradicionais de televisão (Liebowitz & Zentner, 2016, citado por McKenzie et al., 2019), facto que se torna mais prevalente para a população mais jovem e com rendimentos inferiores (Prince & Greenstein, 2017, citado por McKenzie et al., 2019). Consta que 82% do tráfego seja gerado por serviços de subscrição *OTT*, ainda que haja tendência de oscilação deste valor para certos mercados, como é o caso da Índia, dado o choque cultural e prevalência do consumo de conteúdo nacional/local (Nagaraj et al., 2021). Adicionalmente, foi constatado que a elasticidade de preço da procura é maior para filmes vistos no cinema, sendo que para certos segmentos de consumidores, para filmes de maior notoriedade e relevo, a sua sensibilidade de preço é menor (de Roos & McKenzie, 2014, citado por McKenzie et al., 2019). Por outro lado, para os utilizadores de serviços de subscrição, existe uma maior propensão para ver filmes "nicho", "independentes" e/ou mais antigos dada a menor fricção em pesquisá-los, quando presentes no catálogo (Hiller & King and King, 2017, citado por McKenzie et al., 2019). Pode verificar-se que estes serviços permitem a melhor disseminação de títulos com menor *"marketing push"*[5], originários de outros países que não os USA e sem o mesmo alcance que os produzidos pelas principais distribuidoras da indústria (Aguiar & Waldfogel, 2018, citado por McKenzie et al., 2019).

A digitalização da qual os serviços *OTT* tiram proveito reduz o custo fixo de criação de conteúdo (Waldfogel, 2012, citado por Matos et al., 2017) assim como os custos marginais de distribuição (Varian, 2005, citado por Matos et al., 2017). No capítulo da pirataria, estima-se que o usufruto de conteúdo por este meio constitua um terço do excedente do consumidor, ganho à custa dos respectivos produtores e distribuidores (Rob and Waldfogel, 2006, citado por Matos et al., 2017). Estudos demonstram que os esforços anti-pirataria podem aumentar o consumo legal de conteúdo, pelo menos no curto prazo (Danaher et al., 2016, citado por Matos et al., 2017), contudo, nenhuma relação direta foi consensualmente estabelecida entre o consumo via serviços *streaming OTT* e meios ilegítimos (Matos et al., 2017).

[4] *Software*/Plataforma que possibilita o *download* de conteúdo não sancionado protegido por direitos de autor
[5] Esforço a nível de *marketing* para promover determinado conteúdo

2.8 O impacto e disrupção dos *Over-The-Top*

A acessibilidade de interação com o conteúdo oferecido pelos serviços *OTT* revela-se um dos fatores mais preponderantes para o utilizador e principal razão pelos metódos mais convencionais de distribuição de conteúdo televisivo e cinematográfico serem preteridos (Sonnac, 2012, citado por Medina et al., 2019). A facilidade de utilização, a ubiquidade e a poupança de tempo crê-se serem atributos fundamentais para a aceitação e integração de dispositivos *Internet of Things* no contexto do entretenimento (Touzani, 2017, citado por Nagaraj et al., 2021).

Estudos apontam para a reticência de alguns consumidores em subscrever serviços OTT dada a tendência de aumento de comportamentos disfuncionais despoletados por "*binge-watching*", o ato de ver vários episódios de uma série televisiva consecutivamente, sem interrupções (Nagaraj et al., 2021). No âmbito da psicologia, acredita-se que predisposições positivas e negativas quanto à tecnologia co-existem simultaneamente de acordo com os Modelos de Adoção Tecnológica, *TAM*s (Khatri, 2018, citado por Nagaraj et al., 2021). Estes modelos compreendem certos vetores de análise que determinam as motivações para interagir com certa tecnologia: 1) facilidade de utilização percecionada; 2) utilidade prática percecionada; 3) motivação ritualizada para a utilização; 4) motivação instrumentalizada para a utilização; 5) intenção para utilizar determinada tecnologia (Scherer et al., 2019, citado por Camilleri & Falzon, 2020).

Existe uma mudança estrutural e sistémica provocada pelos serviços *OTT* que poderá significar a convergência dos *media*, telecomunicações e entretenimento numa só plataforma (Gimpel, 2015, citado por Medina et al., 2019). Alguns autores alegam poder haver possibilidade de coexistência entre esta e as modalidades de consumo mais convencionais que permitem maximizar a satisfação e capitalizar o tempo de consumo de determinado tipo de conteúdo (Kim, 2016, citado por Medina et al., 2019). Foram atribuídos termos específicos à progressiva transição da televisão por cabo e cinema físico para os serviços OTT, sendo que "*cord-cutting*" significa substituição total e "*cord-shaving*", a relegação para segundo plano (Lee & Lee, 2015, citado por Medina et al., 2019).

O elevado tráfego e atividade na *Internet* registados atualmente implica que existe uma sobrecarga de informação à disposição dos utilizadores, no contexto do entretenimento. A emergência de sistemas de recomendação é um sintoma deste facto e assenta numa abordagem de alavancagem de *big data*. A experiência de um utilizador com um serviço *OTT* influencia o motor de recomendação, que por sua vez é influenciado pelo utilizador através do conteúdo que seleciona, estabelecendo um ciclo. Estes sistemas, apresentando consistentemente conteúdo do interesse do subscritor, contribuem para a diminuição das taxas de cancelamento das subscrições. A grande parte dos trabalhos de investigação neste âmbito utilizam métodos *query-driven* para estabelecer um algoritmo de recomendação. Existem 3 abordagens que alicerçam estes sistemas: "*Collaborative Filtering*", baseada nas preferências de utilizadores com perfis similares por via de aplicação de algoritmos de *machine learning*, "*Content-based filtering*", baseada na atividade do próprio utilizador e "*Demographic filtering*", um misto das duas anteriores

(Awan et al., 2021). A dependência destes sistemas, pode resultar numa "bolha de filtragem" para o utilizador, uma vez que fica dependente do *input* da própria plataforma que disponibiliza o serviço, tornando-se isento de escolha e pautando-se apenas pelo que lhe é sugerido. Idealmente, o utilizador precisará sempre de exposição a opiniões externas assim como de aplicar o seu sentido crítico para decidir o que ver a seguir (Gutzeit et al., 2021).

Os hábitos de consumo e preferências dos utilizadores de serviços *OTT* servem de alicerce para a instauração de algoritmos de recomendação, que permitem chegar ao conteúdo pretendido de forma mais despreocupada. Os dados de que dispõem estes serviços possibilitam a previsão da classificação que um utilizador vai dar a certo filme ou série televisiva. No que diz respeito a *data mining*, estudos indicam que os dados gerados por estes sistemas podem ser obtidos através de *API*s "*open*" ainda que por vezes as permissões sejam bastante restritivas a nível da quantidade acessível destes dados (Duan & Gao, 2021).

2.9 Relevância dos *Over-The-Top* e implicações de investigação

Crê-se que a proliferação de serviços *OTT*, para além de intrinsecamente ligada ao crescente número de conexões via *Internet*, deve-se também ao melhoramento da infraestrutura de redes, avanços tecnológicos e a maior acessibilidade através de dispositivos inteligentes, como *smartphones* ou *smart TVs*. Aliado a isto, está uma mudança de paradigma relativamente aos utilizadores preterirem serem proprietários do conteúdo que consumem em prol da acessibilidade e agilidade auferida por estes novos serviços, disponibilizando bibliotecas de conteúdo em qualquer lugar. Outros *drivers* podem ser enumerados neste âmbito, nomeadamente, a variedade de conteúdo ou o modo de apresentação desse mesmo conteúdo. Adicionalmente, a interatividade entre marca e consumidor são componentes-chave para assistir os respetivos departamentos de *marketing* em assegurar a retenção dos utilizadores e preservar relações transacionais de longo prazo (Habib et al., 2022). As práticas *DMM* ganham relevo neste contexto uma vez que contribuem para uma imagem de marca mais bem sucedida e influenciam diretamente a intenção de compra/subscrição. Entre as razões que fomentam a interação entre consumidor e entidades, ao nível das redes sociais, destacam-se a satisfação, reconhecimento da marca, grau de acessibilidade, qualidade do produto oferecido, conhecimento desse produto e promoções/campanhas realizadas (Rhom et al., 2013, citado por Habib et al., 2022).

As investigações desta natureza não expandem normalmente a análise a múltiplas indústrias de entretenimento dadas as diferenças ao nível da apresentação do conteúdo aos consumidores ainda que por vezes possam ser traçados paralelismos. Estes tipo de estudos são tão ou mais pertinentes quanto os de comparação relativamente a zona geográfica ou choque cultural, conferindo uma visão mais alargada do panorama do entretenimento, percebendo também as nuances que separam as indústrias e razões inerentes. Ainda que exista sobreposição, tanto relativamente à natureza do conteúdo como por ser oferecido pelos mesmos serviços, as indústrias televisiva e cinematográfica são ainda reconhecidas como independentes por entidades como a Federal Communications Comission (Herbert et al., 2019).

Crê-se que o consumo tem impacto na mediação de exercícios de memória relativamente a eventos históricos e entendimento da atualidade mundial (Gambarato et al., 2021), facto ainda mais proeminente em contexto pós-pandémico (Bacon, 2020, citado por Gambarato et al., 2021). Apesar do colapso de memória destes eventos ser associada ao impacto da televisão (Hoskins, 2004, citado por Gambarato et al., 2021), acredita-se que a retrospecção e a nostalgia são dimensões importantes para os serviços *media OTT* tanto em termos do conteúdo produzido como a nível da interação dos utilizadores (Pallister, 2019, citado por Gambarato et al., 2021).

Os serviços *OTT* permitem criar um efeito "*network*" (Cennamo and Santalo, 2013, citado por Gambarato et al., 2021), em que o valor acrecentado aumenta mediante o aumento do número de subscritores dos serviços, uma vez que existem mais possibilidades de interação entre estes (McIntyre and Srinivasan, 2017, citado por Gambarato et al., 2021). A este, acrescenta-se o efeito de "saliência" (Kiousis, 2004, citado por Gambarato et al., 2021), relacionado com a atenção e proeminência que pautam o processo comunicativo, usado por serviços *OTT* para tornar elementos do seu serviço mais evidentes na respectiva plataforma. Ambos os efeitos consolidam as plataformas de forma a manietar o seu conteúdo, favorecendo produções originais em detrimento de conteúdo resultante de parcerias estabelecidas, o que determina aquilo que permanece e desaparece da consciência pública, e que, por sua vez, transforma o utilizador num "*tastemaker*" (Gilchrist and Luca, 2017, citado por Gambarato et al., 2021).

A mini-série dramática "Chernobyl" foi utilizada para demonstrar como estes serviços contribuem para restaurar a memória "cultural" da audiência, transversal a diversas regiões e contextos, apesar do efeito nunca ser tão significativo para mercados "*non-english speaking*" como o africano e asiático-este (Lobato, 2019, citado por Gambarato et al., 2021). Determinou-se que poderá ter potencial para: (1) Sensibilizar a audiência para eventos históricos de relevo; (2) Despoletar discussões sobre tópicos controversos e complexos; (3) Alterar a perceção do passado; (4) Ligar memórias coletivas e pessoais de forma a cultivar a audiência sobre nuances culturais, políticas e económicas pelas quais se rege a atualidade (Pallister, 2019, citado por Gambarato et al., 2021).

A variedade auferida pelas plataformas de entretenimento atualmente, qualquer que seja a modalidade de consumo, resulta nos três seguintes fenómenos:

- Fragmentação, em que os utilizadores são distribuídos pelos diferentes tipos de conteúdo (Napoli, 2003, citado por Suwarto, 2021). São exploradas três perspectivas para o estudo deste conceito: Centrada no tipo de entretenimento ou "*Media-Centric*", para caracterizar a audiência de determinado conteúdo; Centrada no utilizador ou "*User-Centric*, para caracterizar as escolhas realizadas por este; Centrada na audiência, ou "*Audience-Centric*", que examina um tipo específico de *software* de rastreio;
- Polarização, tendência para um grupo de utilizadores focar a sua atenção num determinado produto ou serviço, tendo em conta todas as possibilidades que tem ao seu dispor (Webster & Ksiazek, 2012, citado por Suwarto, 2021);

- "*Duality of media*", processo que agrega as preferências de um certo utilizador em segmentos e os usa para direcionar um utilizador para determinado conteúdo (Webster, 2005, citado por Suwarto, 2021).

São paralelamente enumerados três modelos de fragmentação: *Intramedia*, expansão do consumo através da utilização de várias plataformas de entretenimento simultaneamente; *Intermedia*, expansão de um método de consumo através da promoção da variedade de conteúdo; *Transmedia*, escolha de um tipo de conteúdo com a possibilidade de alteração da sua natureza e propósito (Dow & Arango-Forero, 2017, citado por Suwarto, 2021).

2.10 Métodos empíricos e modelação analítica dos artigos revistos

A utilização da abordagem em estudos de "preferências definidas" permitiu apurar que a predisposição do consumidor para dispender tempo ou dinheiro, está intrinsecamente ligada a índices de satisfação e qualidade percecionada, é positivamente influenciada pela quantidade de conteúdo e disponibilidade e negativamente pela presença de anúncios e necessidade de partilha de informação pessoal do consumidor, ainda que não-identificável (Glasgow & Butler, 2017, citado por McKenzie et al., 2019).

No que diz respeito a modelos de *machine* ou *deep learning* no contexto de análises de tráfego e consumo por parte de utilizadores, regista-se a utilização de modelos como AD, *SVM*, *KNN* e *k-means*. Neste contexto, foram já analisados a utilização e o esforço a nível de *network* de um conjunto de serviços *OTT*, através da utilização de vários algoritmos que trabalharam segundo 3 variáveis de consumo, representação decimal do IP do utilizador, tempo dispendido em cada serviço *OTT* e número de *bytes* usados por segundo por *OTT* (Choi & Kim, 2021).

No que diz respeito à satisfação, esta foi definida como a comparação entre o que uma pessoa tem e o que sente que merece, pode esperar ou aspira a ter (Campbell, 1981, citado por Chick et al., 2020). Foi prevista no contexto de uma análise consoante várias dimensões, nomeadamente, *stress* ou saúde percecionados, participação em atividades de lazer e satisfação das pessoas mais próximas de um certo utilizador. Foram aqui também consideradas caraterísticas pessoais como a idade, rendimento, nível de escolaridade, estado civil, género e situação de residência. São ainda calculados os *eigenvalues* no contexto da análise fatorial e os *VIF*s no contexto da regressão linear múltipla utilizando as variáveis enumeradas anteriormente (Chick et al., 2020).

2.11 Avaliação da qualidade dos artigos revistos

É possível verificar, através da distribuição de cotações tabelada e apresentada em anexo, que as dimensões de investigação não estão representadas igualitariamente. Nomeadamente, o C6 é o único com uma cotação inferior a 10, que constitui um elo em comum entre as publicações científicas consultadas, na medida em que os autores relegam empregar o seu sentido crítico em prol da utilização de ideias apenas quando são suportadas por aquelas que as precederam.

A Q3 é a que tem maior representatividade e embora não apresente a menor média por critério, destaca-se a falta de adequabilidade da RL relativamente à aplicação de modelos preditivos, sendo que a maioria dos artigos não culmina nesse tipo específico de análise. Estruturalmente, a presente proposta de investigação, afasta-se, não propositadamente, de qualquer artigo consultado, quer em termos do explicado anteriormente como a nível da decomposição dos hábitos de consumo e respectiva exploração desses dados, o que explica a máxima cotação para um artigo ser igual a 6. Houve uma flagrante dificuldade em encontrar artigos que seguissem os mesmos moldes da presente investigação, no que diz respeito à definição de variáveis com base em indicadores específicos de consumo e quanto à intenção de prever os indíces de satisfação e qualidade percecionada, que serão basilares à determinação da adequabilidade do conteúdo de entretenimento disponibilizado atualmente. Foi necessária a revisão de artigos com apenas parcial correspondência com a temática mas que almejam tentar prever a satisfação através de vários vetores pré-definidos, nomeadamente ao nível do lazer, mas sem particular incidência no entretenimento.

3 Metodologia

Segue-se a explicação do *modus operandis* da investigação e racionalização dos métodos e ferramentas utilizados. Assente nos princípios por que se rege a metodologia *CRISP-DM*, este estudo é compreendido por 2 fases consecutivas: 1) formulação e distribuição do questionário; 2) tratamento e processamento dos dados compilados, provenientes das respostas ao questionário realizado.

O *survey*, apresentado em anexo, foi formulado pelo mestrando tendo em consideração os artigos científicos revistos anteriormente, quer ao nível do teor das perguntas, quer ao nível da segmentação das questões, ênfase e extensão. Foi produzido via *Google Forms* e disponibilizado a duas pessoas que atuaram como "*focus testers*", de modo a fornecer *feedback* antes da sua distribuição formal. Após *input* e respectiva validação, foram recolhidas 100 respostas, contemplando uma amostra de conveniência e, portanto, de carácter não-probabilístico, e os dados foram extraídos na forma de um ficheiro .csv.

A distribuição foi efetuada: Com o auxílio do departamento de *marketing* da entidade patronal do mestrando, distribuindo pelos colaboradores da organização predispostos a colaborar com a investigação; Comunicando às pessoas mais próximas do mestrando por intermédio das plataformas de mensagens diretas da rede social Facebook, a saber, Messenger e Whatsapp. A amostra contemplada não segue um princípio específico de distribuição nem foi restrita a qualquer segmento populacional, sendo toda a filtragem de conteúdo feita *à posteriori*, na fase de tratamento de dados.

O questionário apresenta duas bifurcações que afetam o número de perguntas respondidas pelos participantes, sendo a primeira na Q4 da secção 2 em que a opção "Nenhum/Terminar *Survey*" significa submissão imediata. A segunda não é transparente para o participante, de forma deliberada, de modo a obter respostas condizentes com o intuito do estudo. Trata-se da escolha efetuada na Q12 da secção 4 onde as opções "Serviços de subcrição" e "Pirateado" levam o inquirido ao preenchimento da secção 5, contrariamente às restantes opções. As perguntas desta secção destinam-se aos utilizadores que têm preferência por serviços de subscrição relativamente às restantes modalidades de consumo, tendo sido determinado que utilizadores mais casuais não teriam as respostas intendidas para esta secção.

Foi também tomada a decisão de não incluir uma quantidade excessiva de perguntas do âmbito pessoal, que poderiam caracterizar a amostra de forma mais exaustiva. Alguns exemplos incluem a escolaridade, situação profissional ou agregado familiar, informação que se interpretou não acrescentar dimensões de análise relevantes o suficiente para serem incluídas. Em contrapartida, a sua exclusão torna o questionário mais conciso, tentando preservar ao máximo o interesse do participante ao longo do preenchimento. Adicionalmente, foi feito um esforço de reduzir ao máximo os estrangeirismos contidos, tanto nas perguntas como nas opções de resposta de forma a não alienar qualquer dos participantes englobados, servindo também para efeitos de coesão e formalidade. Ainda assim, é apresentada uma opção de resposta na Q15, "*Word of Mouth*", cuja tradução literal mais próxima, não conferia o significado pretendido para a mesma.

3.1 Limitações, lapsos e *input* recebido

Da fase preliminar, denominada anteriormente como "*focus testing*", foi recebido *feedback* no sentido de aumentar o grau de formalismo das perguntas e respectivas opções, como é o exemplo da Q12.

No *post mortem* efetuado pelo mestrando foram retiradas algumas ilações relativamente à formulação do questionário. Foi acrescentada a possibilidade de opções de respostas em múltipla seleção, nomeadamente para o caso da Q14. Excecionalmente, foi também explorada a hipótese de adição de uma ordem de preferência mas rapidamente abandonada em prol de formular as questões de um modo distinto ou de adicionar as escalas da secção final. Na Q9, houve falha em não acrescentar a opção "*Rating*", algo que está destacado apenas em alguns contextos, sobretudo ao nível das plataformas dos serviços de subscrição.

Os comentários registados à pergunta 21, de resposta aberta, expõem alguns dos lapsos ou falhas de exploração de alguns tópicos na formulação do questionário, transcritos formalmente e apresentados em anexo. Contudo, findo o processo de distribuição, foi também registado o nível de interesse dos participantes na pergunta 20, gráficamente apresentado também em anexo, indicativo de que a formulação das perguntas vai de alguma forma ao encontro dos tópicos que os consumidores de conteúdo de entretenimento pretendem ver explorados. Este nível de interesse é maioritariamente positivo, o que revalida as intenções da presente proposta de investigação, não descurando as áreas onde o questionário poderia ter sido aprimorado, como apontado por alguns dos participantes.

3.2 Carregamento e tratamento dos dados

Os dados serão extraídos da plataforma *Google Forms* e carregados no editor de texto *Jupyter Notebook*, acessível através da plataforma/*launcher Anaconda Navigator* para posterior validação e manipulação dos mesmos. Esta fase serviu principalmente para aprimoramento da informação gerada via questionário para efeitos da aplicação e construção de modelos preditivos. Foram realizados ajustes e correções no sentido de tornar os dados concordantes com as variáveis de estudo com as quais se pretende trabalhar, as dependentes, qualidade e satisfação, e independentes, por exemplo, o número de géneros consumidos e de serviços de subscrição usufruídos. Esta fase contempla também a identificação de dados omissos ou *outliers* que poderão comprometer as consultas a realizar posteriormente. Esta etapa foi revisitada ao longo do desenvolvimento do trabalho, sempre que se justificou.

O desenvolvimento deste projeto foi feito inteiramente no editor de texto mencionado anteriormente com recurso à linguagem programática *Python* por meio das bibliotecas *NumPy*, *Pandas*, *Matplotlib*, *Seaborn* e *Scikit-learn*, para realização de 80% das consultas. Todos os passos serão documentados mediante os comandos introduzidos na ferramenta de trabalho mas apenas os dados constarão da respectiva secção, remetendo todos os comandos para o anexo deste documento. Dado a maioria dos comandos conter argumentos, parâmetros e funções escritos na língua inglesa, foi feito um esforço acrescido para definir todas as variáveis e apresentar todo o texto em português. Isto também se aplica ao texto de suporte aos dados apresentados, nomeadamente legendas ou títulos de gráficos e tabelas, de

modo a facilitar o acompanhamento de todas as lógicas aplicadas. A indentação apresentada para os comandos diverge daquela introduzida no editor de texto de forma a melhorar a apresentação no documento, tendo sido a formatação original ajustada. O comando apresentado em anexo sofre alterações ao nível dos argumentos e definições consoante a representação gráfica ou modelo, sendo apenas uma iteração possível da totalidade dos que foram usados.

3.2.1 Normalização da base de dados

Pré-carregamento da base de dados, através do ficheiro extraído do *Google Forms*, são importadas as bibliotecas fundamentais para a prossecução de praticamente todas as tarefas desempenhadas ao longo do desenvolvimento do trabalho. É necessário expandir a apresentação das células *output* do editor de texto com vista a facilitar certas consultas e permitir recortar imagens via "*Sniping tool*", a introduzir no presente documento. Alguns dos passos preliminares incluem, a contagem do número de registos, listagem das colunas originais para perceber a necessidade de renomeamento, identificação da tipologia de dados e frequência de valores nulos para cada coluna. As últimas duas dimensões enumeradas, são ilustradas na Tabela 6.

Finda esta primeira fase, eliminam-se as colunas inconsequentes para a análise, que não serão utilizadas em qualquer contexto que sucede esta etapa, a saber "Data e Hora de preenchimento do questionário" [6], "Nível de interesse dos participantes no questionário, como foi formulado" e "Reparos e Oportunidades de Melhoria". Os dados contidos nestas duas últimas colunas servem apenas para registo de *feedback* dos participantes, não possuindo materialidade para o trabalho de investigação a realizar. Posteriormente, são também eliminados os registos para os quais foi selecionada a opção "Nenhum/Terminar *Survey*" na questão Q4, que correspondem a participações de utilizadores para quem o questionário não é destinado. São também substituídos os valores nulos por "Não prioritiza serviços de subscrição" nas respectivas colunas quando o participante seleciona "Televisão por cabo" ou "Pirateado" na Q12[7] com a respectiva conversão da tipologia dessas colunas, de forma a evitar inconsistências nas consultas efetuadas. Nesta fase, são também dados outros passos no sentido de aumentar a formalidade dos dados de forma a aprimorar a apresentação neste documento.

[6] A coluna "*timestamp*" é a única que não consta do questionário, uma vez que é criada aquando da transformação para ficheiro csv na exportação do *Google Forms*.
[7] Apenas possível após a remoção de dados omissos que precede este passo

Tabela 6: Renomeamento e resumo funcional das colunas da base de dados

Nome original	Nome atualizado	Registos nulos	Tipologia
'Género'	Sem alteração	0	*object*
'Faixa etária'	Sem alteração	0	*object*
'Rendimento'	Sem alteração	0	*object*
'Interesse/Investimento em Conteúdo de Entretenimento'	'Interesses'	0	*object*
'Número de horas consumidas/dia (média estimada)'	'Horas de consumo/dia'	12	*object*
'Dias da semana com maior incidência'	'Dias da semana de maior consumo'	12	*object*
'Número de séries vistas/mês (média estimada)'	'Séries vistas/mês'	12	*object*
'Número de filmes vistos/mês (média estimada)'	'Filmes vistos/mês'	12	*object*
'Como decide o que ver a seguir?'	'Critérios para o que ver a seguir'	12	*object*
'Dispositivos utilizados para consumo de entretenimento'	'Dispositivos utilizados'	12	*object*
'Variedade de géneros consumidos'	Sem alteração	12	*object*
'Modo de consumo predominante?'	'Modalidade de consumo predominante'	12	*object*
'De que serviços é o conteúdo de que usufrui?'	'Serviços subscritos e/ou usufruídos'	26	*object*
'Razões que motivam a opção por serviços de subcrição'	Sem alteração	26	*object*
'Como veio a descobrir os serviços de subscrição cujo conteúdo usufrui de momento?'	'Modos de descoberta de serviços de subscrição'	26	*object*
'Montante dispendido em conteúdos de entretenimento/mês?'	'Montante dispendido/mês'	12	*object*
'Grau de satisfação com a acessibilidade ou preço pago (se aplicável) do conteúdo televisivo/cinematográfico disponível atualmente?'	'Satisfação com a acessbilidade e/ou preço pago'	12	*float64*
'Nível de qualidade percecionada do conteúdo televisivo/cinematográfico disponibilizado atualmente?'	'Qualidade percecionada'	12	*float64*
'Grau de satisfação com o conteúdo televisivo/cinematográfico disponível atualmente?'	'Satisfação global'	12	*float64*

3.2.2 Complementos à base de dados

Com base nas últimas três colunas numéricas, nomeadamente, "Satisfação com a acessibilidade e/ou preço pago", "Qualidade percecionada" e "Satisfação global" seguindo sempre a mesma lógica, foram adicionadas novas colunas, que contêm o identificador "(Y/N)". Estas têm o intuito da aplicação na construção de modelos baseados em AD, *RF/ensemble* e *SVM*, cujo *input* para o y terá de ser uma variável binária-alvo, sendo os resultados destes modelos confrontados posteriormente com os obtidos através de modelos de regressão linear. A dicotomia estabelece-se no valor 7 de cada escala, patamar mínimo definido para o participante considerar que está satisfeito ou que o conteúdo disponibilizado tem qualidade. Aquando do resumo estatístico, definindo o valor em 5, o intermédio na escala, verificou-

se que os valores para as variáveis dicotómicas eram bastante tendenciosos, na medida em que os participantes selecionaram maioritariamente, valores entre 5 e 10. Ao nível da modelação preditiva, particularmente, nas AD, *RF* e *SVM*, a análise iria ser bastante desvirtuada e o cálculo de valores enviesado, uma vez que os modelos teriam dificuldades em encontrar valores abaixo de 5.

De seguida, contabilizaram-se os valores dentro duma mesma coluna, nos casos em que a pergunta no questionário possibilita seleção múltipla. Efetuado para "Critérios para o que ver a seguir", "Variedade de géneros consumidos", "Serviços subscritos e/ou usufruídos", "Razões que motivam a opção por serviços de subcrição" e "Modos de descoberta de serviços de subscrição". É necessário aplicar uma condição à expressão, uma vez que algumas opções de resposta exigem um tratamento específico, expresso na Tabela 9.

Em última instância, resta apenas reordenar logicamente as colunas de forma a tornar a navegação pela base de dados mais intuitiva para o mestrando ao longo da elaboração do desenvolvimento do trabalho.

Tabela 7: Mapeamento das novas *features* numéricas a adicionar

Coluna	**Valor na coluna original**	**Valor na coluna nova**
Critérios para o que ver a seguir	"Sem critério"	0
	Qualquer outra resposta	Contabilização dos valores
Variedade de géneros consumidos	"Tudo um pouco"	Soma de todos os géneros listados: 9
	Qualquer outra resposta	Contabilização dos valores
Serviços subscritos e/ou usufruídos Razões que motivam a opção por serviços de subcrição Modos de descoberta de serviços de subscrição	0	"Não usa predominantemente serviços de subscrição"
	Qualquer outra resposta	Contabilização dos valores

3.3 Processamento e apresentação de resultados

Findo o tratamento dos dados, avaliação da elegibilidade e restantes passos complementares, procedeu-se à análise de dados exploratórios. Nesta fase, o intuito passa pela realização de consultas à base de dados e respectiva apresentação gráfica e estatística dos dados compilados de forma a estabelecer relações e padrões entre estes. Trata-se de um preâmbulo para a introdução dos algoritmos de *machine learning* que se segue, identificando e avaliando as variáveis a utilizar. A fase final da investigação englobará a construção de modelos preditivos, utilizando as variáveis a identificar seguidamente, tendo como alvo, a estimação da satisfação e perceção de qualidade mediante os diversos dados de consumo disponbilizados. Esta também contempla a análise fatorial e de correlação das *features* numéricas da base de dados, de forma a determinar a multicolinearidade e homogeneidade dos dados. Foram escrutinadas as principais relações, interpretadas como mais lógicas, entre as variáveis-chave e as independentes, efetuando previsões que serão feitas com o auxílio de modelos de regressão linear, AD, *RF* e *SVM*. Ao nível desta última fase, é importante destacar a distribuição das amostras de treino e teste, respectivamente fixada em 70 e 30%, ligeiramente superior à prática convencionada na maioria das

análises desta natureza, 80 e 20%, de forma a contemplar mais da amostra. Os valores apresentados na componente analítica deste trabalho são todos arredondados às duas e quatro casas decimais, respectivamente, na fase da análise exploratória e modelação preditiva.

4 Resultados e ilações

O capítulo vigente apresenta os principais resultados da análise exploratória assim como da modelação preditiva nos moldes descritos anteriormente. Os resultados são sucedidos pelas observações materialmente relevantes a fazer e se o constatado está concordante com o descrito na proposta de investigação e responde às questões inicialmente formuladas. A tipologia de representação gráfica foi selecionada consoante o tipo de consulta ou conjunto de dados e o interpretado como o mais ilustrativo do que se pretende analisar.

4.1 Sumário estatístico

Realizou-se uma decomposição das métricas para as *features* numéricas que compõem a base de dados e respectivas ilações a retirar.

Tabela 8: Principais métricas das *features* numéricas

Coluna	Contagem de registos	Média	Desvio padrão	Mínimo	Quartil 1	Quartil 2	Quartil 3	Máximo
Número de critérios aplicados	88	2.00	1.27	0.0	1.00	2.0	3.00	6.0
Número de géneros consumidos	88	5.57	2.65	1.0	3.00	5.0	9.00	9.0
Número de serviços subscritos	88	2.33	1.78	0.0	1.00	2.0	3.25	8.0
Número de razões motivadoras das subscrições	88	1.53	1.03	0.0	1.00	1.0	2.00	4.0
Número de modos de descoberta de *OTT*s	88	1.20	0.78	0.0	1.00	1.0	2.00	3.0
Satisfação com a acessibilidade e/ou preço pago	88	7.24	1.79	2.0	7.00	7.0	8.00	10.0
Satisfação com a acessibilidade e/ou preço pago (Y/N)	88	0.48	0.50	0.0	0.00	0.0	1.00	1.0
Qualidade percecionada	88	7.45	1.71	1.0	6.00	8.0	9.00	10.0
Qualidade percecionada (Y/N)	88	0.56	0.50	0.0	0.00	1.0	1.00	1.0
Satisfação global	88	7.44	1.46	4.0	6.75	8.0	8.00	10.0
Satisfação global (Y/N)	88	0.53	0.50	0.0	0.00	1.0	1.00	1.0

Em média, os utilizadores não precisam de mais que dois critérios para decidir o que ver a seguir, o que demonstra relativa assertividade nas escolhas que fazem, considerando o vasto leque de informação que têm ao dispor para tomar a decisão. Tendo sido listado um total de 6 critérios na respectiva questão formulada, é importante realçar, que pelo menos um dos participantes faz uso de todos eles.

No número de géneros consumidos, encontramos a maior dispersão relativamente ao número médio entre as *features* numéricas assim como o maior valor médio registado entre as variáveis explicativas,

excluindo os índices-alvo do final da Tabela 9. Estes números espelham a volatilidade de preferências do consumidor de entretenimento atual, mais predisposto a variar o conteúdo de que usufrui em detrimento de se cingir a um género em particular que sabe à partida ser do seu agrado.

Os utilizadores contemplados na amostra subscrevem ou usufruem em média de 2 serviços, registando-se a segunda maior dispersão deste valor entre os potenciais preditores, havendo participantes que tenham indicado um total de 8 serviços subscritos. Apesar dos serviços *OTT* agruparem benefícios ou ofertas de similar natureza e conferirem um maior nível de personalização, os utilizadores consideram ainda assim necessário contemplar no seu portfólio múltiplos destes, de forma a terem acesso a todo o conteúdo desejado.

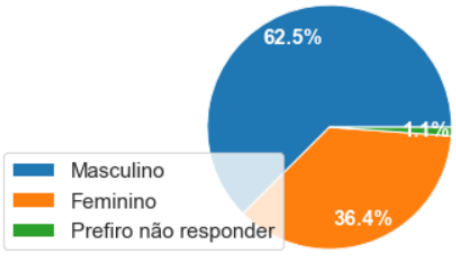
O número de razões para usufruir de serviços *OTT* e número de modos como são descobertos apresentam valores similares, sendo os desta última variável marginalmente inferiores, potencialmente, dado o facto de haver menos opções enumeradas aquando da formulação da pergunta. Regista-se também que o valor máximo de razões identificadas em simultâneo é superior ao valor máximo de modos de descoberta, provavelmente devido à facilidade inerente a cada pergunta. O que leva a apurar que é um exercício mental de menor complexidade identificar as razões que levam o utilizador a subscrever um serviço atualmente do que tentar recordar como veio a saber da existência de um.

Entre os principais índices de análise, verifica-se que a qualidade percecionada tem o valor médio mais alto tanto para a variável original como binária, o que significa que é aquela com mais valores indicados superiores a 7 na escala da variável correspondente. É também a que apresenta a maior amplitude de valores relativamente à média, porém, é a satisfação com a acessibilidade e/ou preço pago a mais desviante ou dispersa relativamente aos valores médios.

4.2 Caracterização do perfil da amostra inquirida

Nesta secção, pretende-se apontar as características distintivas dos participantes do questionário contemplados através de *pie charts*, aferindo sobre a sua adequabilidade para a presente investigação.

Figura 3: Representatividade de género



Dado que a maioria dos participantes do questionário pertence ao setor de atividade das tecnologias de informação, área de atividade do mestrando, a representatividade apresentada é reflexo das circunstâncias do mestrando ainda que não seja tão igualitária quanto pretendido. Estima-se que em

2022, apenas 26% da *workforce*, para posições de carácter tecnológico, seja constituída por mulheres, sendo que esse número é inferior quando ascendemos a hierarquia corporativa (Howarth, 2022).

Figuras 4: Características da amostra inquirida

Figura 5 A – Faixa Etária

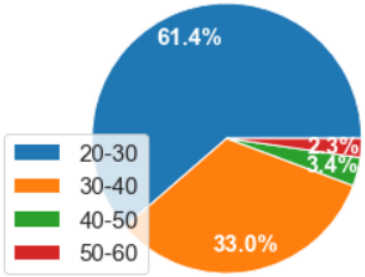


Figura 5 B - Rendimento

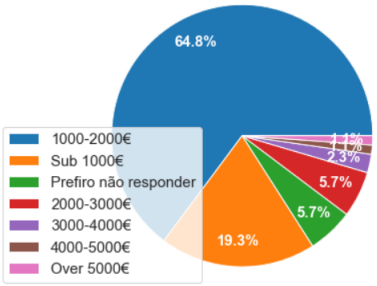
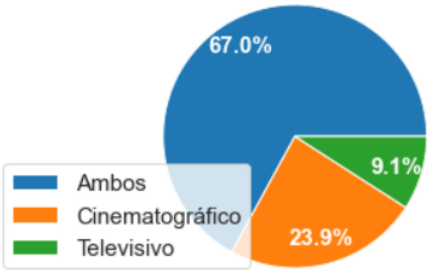


Figura 5 C - Interesses

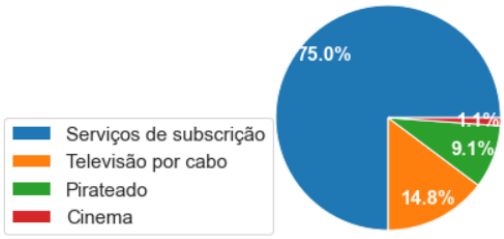


É possível identificar que em todos os gráficos acima apresentados parece haver um valor mais frequente, com uma representatividade idêntica entre eles. As distribuições são igualmente representativas do setor de atividade de grande parte da amostra contemplada assim como das características da entidade patronal do mestrando.

4.3 Análise comparativa das modalidades de consumo de conteúdo de entretenimento

Cálculo das métricas com maior grau de representação para cada coluna ou *feature*, numérica ou não, agrupadas pela modalidade de consumo predominante identificada na Q2 do questionário. São identificadas, para cada uma, as razões explicativas dos resultados obtidos, confrontadas com assunções pré-concebidas.

Figura 5: Modalidade de consumo predominante



O encapsulamento do atributo "Dias de maior consumo", na Tabela 10, significa que existe mais do que um valor mais frequente. Ou seja, para a categoria "Cinema", a moda é constituída por ambos os valores "Sexta" e "Sábado-Domingo".

A distribuição entre as diferentes modalidades de consumo demonstra o alcance e proeminência dos serviços de subscrição na amostra contemplada. De realçar a falta de expressividade da categoria "Pirateado" relativamente aos "Serviços de Subscrição" na medida em que os recursos à disposição dos consumidores neste âmbito são, atualmente, mais numerosos e de mais fácil acesso. A "televisão por

cabo" perde prevalência, dada a sua conhecida rigidez de programação e acessibilidade e por, em muitos casos, ser uma opção mais dispendiosa, exigindo a compra de pacotes que oferecem conteúdo excedentário mas necessário para aceder ao pretendido.

Tabela 9: Apresentação das métricas calculadas – confronto das modalidades de consumo

Atributo/*Feature*	**Estatística**	**Modalidade de consumo**			
		Cinema	Pirateado	Serviços de subscrição	Televisão por cabo
Horas consumidas/dia	Moda	2h-3h	1h ou menos	2h-3h	2h-3h
Séries vistas/mês	Moda	1-2	1-2	1-2	1-2
Filmes vistos/mês	Moda	3-4	1-2	1-2	1-2
Dias da semana de maior consumo	Moda	[Sexta, Sábado-Domingo]	[Sábado-Domingo]	[Sábado-Domingo]	[Sábado-Domingo]
Número de critérios aplicados	Média	2	2	2	2
Critérios utilizados	Moda	[Género, Popularidade]	[[Atores envolvidos, Realizadores], [Género]]	[[Género], [Género, Atores envolvidos, Popularidade]]	[Género, Popularidade]
Dispositivos utilizados	Moda	Televisão	Outros dipositivos móveis	Televisão	Televisão
Número de géneros consumidos	Média	5.00	5.88	5.41	6.23
Número de serviços subscritos/usufruídos	Média	-	2.12	2.85	-
Número de razões motivadoras	Média	-	1.25	1.89	-
Número de modos de descoberta	Média	-	1.38	1.44	-
Montante dispendido/mês	Moda	Menos de 25€	Menos de 25€	Menos de 25€	Menos de 25€
Satisfação com a acessibilidade e/ou preço pago	Média	7.00	5.88	7.50	6.77
Qualidade percecionada	Média	5.00	6.25	7.73	7.00
Satisfação global	Média	4.00	6.25	7.70	7.15

É feito o argumento de que o tempo livre de que devemos dispôr, de forma a maximizar o bem-estar e felicidade, está compreendido no intervalo de 2 a 5 horas, sendo os dados registados, concordantes desta noção (Holmes, 2022). Porém, contrariamente ao esperado, para a categoria "Pirateado", imaginar-se-ia que, pela acessibilidade conferida, este valor seria superior ao das restantes modalidades de consumo, o que não se verifica.

Relativamente ao número de filmes e séries vistos numa base mensal, aponta-se apenas a discrepância para o participante que seleciona "Cinema" como a sua modalidade de consumo predominante com uma incidência maior aos restantes, apesar deste tipo de consumo ser mais dispendioso e menos prático. Os dias úteis são preteridos em termos de consumo de entretenimento, como seria de esperar, mesmo em contexto pós-pandémico, ainda que o facto da maioria da população trabalhar mais tempo em casa e dispor de mais tempo pudesse acrescentar uma variante a esta dimensão da análise. Sexta-feira é também privilegiada pelo cinéfilo da amostra, não sendo registado para mais nenhuma categoria.

Os participantes utilizam em média 2 critérios de decisão para o que consumir a seguir, qualquer que seja a categoria de consumo selecionada. A televisão continua a ser a principal ferramenta de

visualização, com exceção da categoria "Pirateado", que privilegia outros dispositivos como o computador, opção não contemplada na respectiva pergunta do questionário, mas indicada pelo participante na pergunta de resposta aberta. Para as plataformas digitais, possibilitando mais facilmente o consumo em qualquer dispositivo *Smart*, poderiam ter sido registados resultados diferentes mas, nas circunstâncias atuais, em que a população é mais sedentária, os hábitos de consumo permanecem inalterados a este nível.

A televisão por cabo é a categoria onde se regista a maior média de géneros consumidos, apesar da maior acessibilidade e variedade de conteúdo encontrada nas outras modalidades de consumo. Existe, por norma, uma maior desagregação do conteúdo, o que dificulta encontrar o que se pretende, ainda assim, continua a ser o método que permite aos utilizadores inquiridos diversificar mais o seu portfólio de entretenimento.

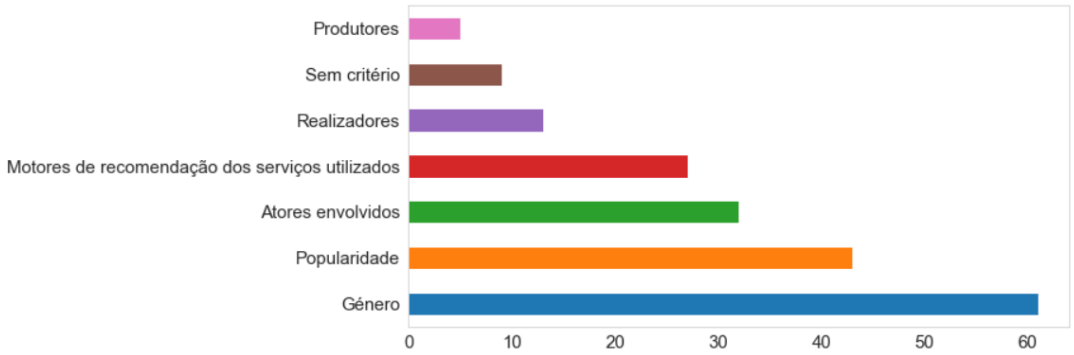
Dentro dos subscritores que usufruem predominantemente de serviços *OTT*, o número médio de serviços é inferior para a modalidade "Pirateado", inesperado à partida visto que têm um número superior à disposição. Adicionalmente, é importante frisar que o valor oferecido por um serviço é considerável, porém, caso o utilizador tenha interesse em conteúdo oferecido em múltiplos, o gasto pode tornar-se bastante significativo. Este facto tem a salvaguarda dos métodos de pagamento que utiliza, base mensal ou anual com ou sem desconto, se está incluído num plano familiar e que patamar da subscrição utiliza, com ou sem anúncios e com mais ou menos opções de visualização, quando aplicável.

O número médio das razões que motivam os utilizadores a subscrever ou usufruir é superior para "Serviços de Subscrição", explicado pelo facto destes utilizadores dispenderem parte do seu rendimento nas subscrições, o que leva à identificação de mais razões que justifiquem esta decisão. Qualquer que seja a modalidade de consumo, a resposta mais frequente entre os participantes para o montante dispendido por mês é menos de 25€. Foi também feito o mesmo exercício em função do rendimento, ainda que não apresentado aqui, e apenas a categoria "3000€-4000€" mostra a mesma frequência para o valores "Menos de 25€" e "25€-50€", o que é reflexo do nível de zelo financeiro dos participantes mas poderá ser representativo do maior critério que os consumidores estão a demonstrar no que diz respeito a gastos "não essenciais".

4.4 Distribuição dos valores para *features* com múltipla seleção

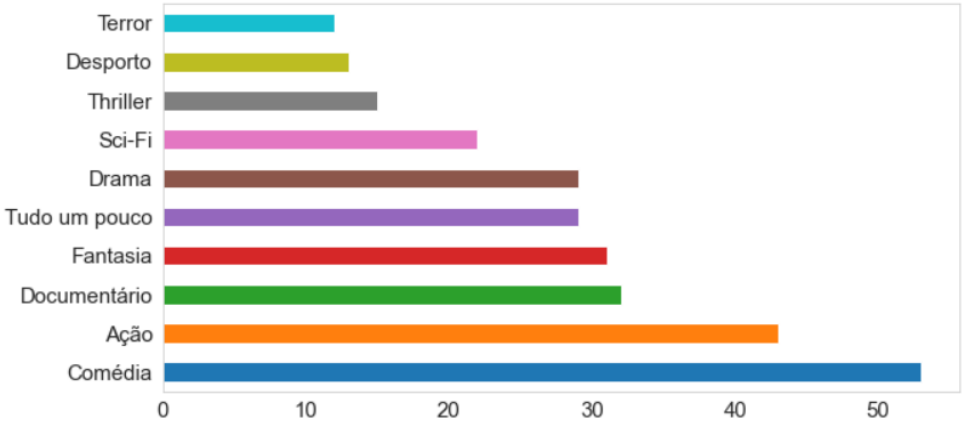
Nesta secção, são indicados valores obtidos através de gráficos de barras, produzidos aquando da fragmentação das respectivas colunas, isto é, para features com seleção múltipla, é necessário separar cada valor através do encapsulamento na base de dados, de forma a contabilizar cada um individualmente. Estes valores são os principais indicadores que distanciam e caracterizam a amostra e procurar-se-á contextualizá-los relativamente à atualidade das indústrias de entretenimento.

Figura 6: Frequência dos diferentes critérios



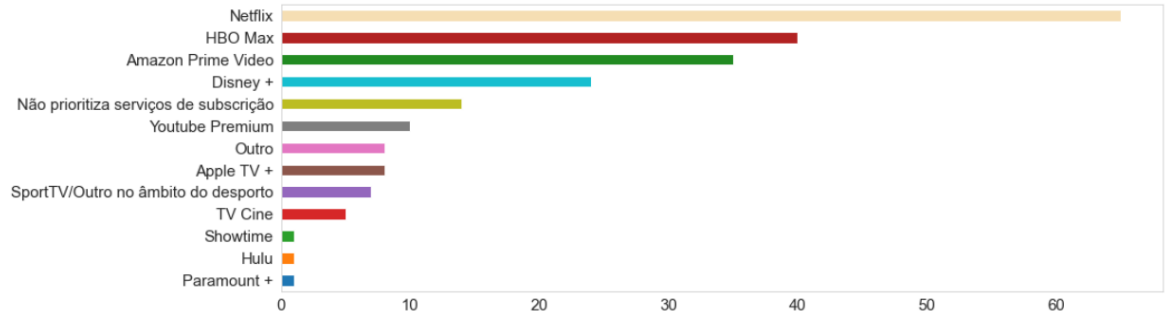
É possível verificar que os critérios "Produtores" e "Realizadores" são os menos representados na amostra, ainda que sejam os elementos mais determinantes para o (in)sucesso de um dado projeto. Estes são os intervenientes mais participativos numa dada produção, tendo mais influência direta na natureza do que é desenvolvido do que os atores, o 3º critério mais utilizado. É possível constatar que o género e a popularidade são os mais evidenciados, tanto pelo gráfico como pela Tabela 10.

Figura 7: Frequência dos diferentes géneros



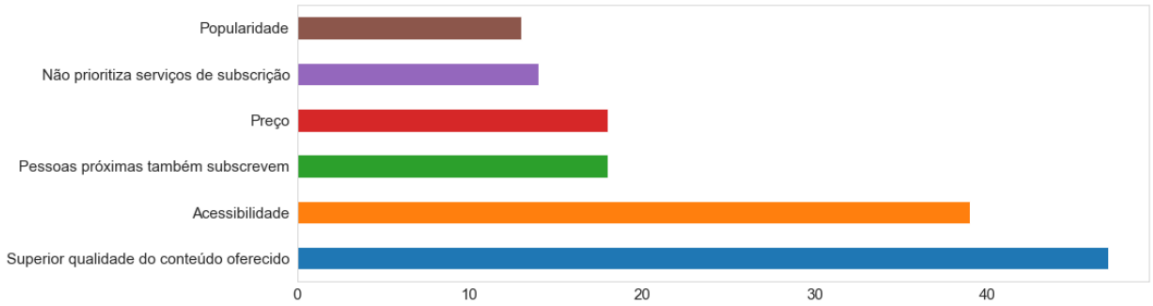
Tendo como base de comparação os dados apresentados pela Statista em 2021, constata-se que a amostra revela uma incidência bastante dispar no género "Drama", que é o que tem maior representatividade para "Os géneros mais populares, para séries originais, nos Estados Unidos" (Stoll, 2022). A restante distribuição está moderadamente alinhada com os dados publicados quer ao nível da Statista como das plataformas digitais mais populares, exemplo do filme "*Red Notice*", da Netflix, pertencente ao género "Ação-Comédia", que é o que tem, atualmente, o maior número de horas de visualizações.

Figura 8: Frequência dos serviços subscritos



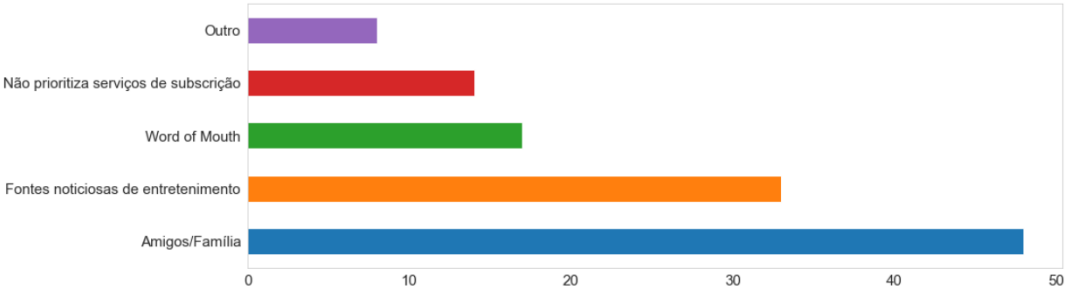
Os valores registados vão ao encontro dos últimos dados publicados até ao primeiro trimestre de 2022 nos Estados Unidos, registando-se as maiores discrepâncias a apontar para o serviços "HBO Max", que ocupa a posição número 4 em vez de 2 e a fraca preponderância da "Hulu" no mercado português, ocupando a posição 5 nos EUA, atrás da "Disney +", de quem é subsidiária (*IGN*, 2022).

Figura 9: Frequência das razões motivadoras para subscrever serviços de subscrição



As duas representações gráficas anteriores são concordantes em termos do quanto o preço não afeta a decisão de subscrever ou usufruir de um dado serviço, uma vez que para além de aparecer na 3º posição entre as "Razões motivadoras" menos preponderantes, a Netflix ocupa a 1ª posição dos mais subscritos, sendo o mais dispendioso da lista, não oferecendo sequer a possibilidade de realizar o pagamento anualmente, com um preço mais reduzido, tendo em comum com os restantes apenas o período experimental gratuito para novos assinantes.

Figura 10: Frequência dos modos de descoberta

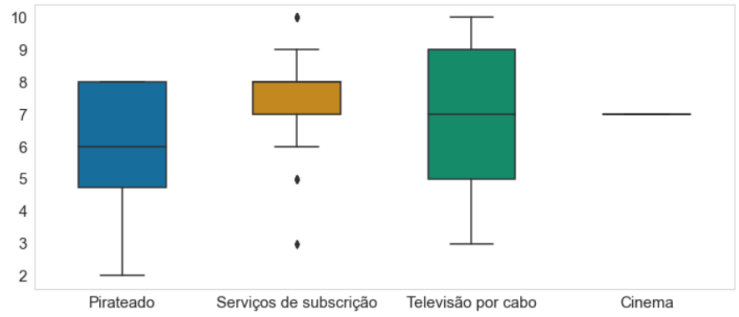


O alcance promocional dos novos serviços subscritos é notório, ainda assim, o principal modo de descoberta registado é "Amigos/Família", elemento influenciador das decisão de subscrever/usufruir de determinado serviço. O esforço de *marketing* das organizações que oferecem estes serviços é agora menos necessário, dada a proeminência, presença digital e o facto da sua disseminação ocorrer de forma orgânica de utilizador para utilizador.

4.5 Amplitude e distribuição de valores para os principais índices de análise

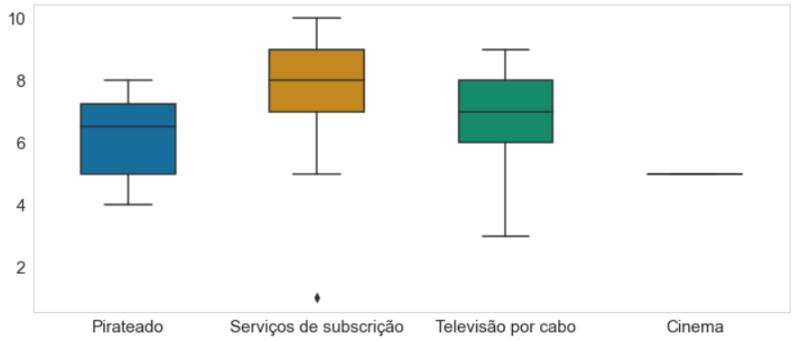
Esta secção apresenta a distribuição dos valores dos principais índices de análise em função da modalidade de consumo predominante, através de *boxplots* ou diagramas de extremos e quartis. Esta secção serviu como um preâmbulo da fase da modelação preditiva.

Figura 11: Satisfação com a acessibilidade e/ou preço pago



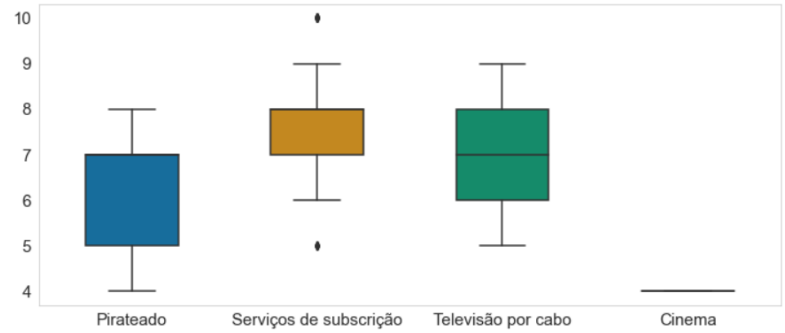
A volatilidade do grau de satisfação verifica-se mais acentuado para os participantes cuja modalidade de consumo predominante é a televisão por cabo, sendo ligeiramente menor para quem usufrui de conteúdo pirateado. Já mencionado anteriormente, é frequente o caso em que, para quem assina pacotes de televisão por cabo, é mais moroso o acesso ao conteúdo verdadeiramente desejado, uma vez que tem de filtrar todo aquele que não o é. O preço pago raramente é representativo da percentagem de conteúdo usufruído mas sim do pacote como um todo, contendo componentes excedentárias. É importante destacar que se trata da única categoria para a qual foi selecionado um nível de satisfação igual a 10 consistentemente, dada a maturidade desta modalidade de consumo e esforço publicitário inerente que exponencia a visibilidade dos utilizadores para os produtos oferecidos. A amplitude entre o primeiro e terceiro quartil é também considerável para a categoria "Pirateado", definitivamente não pelo preço mas talvez pela dificuldade em chegar ao recurso *online*, dentro dos muitos existentes, que disponibiliza o que o utilizador pretende. O valor singular registado para a categoria "Cinema" não é particularmente alto, uma vez que o "preço de admissão" é bastante alto e o usufruto de uma sessão num cinema físico implica a deslocação do utilizador.

Figura 12: Qualidade percecionada



As amplitudes da variação verificadas são praticamente indistinguíveis para todas as modalidades exceto a "Cinema". O valor registado para esta categoria pode ser reflexo da experiência de ir ao cinema físico estar relacionada maioritariamente com o aspecto social em detrimento da qualidade do filme. A categoria "serviços de subscrição" é a única para a qual se registam níveis de qualidade iguais a 10, explicado, entre outros fatores, por destacar o *rating* de determinado conteúdo na sua página principal, ter conteúdo original com elevados níveis de produção e disponibilizar conteúdo "*third-party*" de grande notoriedade ou popularidade com frequência. É importante destacar o valor apresentado para a categoria "Cinema", discordante do verificado no índice diretamente acima, o que poderá ser motivado pelos direitos de transmissão que dita o catálogo das salas de cinema portuguesas, que não está a oferecer conteúdo com a qualidade desejada ou com a duração pretendida. Os valores registados para a categoria "Pirateado" poderão ser explicados pelo fenómeno denominado "*choice paralysis*", em que o utilizador não consegue escolher perante um vasto leque de opções à sua disposição, o que em última instância, pode estar a levá-lo a consumir conteúdo sem a qualidade pretendida.

Figura 13: Satisfação global



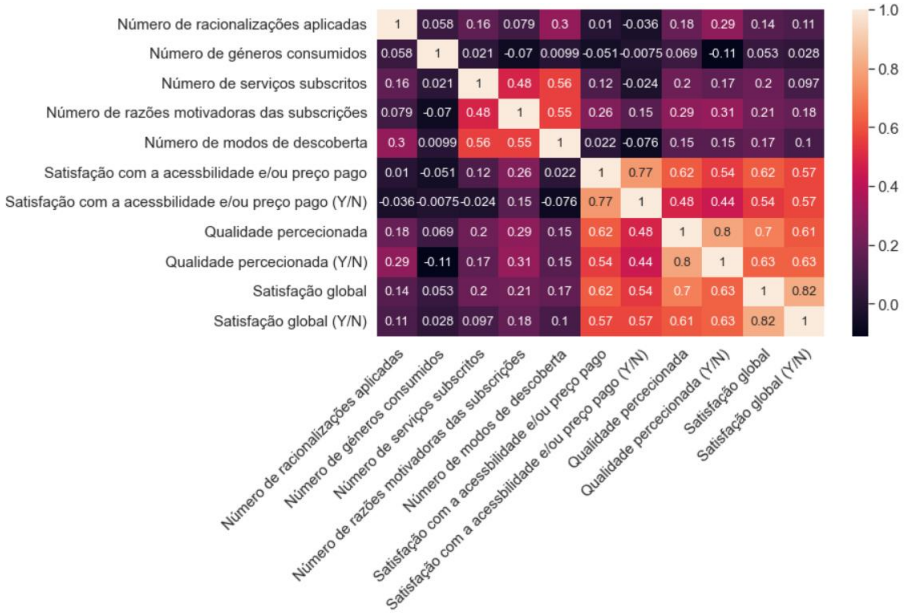
É incontornável estabelecer paralelismos entre esta e a representação gráfica para a satisfação com a acessibilidade e/ou preço pago. Regista-se que o terceiro quartil da modalidade "Pirateado" é igual ao primeiro quartil da modalidade "Serviços de Subscrição", o que espelha a importância do usufruto do conteúdo através da respectiva plataforma para a satisfação global, contrariamente ao acesso via *browser* ou *website* ilegítimo, que acaba por desvirtuar a experiência do utilizador. O valor singular indicado para a modalidade "Cinema" demonstra maior correspondência com o registado na qualidade

percecionada o que pode ser explicado pelos fatores já referidos, preço e falta de comodidade. Apesar de ser um *outlier*, o valor máximo é registado para a modalidade "Serviços de Subscrição" assim como o menor valor da amplitude interquartílica, ilustrativo da consistência do nível de satisfação dos consumidores inquiridos.

4.6 Modelação preditiva

Esta é a fase em que culmina a investigação, composta pela construção de modelos preditivos com base em quatro algoritmos de *machine learning* distintos, a saber, regressão linear, *CART*, *RF e SVM*. Pretende-se primeiramente apresentar as relações entre as variáveis, tanto para os potenciais preditores/variáveis explicativas como para os principais índices de análise/variáveis alvo.

Figura 14: *Heatmap* para os rácios de correlação[8]



A representação de rácios de correlação, utilizando o coeficiente padrão *Pearson*, posiciona e contextualiza a adequabilidade das variáveis numéricas que compõem a base de dados constituída e tratada na fase anterior da investigação. Entre as relações que mais destaque merecem, verificamos a relação mais pronunciada, (0.56), entre o número de serviços subscritos e o número de modos de descoberta destes, que indica que, quanto maior for o grau de exposição a ferramentas de comunicação e canais publicitários e ou de distribuição, maior é a incidência de usufruto ou subscrição de serviços. Não muito distanciada, encontra-se a relação entre o número de razões que motivam a subscrição e os modos de descoberta de serviços, (0.55), que poderá também ser demonstrativo, em concordância com a ilação anterior, do grau de influência de campanhas publicitárias ou de *feedback* de outros utilizadores

[8] Apenas para efeitos demonstrativos, não vinculativo para o estabelecimento das relações "lógicas", diretamente abaixo estabelecidas e empregadas na construção de modelos preditivos

de características e preferências similares, sendo que estes podem ser próximos do consumidor ou presentes nas redes sociais que utiliza.

Menos significativa, é a relação entre as razões que motivam a escolha por serviços *OTT* e o número que de facto subscreve ou usufrui, (0.48), o que não seria, à partida, intuitivamente expectável. O facto do utilizador justificar exaustivamente as suas escolhas poderia ser reflexo da opção por apenas um dos serviços, que reúne todos os requisitos que procura.

Ainda que haja quase uma ausência de relação, nota-se a relação inversa existente entre o número de géneros consumidos e a satisfação com a acessibilidade e/ou preço pago, (-0.051), que se traduz no facto de quanto mais géneros ou mais variado é o conteúdo que um utilizador consome, mais dificuldades têm em encontrar o conteúdo pretendido e/ou está mais descontente com o montante dispendido que lhe permite usufruir do vasto repertório de que dispõe.

Entre os principais índices de análise, verificamos a relação mais forte entre a qualidade percecionada e a satisfação global, (0.7), o que é ilustrativo da menor aversão ao preço dos utilizadores, quando reconhecem o valor inerente de um determinado produto. Desta ilação, não está descartada a possibilidade dos participantes, quando confrontados com respostas em escala consecutivas, serem inevitavelmente condicionados e tenderem a indicar valores próximos ou até mesmo iguais.

Da transformação promovida para as 3 variáveis-alvo, constata-se que a satisfação global foi a menos alterada aquando da normalização, processo de transformação da variável original em binária, na medida em que é a que apresenta uma maior correlação com a derivada, com o sufixo "(Y/N)", sendo a que apresenta a maior diferença, a satisfação com a acessibilidade e/ou preço pago.

Figuras 15: Satisfação global

Figura 16 A - Critérios

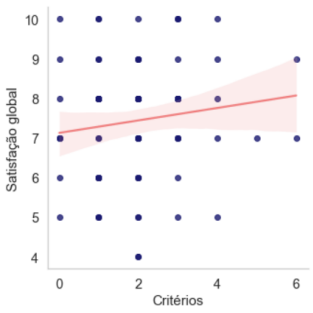


Figura 16 B - Géneros

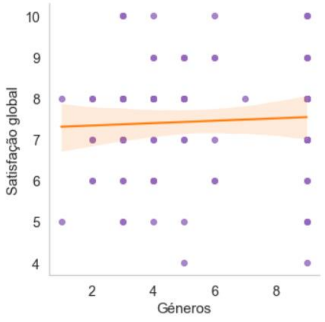


Figura 16 C - Serviços

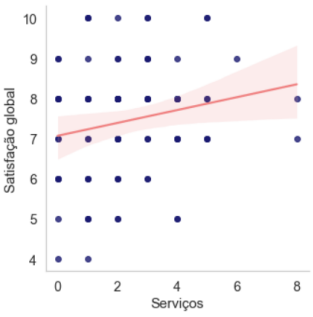


Figura 16 D – Razões *OTT*s

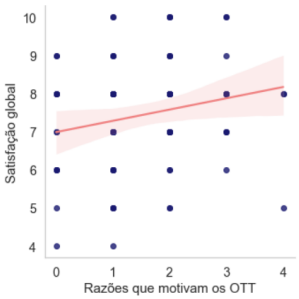
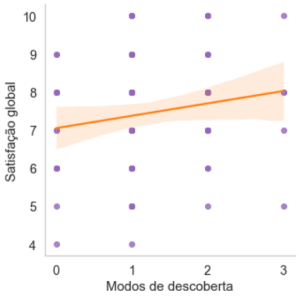


Figura 16 E – Descoberta *OTT*s



Figuras 16: Satisfação com a acessibilidade/preço pago e qualidade percecionada

Figura 17 A – Serviços SAP **Figura 17 B** – Razões OTTs **Figura 17 C** – Descoberta OTTs

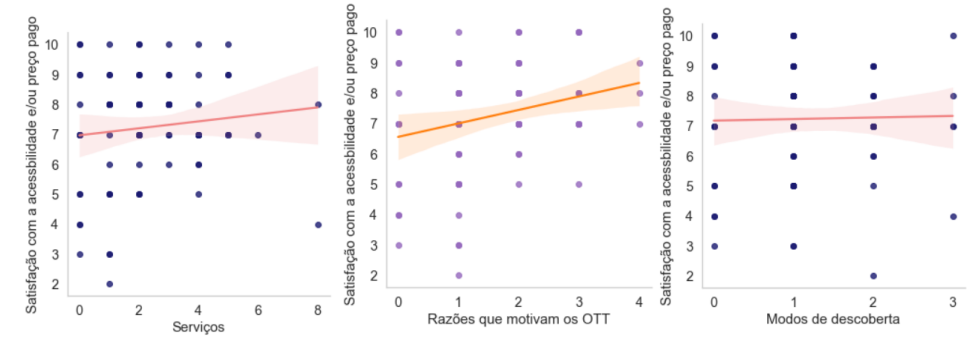
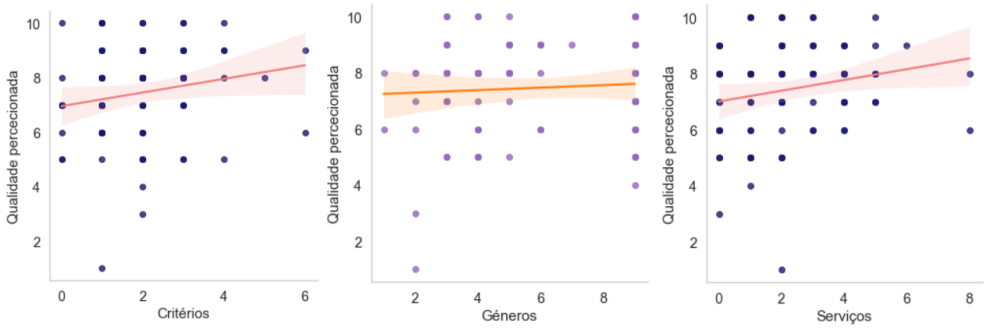
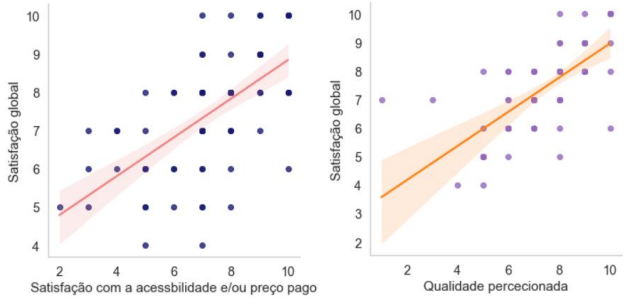


Figura 17 D - Critérios **Figura 17 E** - Géneros **Figura 17 F** – Serviços QP



Figuras 17: Satisfação global em função dos dois outros índices-alvo da análise

Figura 18 A – SAP **Figura 18 B** - QP



As representações gráficas acima espelham a exploração que a apresentar seguidamente. As variáveis y são as identificadas como índices principais entre todas as *features* numéricas e as variáveis x são as interpretadas logicamente como aquelas com mais probabilidade de estarem mais intrinsecamente relacionadas e servirem como melhores preditores nos modelos a construir. Para os gráficos deste tipo, a reta que os atravessa representa a minimização da soma dos quadrados dos desvios entre cada par de valores, *xi* e *yi*, ajustada pelo método dos mínimos quadrados. Existem relações lógicas para além daquelas escrutinadas presentemente, mas para efeitos de objectividade, estas serão o foco do estudo, a partir das quais será feito o confronto de valores entre os modelos de regressão linear e os de árvore de decisão, RF e SVM. Para os últimos 3 algoritmos, são utilizadas as variáveis binárias adicionadas anteriormente, derivadas das apresentadas na Tabela 11. Dado, através da RL, não ter sido possível encontrar estudos que incidam sobre esta temática e em simultâneo procurem construir modelos

com o intuito de prever a satisfação, qualidade percecionada ou índices similares, o mapeamento dos preditores não foi resultado daquele definido em anterios estudos, não existindo aqui uma plataforma de referência ou comparação.

Tabela 10: Mapeamento das relações lógicas entre as variáveis

ID Coluna	*Feature* a prever – Variável Dependente	ID Coluna	Preditor – Variável Independente	ID Modelo
21	Satisfação com a acessibilidade e/ou preço pago - SAP	15	Número de serviços subscritos ou usufruídos - NS	1
		17	Número de razões motivadoras para optar por serviços *OTT* - NR	
		19	Número de modos de descoberta - NM	
23	Qualidade percecionada - QP	9	Número de critérios aplicados para o que ver a seguir - NC	2
		12	Número de géneros consumidos - NG	
		15	Número de serviços subscritos ou usufruídos - NS	
25	Satisfação global - SG	9	Número de critérios aplicados para o que ver a seguir - NC	3
		12	Número de géneros consumidos - NG	
		15	Número de serviços subscritos ou usufruídos - NS	
		17	Número de razões motivadoras para optar por serviços *OTT* - NR	
		19	Número de modos de descoberta - NM	
		21	Satisfação com a acessibilidade e/ou preço pago - SAP	4
		23	Qualidade percecionada - QP	

4.6.1 Regressão linear

Foram construídos modelos e calculadas as principais métricas para cada um dos preditores identificados na Tabela 11, assim como para todos os relacionados em simultâneo. Nestas instâncias, é necessário definir uma variável adicional que liste os atributos mediante o seu encapsulamento. Cada modelo será separado nas amostras de treino e teste como método de validação cruzada, de modo a retornar a raiz quadrada do erro quadrático médio, confrontando com o original, como medida da qualidade do ajustamento.

Os resultados das Tabelas 12 e 13 não demonstram uma forte capacidade preditora das *features* numéricas que compõem a base de dados. Como expectável, os modelos para a satisfação global em função dos outros índices alvo, apresentam um forte ajustamento à variável dependente, e ainda mais com ambas qualidade percecionada e satisfação com a acessibilidade e/ou preço pago combinadas. Nestes modelos, os índices de erro têm os mínimos entre os valores tabelados, também válido após ajuste, através da separação da amostra em treino e teste.

Tabela 11: Resultados da aplicação dos modelos de regressão linear simples

Variáveis - ID		Expressão - Modelo	R^2	Erro Absoluto Médio	Erro Quadrático Médio	Raiz Quadrada do Erro Quadrático Médio	
						Original	Ajustado
Y = 21	X = 15	$\widehat{SAP}$ = [6.9673] + [0.1165] x NS p-value β_{NS} = 3.4408 e^{-20}	0.0134	1.371	3.1165	1.7654	1.7501
	X = 17	$\widehat{SAP}$ = [6.5578] + [0.4438] x NR p-value β_{NR} = 4.8689 e^{-25}	0.0651	1.3446	2.9533	1.7185	1.6829
	X = 19	$\widehat{SAP}$ = [7.1764] + [0.0517] x NM p-value β_{NM} = 7.1563 e^{-23}	0.0005	1.3662	3.1574	1.7769	1.7932
Y = 23	X = 9	$\widehat{QP}$ = [6.9545] + [0.25] x NC p-value β_{NC} = 1.5072 e^{-25}	0.0342	3.3901	13.7562	3.7089	1.6973
	X = 12	$\widehat{QP}$ = [7.2062] + [0.0446] x NG p-value β_{NG} = 4.8369 e^{-31}	0.0048	1.3705	2.8932	1.7009	1.7152
	X = 15	$\widehat{QP}$ = [7.0078] + [0.1918] x NS p-value β_{NS} = 2.8531 e^{-21}	0.0396	1.3137	2.7919	1.6709	1.6552
Y = 25	X = 9	$\widehat{SG}$ = [7.1289] + [0.1571] x NC p-value β_{NC} = 2.8245 e^{-25}	0.0186	1.1643	2.0711	1.4391	1.585
	X = 12	$\widehat{SG}$ = [7.2808] + [0.0292] x NG p-value β_{NG} =1.1883 e^{-31}	0.0028	1.1835	2.1045	1.4507	1.5867
	X = 15	$\widehat{SG}$ = [7.0698] + [0.1603] x NS p-value β_{NS} = 2.3014 e^{-21}	0.0381	1.1663	2.03	1.4248	1.5508
	X = 17	$\widehat{SG}$ = [6.9896] + [0.2957] x NR p-value β_{NR} = 1.0444 e^{-24}	0.0433	1.1409	2.0191	1.421	1.5198
	X = 19	$\widehat{SG}$ = [7.0513] + [0.3254] x NM p-value β_{NM} = 2.5854 e^{-25}	0.0298	1.1619	2.0475	1.4309	1.5427
	X = 21	$\widehat{SG}$ = [3.7796] + [0.5061] x SAP p-value β_{SAP} = 3.3681 e^{-64}	0.3834	0.877	1.3012	1.1407	1.2821
	X = 23	$\widehat{SG}$ = [2.9768] + [0.5991] x QP p-value β_{QP} = 1.3799 e^{-70}	0.4945	0.7935	1.0669	1.0329	1.0423

Para os modelos com apenas uma variável independente, em que esta não é uma das variáveis-alvo, evidencia-se a satisfação com a acessibilidade/preço em função das razões que motivam um utilizador a subscrever serviços *OTT*, que apresenta o maior coeficiente de determinação. Contudo, a satisfação global

em função do número de razões que motivam a escolha por *OTT*s, é o modelo que apresenta o menor erro absoluto médio, o que, intuitivamente, nos diz que quanto mais criterioso for um utilizador na escolha de uma modalidade de consumo, mais probabilidade terá de consumir conteúdo do seu agrado, assunção também válida considerando o menor erro quadrático médio e respectiva raiz quadrada, registados no mesmo modelo. O menor R^2 é registado no modelo que relaciona a satisfação com a acessibilidade e/ou preço pago ao número de modos de descoberta e os maiores índices de erro no modelo que projeta a qualidade percecionada em função do número de critérios utilizados para decidir o que ver a seguir, o que espelha que quanto mais é introduzido neste processo de tomada de decisão, existe maior tendência para a sua qualidade percecionada ser mais favorável.

Tabela 12: Resultados da aplicação dos modelos de regressão linear múltipla

Variáveis - ID		**Expressão - Modelo**	**R^2**	**Erro Absoluto Médio**	**Erro Quadrático Médio**	**Raiz Quadrada do Erro Quadrático Médio**	
Y = 21	X_1 = 15 X_2 = 17 X_3 = 19	$\widehat{SAP}$ = [6.7463] + [0.0675] x NS + [0.5799] x NR + [-0.4602] x NM p-value β_{NS} = 0.09285 p-value β_{NR} = 0.000126 p-value β_{NM} = 0.03804	0.0884	1.3238	2.8798	1.697	1.7529
Y = 23	X_1 = 9 X_2 = 12 X_3 = 15	$\widehat{QP}$ = [6.4484] + [0.2075] x NC + [0.0365] x NG + [0.1665] x NS p-value β_{NC} = 2.1979 e^{-05} p-value β_{NG} = 1.4339 e^{-10} p-value β_{NS} = 2.2896 e^{-04}	0.0667	1.2783	2.7132	1.6472	1.6442
Y = 25	X_1 = 9 X_2 = 12 X_3 = 15 X_4 = 17 X_4 = 19	$\widehat{SG}$ = [6.5006] + [0.1211] x NC + [0.0307] x NG + [0.0858] x NS + [0.2214] x NR + [-0.0083] x NM p-value β_{NC} = 7.985375 e^{-04} p-value β_{NG} = 5.998701 e^{-10} p-value β_{NS} = 3.082678 e^{-01} p-value β_{NR} = 2.405424 e^{-03} p-value β_{NM} = 5.928225 e^{-01}	0.0693	1.1303	1.9641	1.4015	1.5446
	X_1 = 21 X_2 = 23	$\widehat{SG}$ = [2.3892] + [0.2431] x SAP + [0.4419] x QP p-value β_{SAP} = 9.658 e^{-06} p-value β_{QP} = 3.3238 e^{-12}	0.5489	0.7294	0.952	0.9757	1.0943

Entre os modelos com múltiplas variáveis explicativas utilizadas, quando nenhuma delas é uma variável-alvo, a satisfação com a acessibilidade/preço em função do número de serviços, razões que motivam a opção por serviços *OTT* e modos de descoberta destes é o que apresenta um maior coeficiente de determinação, apesar de ser onde se registam os maiores valores de erro. Contudo a satisfação global em função do número de critérios, géneros, serviços, razões que motivam e modos de descoberta dos *OTT*s é onde se verificam os menores índices de erro e o modelo de projecção da qualidade percecionada é o que regista o menor R^2.

Relativamente ao ajuste pelo treino dos modelos, em 5 dos 17 modelos apresentados, a raiz quadrada do erro quadrático médio diminui, sendo a diferença na diminuição mais pronunciada na qualidade percecionada em função do número de critérios, para modelos com uma variável independente, e naqueles com múltiplas variáveis, apenas se verificou uma diminuição para a qualidade percecionada em função de todas as variáveis logicamente associadas.

Pelos *p-value* na regressão linear simples, é possível verificar que todos os coeficientes apresentam valores bastante próximos de 0, sendo que quando apenas uma variável preditora é usada, apresentam-se todas significativas na explicação das variações das variáveis dependentes. Nos modelos de regressão linear múltipla, é possível fazer uma análise comparativa entre todos os preditores, o que nos leva a concluir que no modelo que prevê a satisfação global em função das cinco variáveis independentes, o número de serviços e modos de descoberta dos *OTT*s não são significativas na explicação das variações da variável-alvo. A mesma disparidade de significância não é verificado para os preditores dos restantes modelos especificados.

4.6.2 Diagnóstico da multicolinearidade e homogeneidade

Foram calculados os *VIF*s para cada variável numérica logicamente relacionada na Tabela 11, com o intuito de determinar a multicolinearidade entre os preditores para os modelos de regressão linear múltipla. Adicionalmente, é calculada a homogeneidade entre as variáveis com as quais foram construídos os modelos-chave desta investigação, através do teste de adequação de Bartlett, com o objectivo de verificar se a variância dos erros aleatórios é constante para as *features* numéricas mapeadas.

Tabela 13: Apresentação dos resultados do diagnóstico

ID Modelo	Variáveis preditoras dos modelos	VIF	Teste de Bartlett	
			Estatística de teste	*P-Value*
1	Número de serviços subscritos	4.1061	62.8487	2.2520 e^{-14}
	Número de razões motivadoras das subscrições	4.4888		
	Número de modos de descoberta	5.2859		
2	Número de critérios aplicados	2.9444	45.994	1.0293 e^{-10}
	Número de géneros consumidos	2.9351		
	Número de serviços subscritos	2.4535		
3	Número de critérios aplicados	3.4175	158.2940	3.3945 e^{-33}
	Número de géneros consumidos	3.1157		
	Número de serviços subscritos	4.2201		
	Número de razões motivadoras das subscrições	4.6243		
	Número de modos de descoberta	6.3882		
4	Satisfação com a acessbilidade e/ou preço pago	24.8753	0.1493	0.6992
	Qualidade percecionada			
	Satisfação com a acessbilidade e/ou preço pago (Y/N)	2.1238	0.002583	0.9594
	Qualidade percecionada (Y/N)			

É possível verificar que o *p-value*, para as combinações de preditores do M1 a M3, está abaixo do nível de significância de 0.05, o que leva à rejeição da H0, na medida em que a variância não é igual para todas as variáveis ou grupo de valores. No que diz respeito aos *VIF*s obtidos, podemos verificar um problema de multicolinearidade para as variáveis do M4 usadas na regressão linear múltipla, antes da sua normalização, e para o número de modos de descoberta, preditor do M1 e M3, uma vez todas estas apresentam valores superiores a 5.

Para o M1 e M2, verifica-se que o número de serviços subscritos é o que apresenta menor correlação com os restantes preditores. No caso do M3, o menor valor de correlação regista-se no número de géneros. As correlações mais pronunciadas apresentam-se no número de razões que motivam a escolha pelos *OTT*s, para o M1 e M3, e no número de critérios aplicados, para o M2, estando ainda assim, abaixo do valor máximo, indicador de um potencial problema a nível da construção de modelos preditivos.

Uma vez que este é um problema a nível de regressão, no contexto do aplicação do algoritmo *CART* nas árvores de decisão, os modelos não irão incluir as variáveis preditoras para as quais se identificou a existência de multicolinearidade, sendo posteriormente confrontados os valores para as principais métricas calculadas, a saber, índices de erro. Por conseguinte, o M1 e M3 não incluem a variável "número de modos de descoberta" e o M4 utiliza as variáveis binárias, identificadas por "Y/N", e não as originais. Para efeitos de classificação, através da obtenção da matriz de confusão, o diagnóstico apresentado é inconsequente e não irá impactar os modelos especificados seguidamente, quer a nível dos métodos utilizados quer a nível das variáveis utilizadas.

4.6.3 Árvores de decisão

São, seguidamente, apresentados os modelos que agregam múltiplas variáveis independentes, já utilizados aquando da LR. Para cálculos de regressão é usada a função *DecisionTreeRegressor*, com os preditores a utilizar ajustados, como explicado anteriormente, e a função *DecisionTreeClassifier* para determinação das métricas que advêm da matriz de confusão. Foram aplicados diferentes graus de profundidade às árvores para cada modelo, o valor *default* sem qualquer definição de parâmetros e um ajustado, que facilitará a leitura e interpretação da árvore, realizando o "*tree pruning*". O valor ajustado foi fixado após realizar vários testes e ter-se chegado ao valor mínimo que maximiza o PECC e onde ainda se obtêm valores para todos os quadrantes da matriz de confusão, procurando assim maximizar a robustez da análise comparativa. Dito isto, a Tabela 15 apresenta valores para a profundidade padrão, sem acrescento de qualquer argumento. O algoritmo aplicado neste contexto é uma versão otimizada do *CART*, aquele suportado na biblioteca programática utilizada no contexto desta investigação, a saber *SciKit-Learn*. As variáveis dependentes a utilizar neste âmbito são as binárias, derivadas das variáveis-alvo originais, mapeadas na Tabela 11 como *features* a prever. Estas incluem a "Satisfação com a acessbilidade e/ou preço pago (Y/N)", ID = 22, a "Qualidade percecionada (Y/N)", ID = 24 e a "Satisfação global (Y/N)", ID = 26, com as alterações descritas ao nível dos preditores, dependendo do tipo de análise. O diagrama das ávores de decisão é apresentado no anexo deste documento para efeitos de estruturação do documento e de forma a facilitar a leitura ou consulta.

Tabela 14: Resultado dos cálculos de regressão para modelos de árvores de decisão

ID Modelo	Variáveis - ID		Erro Absoluto Médio	Erro Quadrático Médio	Raiz Quadrada do Erro Quadrático Médio
1	Y = 22	X_1 = 15	0.4317	0.2627	0.5125
		X_2 = 17			
2	Y = 24	X_1 = 9	0.5185	0.4815	0.6939
		X_2 = 12			
		X_3 = 15			
3	Y = 26	X_1 = 9	0.6111	0.6019	0.7758
		X_2 = 12			
		X_3 = 15			
		X_4 = 17			
4		X_1 = 22	0.2474	0.1062	0.3259
		X_2 = 24			

Globalmente, constata-se que os modelos apresentam um fraco ajustamento estimado às variáveis dependentes, à exceção do modelo 4, o que relaciona as variáveis-alvo binárias. É importante mencionar que aqui se removem os valores do R^2 do estudo comparativo e que, de acordo com a normalização, as escalas de 0 de 10 são transformadas. O modelo 1 é o que aqui apresenta o menor erro e o modelo 3, o maior sendo que as disparidades entre aplicação de algoritmos é verificada através do facto de que, na LR, o modelo 1 apresenta o maior erro e o modelo 3, o menor.

4.6.3.1 Modelo 1: Satisfação com a acessibilidade e/ou preço pago

Uma vez que para este modelo, a profundidade padrão já maximiza o PECC, esta definiu-se através do segundo maior valor de precisão, diretamente a seguir ao obtido inicialmente. Os valores 1 significam "Satisfeito", os 0 "Não Satisfeito".

Tabela 15: Principais métricas para o modelo 1

Modelos		Precisão	F1 score	PECC	Sensibilidade	Especificidade
Profundidade	**4**	0.6	0.3333	0.5556	0.2308	0.8571
	5	0.6667	0.4211	0.5926	0.3077	0.8571

Na sua forma padrão, dos participantes que o modelo prevê estarem satisfeitos com a acessibilidade e preço pago, 66% estão realmente. Isto aliado ao facto do *F1 Score* para ambas as soluções não ser óptimo, conseguimos verificar tanto por estes valores como pelos da Tabela 15, que a aplicação do modelo não é bem sucedida, nem as variáveis independentes fortes explicadoras da satisfação com a acessibilidade e/ou preço pago. O *tree pruning* nesta situação não produziu o efeito desejado, sendo que diminui a proporção de verdadeiros positivos e aumentou a de falsos negativos.

Verificando os nós "*leaf*", os finais da árvore que já não se separam, e os níveis de impureza de cada separação, começa a registar-se um padrão tendencial entre as características de consumo. No caso em que haja nós "*leaf*" com igual nível de impureza, verifica-se o nó "decisional" diretamente acima. Utilizadores tendencialmente satisfeitos com a acessibilidade e ou preço/pago, descobriram serviços de subscrição por intermédio de menos de 1.5 métodos de descoberta, subscrevem entre 0.5 a 2.5 serviços e identificam mais do que 1.5 razões que motivam estas subscrições, expresso no nó "*leaf*" 3, a contar da esquerda. Em contrapartida, utilizadores não satisfeitos, diferem na medida em que subscrevem entre 3.5 a 4.5 serviços não havendo qualquer diferença a nível dos restantes preditores, expresso no nó "*leaf*" 5, a contar da esquerda. Nesta instância, o preço poderá ter tido um papel mais determinante neste grau de satisfação, na medida em que mais serviços são sinónimo de maior montante dispendido em conteúdo de entretenimento, salvaguardando o modo como acedem a esse conteúdo, via plataformas que pirateiam o conteúdo ou por métodos legítimos, mas por intermédio de contas de pessoas próximas e obtenção de descontos através da inclusão num plano familiar.

4.6.3.2 Modelo 2: Qualidade percecionada

A partir do nível de profundidade 5, o PECC é maximizada para este modelo, sendo que foram confrontados os valores para os níveis 5 e 6, que tem o valor de precisão diretamente inferior ao anterior. Para a representação gráfica do modelo, apresenta-se a árvore para nível de profundidade 4 de modo a

facilitar a consulta, uma vez que o valor do PECC é igual. Os valores 1 neste contexto significam "O utilizador perceciona que o conteúdo disponibilizado atualmente tem qualidade" e valores 0 o inverso.

Tabela 16: Principais métricas para o modelo 2

Modelos		Precisão	F1 score	PECC	Sensibilidade	Especificidade
Profundidade	**5**	0.5	0.4828	0.4444	0.4667	0.4167
	6	0.4615	0.4286	0.4074	0.4	0.4167

O resultado do "*tree pruning*" nesta situação é vantajoso, na medida em que apenas a especificidade se mantém imutável, sendo que todos os outros indicadores aumentam. O modelo ajustado tem mais 4% de capacidade de prever que um participante considera que o conteúdo de entretenimento atual tem qualidade, expresso na métrica da precisão e há uma menor presença de falsos negativos, demonstrado na sensibilidade e PECC.

Aplicando a mesma lógica descrita anteriormente, os participantes que consideram haver qualidade no conteúdo produzido atualmente têm entre 1.5 a 2.5 serviços subscritos, menos de 6.5 géneros consumidos e usam menos de 0.5 critérios para decidir o que ver a seguir, expresso no nó "*leaf*" 4 final, a contar da direita. Por outro lado, os participantes com opinião contrária têm entre 2.5 a 7 géneros consumidos, aplicam mais de 2.5 critérios para decidir o que ver a seguir e subscrevem ou usufruem de menos de 1.5 serviços. Constata-se que os utilizadores mais deliberados e criteriosos na suas decisões do que ver e subscrever, têm uma perceção menos favorável do conteúdo com o qual interagem.

4.6.3.3 Modelo 3: Satisfação global em função dos preditores selecionados

Neste modelo, o PECC é maximizado na profundidade 6, sendo que o segundo valor mais alto é atingido na profundidade 7, constituindo assim o confronto de valores a fazer seguidamente. À semelhança do modelo 1, valores 1 significam que o participante está satisfeito e 0 o contrário.

Tabela 17: Principais métricas para o modelo 3

Modelos		Precisão	F1 score	PECC	Sensibilidade	Especificidade
Profundidade	**6**	0.4118	0.4516	0.3704	0.5	0.2308
	7	0.3571	0.3571	0.3333	0.3571	0.3077

O exercício de "*tree pruning*" demonstrou-se vantajoso, na medida em que o PECC não foi maximizado na solução padrão. Confrontando as duas soluções apresentadas, para a profundidade 6 é

onde se regista a menor proporção de falsos negativos e o modelo tem maior capacidade de prever os participantes satisfeitos globalmente.

Os participantes satisfeitos globalmente consumem menos de 4.5 géneros, usufruem de menos de 4.5 serviços de subscrição, utilizam entre 1.5 a 3.5 critérios para decidir o que ver a seguir e identificam mais de 1.5 razões para usufruir de *OTT*s. Por outro lado, os participantes não satisfeitos, tendencialmente, consumem menos de 7.5 géneros, utilizam entre 1.5 a 2.5 critérios para decidir o que devem ver a seguir, descobriram serviços *OTT*s através de menos de 0.5 meios e identificam menos de 1.5 razões para usufruir de *OTT*s. Ao contrário do verificado no modelo 2, os utilizadores mais satisfeitos aqui são os que diversificam menos o seu repertório de conteúdo, em termos de géneros consumidos, e aplicam mais critérios quando confrontados com a decisão do que ver a seguir.

4.6.3.4 Modelo 4: Satisfação global em função das outras variáveis-alvo

Da mesma forma que o modelo anterior, o PECC é já ótimo para a solução padrão, de profundidade 4 sem manipulação da profundidade máxima atribuída. Os valores irão ser confrontados com o valor diretamente abaixo, neste caso para profundidade 2, sendo o diagrama da árvore representado para a solução padrão.

Tabela 18: Principais métricas para o modelo 4

Modelos		Precisão	F1 score	PECC	Sensibilidade	Especificidade
Profundidade	**2**	0.7857	0.7857	0.7778	0.7857	0.7692
	4	0.8462	0.8148	0.8148	0.7857	0.8462

O "*tree pruning*" não conferiu os resultados pretendidos para este caso, registando-se um aumento da proporção de falsos positivos e uma menor capacidade para prever que os participantes estão de facto satisfeitos. O diferencial diminuto do PECC é devido ao igual número dos falsos negativos e imutabilidade da sensibilidade.

Os participantes globalmente satisfeitos têm uma qualidade percecionada superior a 9 e uma satisfação com a acessibilidade e/ou preço pago entre 7.5 a 8.5. Os participantes do outro lado do espectro apresentam uma qualidade percecionada entre 6.5 a 7.5 e uma satisfação com a acessibilidade e/ou preço pago menor que 6. Valores dentro do intervalo expectável e que enfatizam a importância de estabelecer a dicotomia para as variáveis em utilização nesta secção no valor 7, em detrimento do valor intermédio da escala definida no questionário.

4.6.4 *Random forest*

Na utilização deste modelo analítico, foi usado o número *default* de estimadores, igual a 100, e relativamente aos comandos utilizados, apenas diferem na função utilizadas para o classificador, o *RandomForestClassifier*, assim como no parâmetro *stratify* que não é aqui utilizado na separação da amostra, sendo que o restante processo de construção do modelo permanece inalterado. Este método agrega múltiplos algoritmos, que resultam na construção de múltiplas árvores de decisão e irá também ser aplicado a nível de classificação.

Tabela 19: Principais métricas para os diferentes modelos através de *random forest*

ID Modelo	Precisão	F1	PECC	Sensibilidade	Especificidade
1	0.5	0.4	0.5556	0.3333	0.7333
2	0.6667	0.6250	0.5556	0.5882	0.5
3	0.4286	0.4	0.3333	0.3750	0.2727
4	1	0.5455	0.6296	0.3750	1

Comparando os valores estimados através da matriz de confusão, para o M1, o modelo construído através deste algoritmo têm menor capacidade de prever que os participantes estão satisfeitos e todos os valores são inferiores exceto a sensibilidade pela menor presença de falsos negativos. Para o M2, temos a situação quase inversa em que todos os valores são superiores. Para o M3, o modelo *RF* tem melhor capacidade preditiva e todos os valores são superiores exceto a especificidade, pela maior proporção de falsos positivos. A maior discrepância apresenta-se no M4, em que os resultados não vão ao encontro do pretendido de modo a conseguir-se fazer um confronto de valores, uma vez que o modelo não retorna qualquer falso positivo, depois de vários testes realizados aquando da construção do modelo, quer a nivel da manipulação do número de estimadores, tamanho da amostra e parâmetro não determinístico da função, aquele que assegura que os mesmos valores são produzidos a cada instância de execução. A precisão e especificidade apresentam valores máximos, sendo as restantes métricas inferiores, dada a presença de mais falsos positivos.

4.6.5 *Support vector machine*

Aplicando a mesma lógica programática que nos dois anteriores tipos de modelação, foi seguidamente utilizado um método de classificação linear binária, que procura minimizar o erro de generalização. Este algoritmo necessita da identificação de um *kernel*, que transforma os dados de *input* em determinadas formas, criando uma dimensão de *output* maior. Originalmente, foi usado o *kernel* "*Radial Basis Function*", o mais adequado em termos de classificação. A única variante introduzida por este modelo é a "normalização dos dados", que involve transformar os valores das *features* numéricas em intervalos entre 0 e 1, convertendo o *kernel* para "Linear", de forma a potencialmente otimizar os modelos, sendo que ambas as soluções serão apresentadas.

Tabela 20: Principais métricas para os diferentes modelos através de *support vector machine*

Fase	ID Modelo	Precisão	F1	PECC	Sensibilidade	Especificidade
Pré-Normalização	1	0.375	0.2857	0.4444	0.2308	0.6429
	2	0.5	0.5455	0.4444	0.6	0.25
	3	0.5	0.5882	0.4815	0.7143	0.2308
	4	0.8	0.8276	0.8148	0.8571	0.7692
Normalizado	1	0.8	0.4444	0.6296	0.3077	0.9286
	2	0.65	0.7429	0.6667	0.8667	0.4167
	3	0.5556	0.625	0.5556	0.7143	0.3846
	4	0.65	0.7647	0.7037	0.9286	0.4615

O exercício de normalização demonstrou-se vantajoso para todos os modelos, à exceção do M4, em que apenas aumentou a sensibilidade e especificidade, tendo-se registado uma diminuição da precisão e PECC. Ao estabelecer uma comparação entre os modelos normalizados e os de árvores de decisão, regista-se aqui um aumento de todas as métricas calculadas através da matriz para os modelos M1 a M3, com a única igualdade no M3 ao nível da sensibilidade. No M4, apenas o valor da sensibilidade é maior, dada a menor presença de falsos negativos.

Ao realizar a mesma análise mas comparando com os modelos *random forest*, à semelhança do confronto anterior, no M1 regista-se um aumento de todos valores excepto para a sensibilidade. Para o M2 a M4, o modelo SVM apresenta menor precisão e especificidade, ainda que globalmente, o PECC tenha aumentado.

4.7 Resultados relevantes dos artigos revistos e validação das intenções de estudo

O esforço de network e performance de serviços *OTT* foi calculado, com recurso ao *Scikit-learn* e aos algoritmos de *machine learning J48*, *KNN*, *PART* e *SVM*, tendo-se evidenciado este último por apresentar os valores mais elevados para sensibilidade, precisão e *F1 Score* (Choi & Kim, 2021). Apesar de não ser possível fazer um paralelismo direto dada a natureza dos dados do presente estudo e algoritmos utilizados, verificou-se que os valores apresentados para *SVM* são tendencialmente superiores aos restantes, à exceção daqueles obtidos no M4.

A satisfação ao nível do lazer foi estimada com recurso a modelos de regressão linear múltipla e diagnósticos de correlação entre as diferentes variáveis. Foi demonstrado que fatores sócio-demográficos têm pouco poder preditivo e circunstâncias de *stress* e saúde são preditores bastante mais eficazes e impactantes para as projecções de satisfação no âmbito do lazer (Chick et al., 2020).

Apesar de existir na amostra, um segmento de participantes, de representatividade superior a 60%, que apresenta caraterísticas socio-demográficas similares e a mesma modalidade de consumo

predominante elegida, são detetadas algumas diferenças nas caraterístias de consumo consoante a modalidade predominante dos participantes. Os participantes que usufruem de serviços *OTT* consumem menos géneros que os que maioritariamente vêm televisão por cabo e especificamente os que pagam por *OTT*s apresentam valores médios superiores para os principais índices de análise, satisfação e qualidade. As distribuições de valores demonstram também que a qualidade percecionada e satisfação global são consistentemente superiores para quem paga por *OTT*s, pelas menores amplitudes e valores superiores dos quartis 1 e 3.

A aplicação dos algoritmos de *machine learning* e variáveis empregues nos respectivos modelos demonstram, através dos valores obtidos, uma falta de capacidade preditiva e ajustamento dos modelos às respectivas variáveis dependentes. Isto é válido quer para os cálculos feitos através de regressão como para classificação, expressos por R^2 ou índices de erro e PECC, respectivamente. O diagnóstico de multicolinearidade e correlações de Pearson representadas no *heatmap* permitem verificar um problema para o caso do número de modos de descoberta no M1 e M3 e no caso das variáveis-alvo, quando relacionadas no M4.

5 Conclusões

Este estudo propôs-se determinar o grau de satisfação e qualidade percecionada dos consumidores de conteúdo televisivo e cinematográfico consoante os seus hábitos, preferências e características de utilização. Para o efeito, houve um particular foco nos serviços de subscrição nas diferentes análises efetuadas, com vista também a perceber o seu impacto quer em termos da satisfação quer em termos da qualidade percecionada. Adicionalmente, houve intenção de analisar que tipos de utilizador mostram maior incidência numa determinada modalidade de consumo. Em última instância, o intuito passou por verificar que projeções da adequação do conteúdo consumido poderiam ser feitas mediante as diferentes caraterísticas de consumo, que dizem respeito ou não, a serviços subscritos. A significância dos resultados demonstrou-se inferior ao inicialmente esperado e delineado quer na análise exploratória quer na modelação preditiva mas foi possível apurar alguns indicadores importantes sobre as caraterísticas de consumo de conteúdo de entretenimento.

Ao nível da ADE, os participantes revelam assertividade pouco pronunciada para decidir o que ver a seguir e uma grande amplitude de variabilidade do conteúdo consumido, expresso no número médio de critérios e géneros, respectivamente. É de notar que a escolha pela subscrição ou usufruto de serviço é normalmente acompanhada por uma outra, expresso na média do número de *OTT*s, o que significa que esta modalidade de consumo poderá também ser uma solução sub-ótima no que diz respeito ao consumo de entretenimento, uma vez que obriga o utilizador a interagir com várias plataformas e a gastar mais dinheiro.

Os participantes desta amostra são maioritariamente jovens adultos, entre os 20 e 30 anos de idade, do sexo masculino, a auferir rendimentos entre os 1000 e 2000 euros, com interesse por conteúdo cinematográfico e televisivo e são estes os consumidores-alvo de serviços de subscrição e em linha com a evidência apresentada, nos vários estudos consultados aquando da revisão da literatura. Entre outras caraterísticas distintivas, verifica-se que a televisão, ainda é o dispositivo primordial para consumo, apesar de existir o argumento da acessibilidade acrescida dos novos serviços, referido ao longo deste trabalho, que possibilita a utilização de outros dispositivos. Ao nível dos principais indicadores-alvo da análise, destaca-se o nível de máximo da qualidade percecionada para "serviços de subscrição" e a satisfação com a acessibilidade e/ou preço pago para a "televisão por cabo".

No que diz respeito à modelação preditiva, não é possível aferir que as *features* numéricas que compõem a base de dados, mapeadas logicamente para a aplicação dos diferentes algoritmos, sejam preditores adequados o suficiente para significativamente influenciar quer a qualidade percecionada como as satisfações, com a acessibilidade e/ou preço pago e global. Os cálculos de regressão para a LR foram os que apresentaram valores mais explicativos da influência dos preditores, expresso pelos maiores coeficientes de determinação e erros proporcionalmente mais baixos, tanto em termos absolutos como quadráticos, ainda que distanciados de valores que se podem considerar como tendo um efeito considerável na variável-alvo. A nível das métricas deduzidas das matrizes de confusão, os resultados são mistos em termos da relação algoritmo-modelo, utilizando a precisão e PECC como elementos que

que desempatam a decisão de qual a solução mais ilustrativa. Os modelos elaborados via *support vector machine* são os que apresentam melhor capacidade preditiva, após a normalização dos dados, excetuando o M4, em que o valor máximo da PECC foi atingido via *CART*. Para o M1, M3 e M4, os valores mínimos da PECC foram obtidos via *random forest* e para o M2 via *CART*. Globalmente, estes resultados não permitem consentâneamente afirmar qual o algoritmo mais bem sucedido a fazer previsões consoante as variáveis contempladas. Os algoritmos *random forest* e *CART* apresentam a mesma comunalidade relativamente ao modelo com melhor ajustamento à respetiva variável dependente, sendo a satisfação com a acessibilidade e/ou preço pago a mais intrinsecamente influenciada pelos respectivos preditores, número de serviços, razões que motivam o usufruto e modos da sua descoberta. Contrariamente, a satisfação global é menos influenciada pelo número de critérios por que se rege o utilizador para decidir o que ver a seguir, número de géneros consumidos, número de serviços usufruídos razões que motivam o usufruto e modos da sua descoberta. Contudo, nenhum dos resultados apresentados permitem categoricamente afirmar que os preditores selecionados têm uma forte relação às variáveis-alvo.

5.1 Contributos

Não obstante os resultados obtidos e destacados anteriormente, defende-se a premissa sobre qual assenta este estudo. A investigação poderá servir como ponto de referência basilar para outras investigações desta natureza, que tenham a ambição de decompôr as caraterísticas de consumo de um segmento da população e qual o impacto destas na satisfação dos utilizadores. Para as entidades distribuidoras de conteúdo televisivo e cinematográfico, a intenção foi de fornecer indicadores importantes baseados nestas caraterísticas que pudessem auxiliar na definição da sua estratégia, nomeadamente, ao nível da produção pela maior incidência num certo tipo de género ou nível do *marketing*, percebendo qual o modo de descoberta de serviços de subscrição mais prevalente. A pertinência advém da assunção de que o conteúdo disponibilizado atualmente é diverso e muito dele de elevado nível de qualidade e produção. Isto é ilustrado pela série televisiva "*Rings of Power*" da Amazon, que se estima ter um orçamento alocado de 500 milhões de doláres, apenas para a primeira temporada, ainda que a sua qualidade seja altamente discutível (Weaver, 2022). Porém, interpreta-se que os utilizadores parecem agora mais insatisfeitos e exigentes do que anteriormente, pré-proliferação dos novos serviços de subscrição. Esta interpretação é suportada por fénomenos como o *"review bombing"*, que consiste nos utilizadores deixarem críticas desfavoráveis nos *websites* de agregação e apresentação de *ratings* como o *IMDB* ou o *Rotten Tomatoes*, mesmo antes de sequer terem assistido a um determinado conteúdo, aquando do seu lançamento. Pode ser feito o argumento de que este comportamento é fruto destas plataformas existirem e terem um nível de popularidade e alcance considerável mas é também possível a explicação ter mais nuances.

A abordagem metodológica assumida, no que diz respeito ao software utilizado, considera-se uma alternativa viável à utilização de programas como o *SPSS Modeler* ou *Statistics*, apesar de ser o caminho

de maior fricção, na medida em que tudo o que é produzido tem de ser escrito ao contrário destas plataformas, que já apresentam as opções necessárias para a realização de determinada análise. Contudo, o método eligido confere um maior nível de personalização e o utilizador não se encontra confinado a certo *template*, estrutura ou número de opções.

5.2 Limitações e futuras pistas de investigação

Para além das limitações já enaltecidas na metodologia, são evidentes as restantes áreas onde a presente investigação poderia ter sido aprofundada. A amostra de participantes contemplada, apesar de não ter tido grande influência na exploração da análise de dados, não é expressiva. Na secção que enfatiza a ADE, interpreta-se que muitos dos resultados produzidos seriam passíveis de grande alterações mediante uma amostra mais numerosa. Isto deve-se ao facto do menor número de participantes que selecionaram "Cinema" e "Televisão por cabo" na categoria "Modalidade de consumo predominante", não ter permitido um confronto com maior fiabilidade. Ao nível da modelação preditiva, o número de registos contemplados em cada um dos algoritmos aplicados está acima do mínimo praticado pelas "*rules of thumb*" convencionadas, mas é também possível verificar a falta de expressividade. A amostra contemplada e o facto da distribuição do questionário não se reger por qualquer princípio previamente definido obrigou ainda a um esforço acrescido na fase de normalização, no sentido de aprimorar a adequabilidade dos dados.

A revisão da literatura não se revelou uma plataforma de referência para o que é a realização de uma investigação que incide sobre esta temática no âmbito da presente área científica. Foi resumido principalmente o panorama de consumo atual e o efeito disruptivo dos serviços de subscrição emergentes, o que foi deficitário e em retrospectiva, acabou por não pautar muito do que foi o desenvolvimento deste trabalho, sobretudo no que diz respeito aos princípios metodológicos.

Os valores obtidos na última secção do trabalho são difíceis de explicar na medida em que se podem dever a uma pluralidade de fatores, nomeadamente, modo de construção dos modelos, os algoritmos contemplados e a formulação de perguntas e respectiva transformação das colunas da base de dados em preditores. Este último ponto merece particular destaque para instâncias de investigação futuras, que deverão dar mais incidência em perguntas que se traduzam de forma bem sucedida para *features* numéricas com maior capacidade preditora.

Futuros estudos, nos moldes da presente investigação, para alem de extender o portfólio de *features* numéricas a utilizar, deverão alargar a amostra inquirida a várias regiões ou países, de modo a confrontar onde se registam os maiores ou menores valores de satisfação ou qualidade percecionada, e quais os preditores mais influenciadores. O *rating* dos conteúdos televisivos deverá também ser considerado como preditor destas variáveis-alvo, uma vez que se encontra tão latente a um determinado conteúdo, aquando do seu lançamento.

Referências bibliográficas

Alcolea-Diaz, G. Marin-Lladó, C. Cervi, L. (2021). Expansion of the core business of traditional media companies in Spain through SVOD services. *Communication & Society,* 35(1): 163-175.

Alforova, Z. Marchenko, S. Kot, H. Medvedieva, A. Moussienko, O. (2021). Impact of Digital Technologies on the Development of Modern Film Production and Television. *Rupkatha Journal on Interdisciplinary Studies in Humanities,* 13 Nº 4: 1-11.

Awan, M. Khan, R. Nobanee, H. Yasin, A. Anwar, S. Naseem, U. Singh, V. (2021). A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. *Electronics*, 10, 1215.

Camilleri, M. Falzon, L. (2020). Understanding motivations to use online streaming services: integrating the technology acceptance model (TAM) and the uses and gratifications theory (UGT). *Spanish Journal of Marketing,* 25(2): 217-238.

Chen, Y. (2019). Competitions between OTT TV platforms and traditional television in Taiwan: A Niche analysis. *Telecommunications Policy*, 43: 101793.

Chick, G. Dong, E. Yeh, C. Hsieh, C. (2020). Cultural consonance predicts leisure satisfaction in Taiwan. *Leisure Studies*, 40(2): 183-198.

Choi, J. Kim, Y. (2021). A Heterogeneous Learning Framework for Over-the-Top Consumer Analysis Reflecting the Actual Market Environment. *Applied Sciences*, 11, 4783.

Dang, C. Moreno-García, M. De La Prieta, F. (2021). Using Hybrid Deep Learning Models of Sentiment Analysis and Item Genres in Recommender Systems for Streaming Services. *Electronics*, 10(20), 2459.

Duan, J. Gao, R. (2021). Research on English Movie Resource Information Mining Based on Dynamic Data Stream Classification. *Security and Communication Networks*, Wiley, Hindawi.

Gambarato, R. Heuman, J. Lindberg, Y. (2021). Streaming media and the dynamics of remembering and forgetting: The Chernobyl case. *Memory Studies*, 15(2): 271-286.

Griffin, D. (2022). The State of Streaming in 2022: Netflix's Decline, HBO Max's Future, and More! *IGN*.

Gutzeit, J. Dorsch, I. Stock, W. (2021). Information Behavior on Video on Demand Services: User Motives and Their Selection Criteria for Content. *Information*, 12, 173.

Habib, S. Hamadneh, N. Hassan, A. (2022). The Relationship between Digital Marketing, Customer Engagement, and Purchase Intention via OTT Platforms. *Journal of Mathematics*, 1-12, Hindawi.

Herbert, D. Lotz, A. Marshall, L. (2018). Approaching media industries comparatively: A case study of streaming. *International Journal of Cultural Studies*, 22(3): 349-366.

Holmes, C. (2022). 'Too much free time won't make you happier', says psychologist – how many hours you really need in a day. *CNBC MakeIt*

Howarth, J. (2022). 74+ Shocking Women in Tech Statistics. *Exploring Topics.*

Jang, M. Baek, H. Kim, S. (2021). Movie characteristics as determinants of download-to-own performance in the Korean video-on-demand market. *Telecommunications Policy*, 45(1): 102140.

Lee, S. Lee, S. Joo, H. Nam, Y. (2021). Examining Factors Influencing Early Paid Over-The-Top Video Streaming Market Growth: A Cross-Country Empirical Study. *Sustainability*, 13, 5702.

Matos, M. Ferreira, P. Smith, M. (2017). The Effect of Subscription Video-on-Demand on Piracy: Evidence from a Household Level Randomized Experiment. *Management Science*, 64(12): 5610-5630.

McKenzie, J. Crosby, P. Cox, J. Collins, A. (2019). Experimental evidence on demand for "on-demand" entertainment. *Journal of Economic Behavior and Organization*, 161: 98-113.

Medina, M. Herrero, M. Portilla, I. (2019). The evolution of the pay TV market and the profile of the subscribers. *Revista Latina de la Comunicación Social*, 74: 1761-1780.

Nagaraj, S. Singh, S. Yasa, V. (2021). Factors affecting consumers' willingness to subscribe to over-the-top (OTT) video streaming services in India. *Technology in Society*, 65: 101534.

Nijhawan, G. Dahiya, S. (2020). Role of Covid as a catalyst in increasing adoption of OTTs in India: A study of evolving consumer consumption patterns and future business scope. *Journal of Content, Community & Communication*, 13: 298-311.

Pisarek, J. Zabielska-Mendyk, E. (2022). Film preferences in the pandemic: psychological resilience or an escape from the pandemic reality? *Annals of Psychology*, 14 3-4: 227-241.

Stoll, J. (2022). Most popular digital original series genres based on audience demand in the United States in 2021. *Statista – The Statistics Portal for Market Data, Market Research and Market Studies*. https://www.statista.com/statistics/715161/most-in-demand-tv-genre-in-north-america/#statisticContainer

Suwarto, D. Setiawan, B. Adikara, G. (2021). The Fragmentation of Indonesian Film Audience. *Jurnal Komunikasi: Malaysian Journal of Communication*, 37(1): 74-87.

Udoakpan, N. Tengeh, R. (2020). The Impact of Over-the-Top Television Services on Pay-Television Subscription Services in South Africa. *Journal of Open Innovation: Technology, Market and Complexity*, 6(4), 139.

Weaver, J. (2022). Rings of Power could cost $1B. What's making TV so expensive? *CBC – Canada's Public Broadcaster for Radio and Television.* https://www.cbc.ca/news/entertainment/rings-of-power-expensive-1.6571965

Anexos

A - Comandos introduzidos no *jupyter notebook*

Comando 1: Importação das bibliotecas basilares necessárias para a prossecução das tarefas

import numpy as np import pandas as pd

Comando 2: Adição de opções para expandir a apresentação dos dados nas células de *output*

pd.options.display.max_rows = 100 pd.options.display.max_columns = 30 pd.options.display.max_colwidth = 100

Comando 3: Carregamento do *dataset*-objecto de estudo

df = pd.read_csv('Serviços de Entretenimento.csv')

Comando 4: Listagem das colunas originais para efeitos de contextualização

for col in df.columns: print(col)

Comando 5: Contagem do número de registos

num_rows = len(df.index) print('Número de registos:', num_rows)

Comando 6: Eliminação de colunas excendentárias que não acrescentam valor à análise

df.drop(['Timestamp', 'Nível de interesse neste questionário, da maneira como foi formulado?', 'Reparos e oportunidade de melhoria'], axis=1, inplace=True)

Comando 7: Renomeamento das colunas consoante a listagem original para facilitar a leitura

df = df.rename(columns={Coluna original: Coluna renomeada,...})

Comando 8: Listagem do número de registos nulos e identificação da tipologia de dados

df.isnull().sum() df.dtypes

Comando 9: Frequência da opção "Nenhum/Terminar *Survey*" selecionada na Q4

nenhum_interesse_registado_count = df.apply(lambda x : True if x['Interesses'] == "Nenhum/Terminar Survey" else False, axis = 1) num_rows_nenhum_interesse_registado = len(df[nenhum_interesse_registado_count == True].index) print('Número de registos em que não se regista interesse:', num_rows_nenhum_interesse_registado)

Comando 10: Eliminação das respectivas entradas identificadas no comando 9

interesses_indexes = df[df['Interesses'] == 'Nenhum/Terminar Survey'].index

df.drop(interesses_indexes, inplace=True)

Comando 11: Frequência das opções "Televisão por cabo" ou "Cinema" selecionadas na Q12

opções_serviços = ["Televisão por cabo","Cinema"] contagem = df["Modalidade de consumo predominante"].value_counts() print('Número de registos com valores nulos:', contagem[opções_serviços].sum())

Comando 12: Substituição dos valores nulos

df[['Serviços subscritos e/ou usufruídos', 'Razões que motivam a opção por serviços de subcrição', 'Modos de descoberta de serviços de subscrição']] = df[['Serviços subscritos e/ou usufruídos', 'Razões que motivam a opção por serviços de subcrição', 'Modos de descoberta de serviços de subscrição']].fillna('Não prioritiza serviços de subscrição')

Comando 13: Conversão da tipologia das últimas 3 colunas

df['Satisfação com a acessbilidade e/ou preço pago'] = df['Satisfação com a acessbilidade e/ou preço pago'].astype('int64') df['Qualidade percecionada'] = df['Qualidade percecionada'].astype('int64') df['Satisfação global'] = df['Satisfação global'].astype('int64')

Comando 14: Adição de variáveis dicotómicas

nome_variável_conditions = [(df['Respectiva coluna'] <= 7), (df['Respectiva coluna'] > 7)] nome_variável _values = ['0','1'] df[‘Respectiva coluna' (Y/N)'] = np.select(nome_variável _conditions, nome_variável _values).astype('int64')

Comando 15: Adição de novas colunas e conversão da tipologia de dados

df[‘Coluna nova'] = np.where(df[‘Coluna original’].str.contains('Valor condicional'), Valor na coluna nova, df[‘Coluna original’].str.split(';').str.len()).astype('int64')

Comando 16: Reordenação lógica das colunas

df = df.reindex(columns=[‘Colunas’])

Comando 17: Sumário estatístico das variáveis numéricas

df.describe().transpose()

Comando 18: Importação das bibliotecas necessárias para a análise de dados exploratórios

import matplotlib.pyplot as plt import seaborn as sns %matplotlib inline

Comando 19: Construção de *pie-charts* para caracterização da amostra

df[‘Coluna’].value_counts().plot(kind='pie', labeldistance=None, autopct="%.1f%%", fontsize=12, figsize=(4,4), pctdistance = 0.8) plt.legend(loc="upper right", bbox_to_anchor=(0,0), loc="lower left", bbox_transform=plt.gcf().transFigure) plt.ylabel("Coluna", loc="top", rotation=0)

Comando 20: Cálculo das métricas a apresentar na análise de dados exploratórios

df[‘Coluna’] = df[‘Coluna’].str.split(';') nome_variável = df.groupby('Modalidade de consumo predominante')[‘Coluna’].agg(pd.Series.mode)/.mean() print(nome_variável)

Comando 21: Representação gráfica em barras para *features* de múltipla seleção e *boxplot*

plt.figure(figsize=(10,5)) plt.xticks(rotation=0)

df[‘Coluna’].str.split(';').explode(‘Coluna’).value_counts().sort_index().plot(kind='barh', color=['tab:blue', 'tab:orange',

'tab:green', 'tab:red', 'tab:purple', 'tab:brown', 'tab:pink', 'tab:gray', 'tab:olive', 'tab:cyan']) plt.grid(False) fig, ax = plt.subplots(figsize=(8, 5)) graph_nome_coluna = sns.boxplot(y=coluna y, x='Modalidade de consumo predominante', data=df, ax=ax, width=0.5, palette='colorblind') plt.xlabel('') plt.ylabel('')

Comando 22: Construção do mapa de verificação de relações entre as *features* numéricas

fig, ax = plt.subplots(figsize=(10, 5)) sns.heatmap(df.corr(), annot=True, ax=ax) plt.xticks(rotation=45, rotation_mode='anchor', ha='right')

Comando 23: Construção dos gráficos *lineplot*

sns.lmplot(x=coluna x, y=coluna y, data=df, scatter_kws={'color': 'tab:blue'}, line_kws={'color': 'tab:red'}) sns.set_style('whitegrid')

Comando 24: Teste de adequação e diagnósticos

import scipy.stats as stats import statsmodels.api as api from statsmodels.stats.outliers_influence import variance_inflation_factor stats.bartlett('Colunas') colunas_fatores = df[['Colunas']] vif_dados = pd.DataFrame() vif_dados["Feature"] = colunas_fatores.columns vif_dados["VIF"] = [variance_inflation_factor(colunas_fatores.values, i) for i in range(len(colunas_fatores.columns))] print((vif_dados).round(4)) p_values = modelorl.summary2().tables[1]['P>|t|']

Comando 25: Importação das bibliotecas afetas à regressão linear

from sklearn.linear_model import LinearRegression from sklearn import metrics from sklearn.model_selection import train_test_split

Comando 26: Construção dos modelos de regressão linear e respectiva apresentação

variávelx_rl = df.iloc[:,posição coluna x].values.reshape(-1, 1) variávely_rl = df.iloc[:,posição coluna y].values.reshape(-1, 1) rl = LinearRegression() modelo_rl = rl.fit(variávelx_rl, variávely_rl) print("variável y =", rl.intercept_.round(4), "+", rl.coef_[0].round(4), "x variável x")

Comando 27: Cálculo das principais métricas dos modelos preditivos

modelo_rl.score(variávelx_rl, variávely_rl).round(4) variávely_observada = variávely _rl variávely _prevista = modelo_rl.predict(variávelx _rl) metrics.mean_absolute_error(variávely_observada, variávely_prevista).round(4) metrics.mean_squared_error(variávely_observada, variávely_prevista).round(4) np.sqrt(metrics.mean_squared_error(variávely_observada, variávely_prevista)).round(4)

Comando 28: Separação da amostra em treino e teste para os modelos preditivos

def train_test_rmse(variávelx_rl, variávely_rl):

variávelx_train, variávelx_rl_test, variávely_rl_train, variávely_rl_test = train_test_split(variávelx_rl, variávely_rl, random_state=1) rl_treino_teste = LinearRegression() rl_treino_teste.fit(variávelx_rl_train, variávely_rl_train) variávely_prevista = rl_treino_teste.predict(variávelx_rl_test) return np.sqrt(metrics.mean_squared_error(variávely_rl_test, variávely_prevista)) train_test_rmse(variávelx_rl, variávely_rl).round(4)

Comando 29: Instalação das ferramentas que possibilitam a construção de árvores de decisão

pip install graphviz pip install pydotplus conda install graphviz conda install python-graphviz conda update -n base -c defaults conda

Comando 30: Importação das bibliotecas afetas aos algoritmos de *machine learning*

from sklearn.tree import DecisionTreeClassifier from sklearn.metrics import classification_report, confusion_matrix from sklearn.tree import DecisionTreeRegressor from sklearn.tree import export_graphviz from six import StringIO from IPython.display import Image import pydotplus from sklearn.preprocessing import MinMaxScaler from sklearn.ensemble import RandomForestClassifier from sklearn.preprocessing import StandardScaler from sklearn.ensemble import RandomForestRegressor from sklearn.svm import SVC

Comando 31: Definição das variáveis a utilizar e separação nas amostras treino e teste

atributos_ad = [‘Variáveis independentes'] fatores = df[atributos_ad] alvo = df['Variável-alvo (Y/N)'] fatores_train, fatores_test, alvo_train, alvo_test = train_test_split(fatores, alvo, test_size = 0.3, stratify = alvo, random_state = 1)

Comando 32: Definição da profundidade das árvore de decisão e chamada das variáveis

classificador = DecisionTreeClassifier(random_state = 5, max_depth = 3) classificador.fit(fatores_train, alvo_train) alvo_previsão = classificador.predict(fatores_test)

Comando 33: Construção da matrix de confusão, relatório de valores e representação gráfica

matrix_satisfação = confusion_matrix(alvo_test, alvo_previsão, labels=[0,1]) rótulos_satisfação = pd.DataFrame(matrix_satisfação, index=['Rótulo', 'Rótulo'], columns=['Previsto - Rótulo', 'Previsto - Rótulo']) fig, ax = plt.subplots(figsize=(7, 5)) sns.heatmap(rótulos, annot = True, fmt = "d", center = 1) print(classification_report(alvo_test, alvo_previsão))

Comando 34: Construção da árvore

dot_data = StringIO() export_graphviz(classificador, out_file=dot_data, filled=True, rounded=True, special_characters=True,feature_names = atributos_ad,class_names=['0','1']) graph = pydotplus.graph_from_dot_data(dot_data.getvalue()) graph.write_png('rótulo.png') Image(graph.create_png())

B - Fórmulas

Fórmula 1: Precisão

$$\frac{VP}{(VP + FP)} \tag{1}$$

Fórmula 5: Sensibilidade

$$\frac{VP}{(VP + FN)} \tag{2}$$

Fórmula 3: *F1 score*

$$\frac{(2 \times Precisão \times Sensibilidade)}{(Precisão + Sensibilidade)} \tag{3}$$

Fórmula 4: PECC

$$\frac{(VP + VN)}{(VP + VN + FN + FP)} \tag{4}$$

Fórmula 6: Especificidade

$$\frac{VN}{(VN + FP)} \tag{5}$$

C – Artigos científicos revistos

ID	Mês	Ano	Título	Autores	Publicação	Localização
1	Abril	2022	The Relationship between Digital Marketing, Customer Engagement, and Purchase Intention via OTT Platforms	Habib, S. Hamadneh, N. Hassan, A.	Journal of Mathematics	Riyadh, Arábia Saudita.
2	Fevereiro	2022	Film preferences in the pandemic: psychological resilience or an escape from the pandemic reality?	Pisarek, J. Zabielska-Mendyk, E.	Annals of Psychology 24	Lublin, Polónia.
3	Novembro	2021	Impact of Digital Technologies on the Development of Modern Film Production and Television	Alforova, Z. Marchenko, S. Kot, H. Medvedieva, A. Moussienko, O.	Rupkatha Journal 13	Kharkiv, Ucrânia. Kyiv, Ucrânia.
4	Outubro	2021	Using Hybrid Deep Learning Models of Sentiment Analysis and Item Genres in Recommender Systems for Streaming Services	Dang, C. Moreno-García, M. De La Prieta, F.	Electronics 10	Ho Chi Minh, Vietname. Salamanca, Espanha.
5	Setembro	2021	Expansion of the core business of traditional media companies in Spain through SVOD services	Alcolea-Diaz, G. Marin-Lladó, C. Cervi, L.	Communication & Society 35	Madrid, Espanha. Barcelona, Espanha.
6	Agosto	2021	Streaming media and the dynamics of remembering and forgetting: The Chernobyl case	Gambarato, R. Heuman, J. Lindberg, Y.	Memory Studies 15	Jonkoping, Suécia.
7	Maio	2021	A Heterogeneous Learning Framework for Over-the-Top Consumer Analysis Reflecting the Actual Market Environment	Choi, J. Kim, Y.	Applied Sciences 11	Gyungbuk, Coreia. Seoul, Coreia.
8	Maio	2021	A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach	Awan, M. Khan, R. Nobanee, H. Yasin, A. Anwar, S. Naseem, U. Singh, V.	Electronics 10	Lahore, Paquistão. Abu Dhabi, Emirados Árabes Unidos. Oxford, Reino Unido. Liverpool, Reino Unido. Islamabad, Paquistão. Taxila, Paquistão. Sydney, Austrália. Delhi, Índia.
9	Maio	2021	Examining Factors Influencing Early Paid Over-The-Top Video Streaming Market Growth: A Cross-Country Empirical Study	Lee, S. Lee, S. Joo, H. Nam, Y.	Sustainability 13	Seoul, República da Coreia.
10	Abril	2021	Information Behavior on Video on Demand Services: User Motives and Their Selection Criteria for Content	Gutzeit, J. Dorsch, I. Stock, W.	Information 12	Duseldorf, Alemanha. Graz, Aústria.
11	Março	2021	The Fragmentation of Indonesian Film Audience	Suwarto, D. Setiawan, B. Adikara, G.	Jurnal Komunikasi	Yogyakarta, Indonésia.

12	Março	2021	Movie characteristics as determinants of download-to-own performance in the Korean video-on-demand market	Jang, M. Baek, H. Kim, S.	Telecommunications Policy 45	Daejeon, República da Coreia. Seoul, República da Coreia.
13	Março	2021	Research on English Movie Resource Information Mining Based on Dynamic Data Stream Classification	Duan, J. Gao, R.	Security and Communication Networks	Yunnan, China. Tóquio, Japão.
14	Fevereiro	2021	Factors affecting consumers' willingness to subscribe to over-the-top (OTT) video streaming services in India	Nagaraj, S. Singh, S. Yasa, V.	Technology in Society 65	Hyderabad, Índia.
15	Dezembro	2020	Role of Covid as a catalyst in increasing adoption of OTTs in India: A study of evolving consumer consumption patterns and future business scope	Nijhawan, G. Dahiya, S.	Journal of Content, Community & Communication 13	Delhi, Índia.
16	Novembro	2020	Understanding motivations to use online streaming services: integrating the technology acceptance model (TAM) and the uses and gratifications theory (UGT)	Camilleri, M. Falzon, L.	Spanish Journal of Marketing	Edinburgh, Reino Unido. Msida, Malta.
17	Novembro	2020	The Impact of Over-the-Top Television Services on Pay-Television Subscription Services in South Africa	Udoakpan, N. Tengeh, R.	Journal of Open Innovation: Technology, Market and Complexity 6	Cidade do Cabo, África do Sul.
18	Setembro	2020	Cultural consonance predicts leisure satisfaction	Chick, G. Dong, E. Yeh, C. Hsieh, C.	Leisure Studies 40	Pennsylvania, Estados Unidos da América. Arizona, Estados Unidos da América. Hualien, Taiwan. Taichung, Taiwan.
19	Novembro	2019	The evolution of the pay TV market and the profile of the subscribers	Medina, M. Herrero, M. Portilla, I.	Revista Latina de Comunicación Social 74	Navarra, Espanha.
20	Outubro	2019	Competitions between OTT TV platforms and traditional television in Taiwan: A Niche analysis	Chen, Y.	Telecommunications Policy 43	Taipei City, Taiwan.
21	Março	2019	Experimental evidence on demand for "on-demand" entertainment	McKenzie, J. Crosby, P. Cox, J. Collins, A.	Journal of Economic Behavior and Organization 161	Macquarie, Austrália. Athabasca, Austrália. Nottingham, Reino Unido.
22	Novembro	2018	Approaching media industries comparatively: A case study of streaming	Herbert, D. Lotz, A. Marshall, L.	International Journal of Cultural Studies 22	Michigan, Estados Unidos da América. Queensland, Austrália. Bristol, Reino Unido.
23	Novembro	2017	The Effect of Subscription Video on-Demand on Piracy: Evidence from a Household Level Randomized Experiment	Matos, M. Ferreira, P. Smith, M.	Articles in Advance	Lisboa, Portugal. Pittsburgh, Estados Unidos da América.

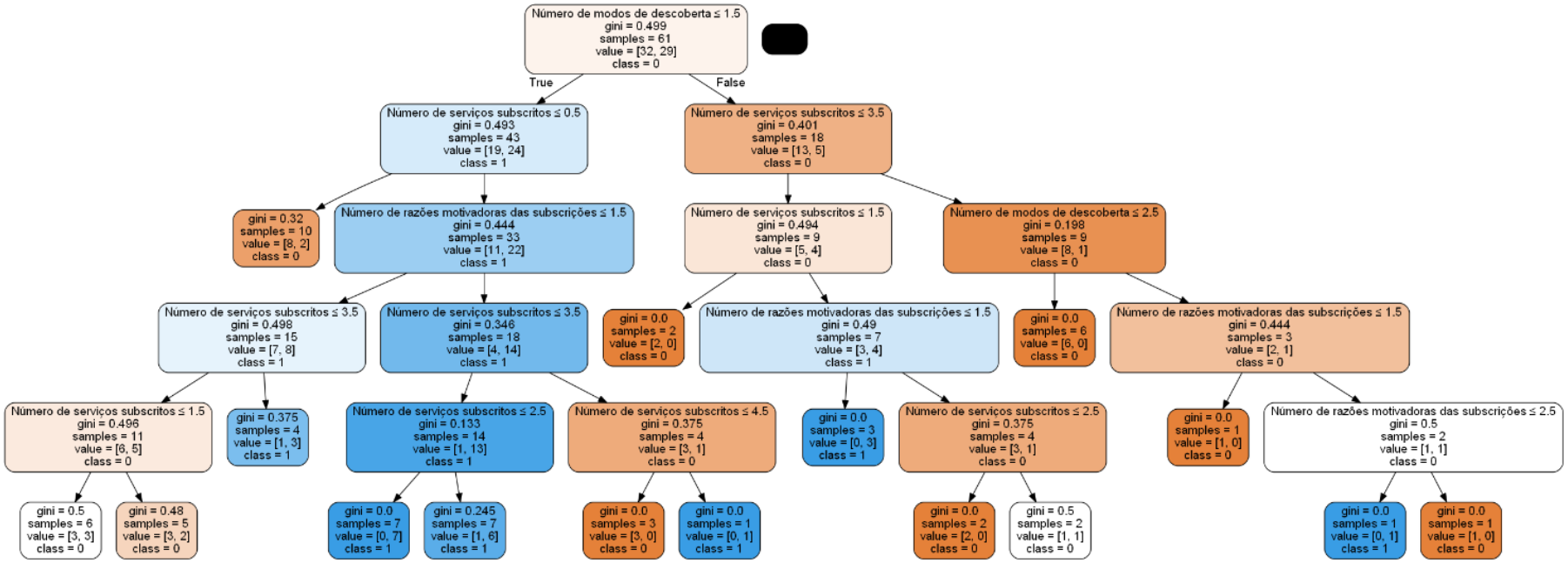
D – Apresentação do questionário

Título			
Conteúdo de entretenimento - Hábitos de consumo e interesses			
Questões	**Denominação**	**Tipo**	**Opções**
Secção 1 - Dados sociodemográficos			
1	Género	Escolha Múltipla	Masculino; Feminino; Não-binário; Prefiro não responder
2	Faixa Etária	Escolha Mútlipla	[20-30[; [30-40[; [40-50[; [50-60[; [60-70[; Prefiro não responder
3	Rendimento	Escolha Múltipla	<1000€; [1000-2000€[; [2000-3000€[; [3000-4000€[; [4000-5000€[; >5000€; Prefiro não responder
Secção 2 - Definição de preferências			
4	Em que tipo de conteúdo de entretenimento o participante tem interesse/investimento pessoal	Escolha Múltipla	Televisivo; Cinematográfico; Ambos; Nenhum/Terminar Survey
Secção 3 - Dados de consumo			
5	Número de horas consumidas/dia em média	Escolha Múltipla	<=1; [2-3h[; [4-5h[; >5
6	Dias da semana com maior incidência	*Checklist*	Segunda ou Terça-feira; Quarta ou Quinta-feira; Sexta; Sábado-Domingo
7	Número de séries vistas/mês em média	Escolha Mútipla	=0; [1-2[; [3-4[; >=5
8	Número de filmes vistos/mês em média	Escolha Múltipla	=0; [1-2[; [3-4[; >=5
9	Como decide o participante o que ver a seguir	*Checklist*	Género; Atores envolvidos; Realizadores; Produtores; Popularidade; Motores de recomendação dos serviços utilizados; Sem critério
10	Dispositivos utilizados para consumo de entretenimento	*Checklist*	Televisão; Telemóvel; Tablet; Outros dispositivos móveis; Projetor
11	Variedade de géneros consumidos	*Checklist*	Tudo um pouco; Ação; Comédia, *Thriller*; Documentário; Desporto; Fantasia; Sci-Fi; Terror; Drama
Secção 4 - Tipos de utilização			
12	Modo de consumo predominante	Escolha Múltipla	Serviços de subscrição; Televisão por cabo; Cinema; Pirateado

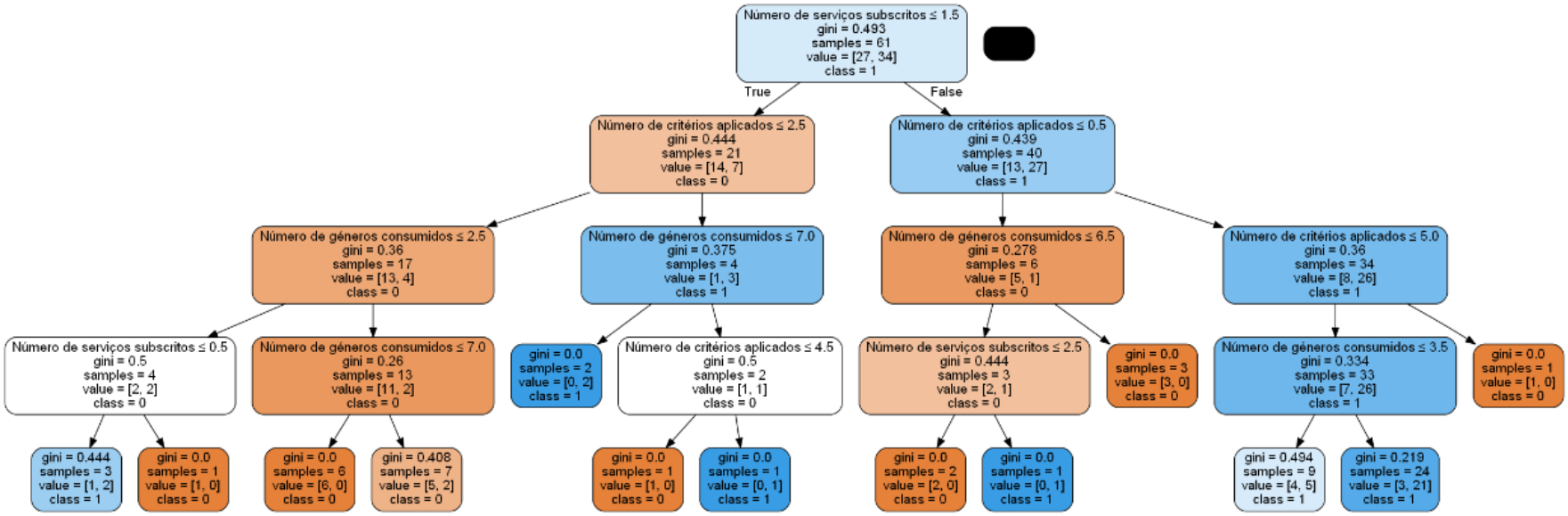
Secção 5 - Serviços de subscrição			
13	A que serviços pertence o conteúdo de que usufrui	*Checklist*	Netflix; HBO Max; Amazon Prime Video; Disney +; Apple TV +; Paramount +; Hulu; Peacock; Showtime; Youtube Premium; TV Cine; SportTV/Outro no âmbito do desporto; Outro
14	Razões que motivam a opção por serviços de subscrição	*Checklist*	Superior qualidade do conteúdo oferecido; Pessoas próximas também subscrevem; Preço; Acessbilidade; Popularidade
15	Descoberta dos serviços de subscrição cujo conteúdo usufrui de momento	*Checklist*	"*Word of Mouth*"; Fontes noticiosas de entretenimento; Amigos/Família; Outro
Secção 6 - Medição da temperatura/Perceções gerais			
16	Montante dispendido em conteúdos de entretenimento/mês	Escolha Múltipla	<25€; [25-50€[; [50-100€[; >=100€
17	Grau de satisfação com a acessibilidade ou preço pago (se esta segunda dimensão for aplicável) do conteúdo televisivo/cinematográfico disponível atualmente	Escala de 0 a 10	Insatisfeito - Valor 0
			Satisfeito - Valor 10
18	Nível de qualidade percecionada do conteúdo televisivo/cinematográfico disponibilizado atualmente	Escala de 0 a 10	Qualidade mínima - Valor 0
			Qualidade máxima - Valor 10
19	Grau de satisfação com o conteúdo televisivo/cinematográfico disponível atualmente	Escala de 0 a 10	Insatisfeito - Valor 0
			Satisfeito - Valor 10
Secção 7 – *Feedback* do questionário			
20	Nível de interesse neste questionário, da maneira como foi formulado	Escala de 0 a 10	Interesse mínimo - Valor 0
			Interesse máximo - Valor 10
21	Reparos e Oportunidades de melhoria	Resposta Aberta	N/A

E – Apresentação das árvores de decisão

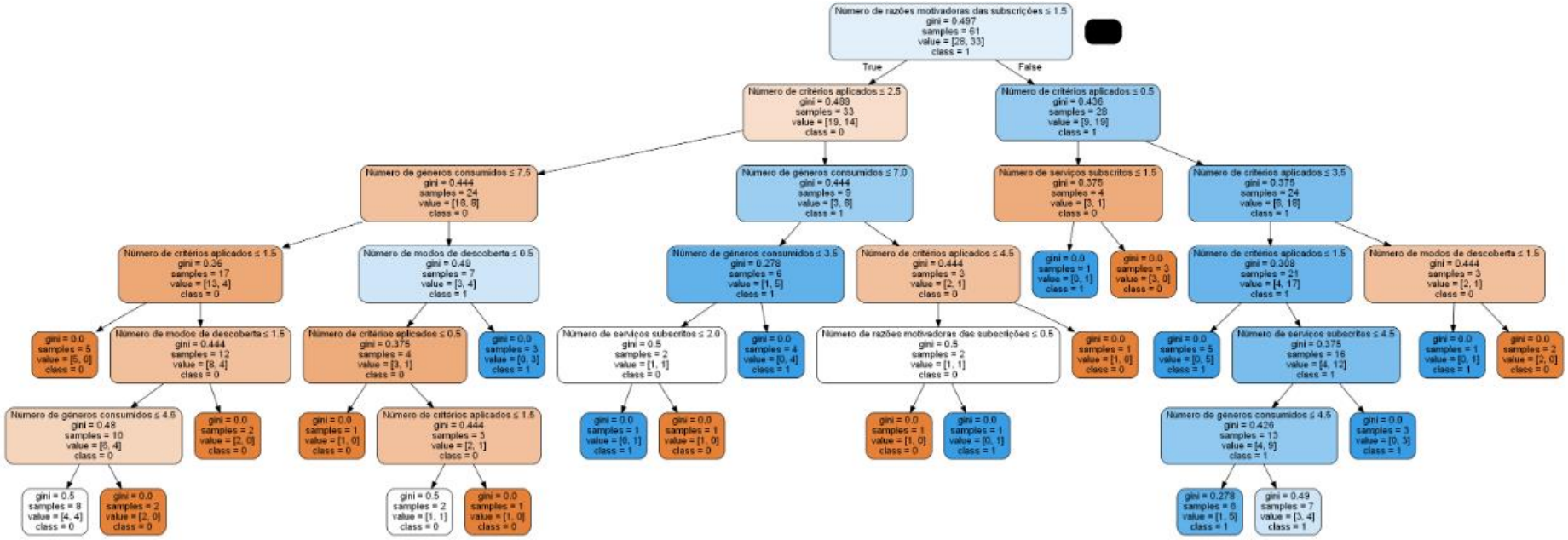
Árvore de decisão para o modelo 1 – Profundidade 5



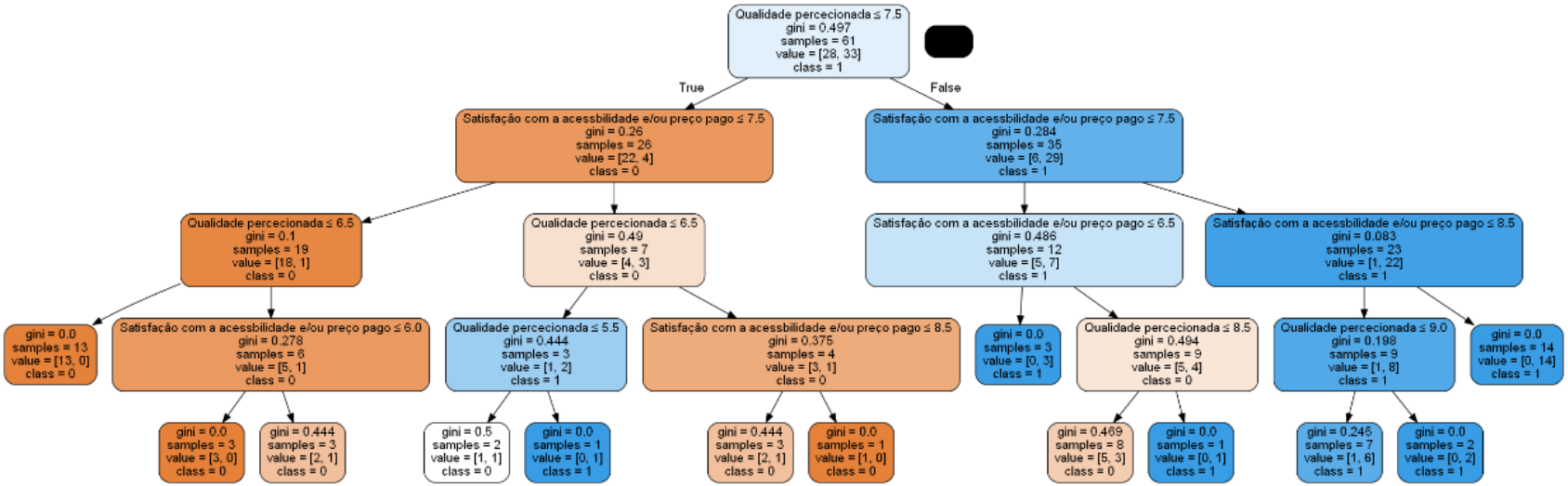
Árvore de decisão para o modelo 2 – Profundidade 4



Árvore de decisão para o modelo 3 – Profundidade 6



Árvore de decisão para o modelo 4 – Profundidade 4



F – Cotações atribuídas aos artigos revistos em função dos critérios e questões

Questões	Q1		Q2		Q3			Q4	
Critérios / Artigos	C1	C2	C3	C4	C5	C6	C7	C8	Total
1	0.5	0.5	0	0.5	1	1	1	0	4.5
2	0	0	0.5	0.5	0	0	0.5	1	2.5
3	1	0	0	0.5	0	1	1	0	3.5
4	0	1	1	0.5	1	0	0.5	1	5
5	1	0.5	0.5	0.5	0	1	0.5	0	4
6	0	0.5	0.5	0	0	0	1	0	2
7	1	1	1	1	1	0	0.5	0.5	6
8	0.5	0.5	0.5	1	1	0	0	1	4.5
9	1	0.5	0	0.5	1	0	0.5	1	4.5
10	0	0.5	0	0	0.5	0.5	0	0	1.5
11	0.5	0.5	0.5	0	0	0	0.5	0	2
12	0.5	0.5	0.5	1	1	0	1	1	5.5
13	0.5	0.5	1	1	0.5	0.5	0.5	0	4.5
14	0	0	1	0.5	1	0	0	0.5	3
15	1	1	0.5	0.5	0	1	1	1	6
16	0	0.5	1	0.5	1	0.5	1	1	5.5
17	1	0.5	0.5	0.5	1	0.5	1	1	6
18	0	1	0.5	0.5	1	1	0.5	1	5.5
19	0.5	1	1	0	1	0	0.5	1	5
20	0	1	1	0.5	1	0.5	1	1	6
21	1	1	1	1	0.5	1	0	0.5	6
22	1	0.5	0	0	0	0	0.5	0	2
23	0	0.5	1	0.5	1	0.5	1	1	5.5
Total	11	13.5	13.5	11.5	14.5	9	14	13.5	
Total/questão	24.5		25		37.5			13.5	
Média/questão	12.25		12.5		12.5			13.5	

G – Outros suportes e consultas

Matriz de confusão – *Scikit-learn*

		Valor Previsto	
		Negativo	**Positivo**
Valor Observado	**Negativo**	Verdadeiro Negativo	Falso Positivo
	Positivo	Falso Negativo	Verdadeiro Positivo

Reparos e melhorias – *Feedback* questinário

ID	**Obsevação**
1	Ausência da opção para a faixa etária de 40 a 50 na Q2
2	Ausência da opção para visualização através do computador, portátil ou desktop, na Q10
3	Necessidade de maior granularidade de intervalos, na Q5, onde não são contemplados todos os intervalos possíveis
4	Confronto entre os preços praticados e conteúdo disponbilizado mediante região, o que envolveria contemplar um conjunto de participantes mais abrangente e distintos entre si
5	Confronto do nível de satisfação entre os diversos serviços de subscrição, o que não foi considerado dada a discrepância de popularidade entre eles, sobretudo em contexto nacional
6	Maior grau de formalidade na redação das perguntas

Nível de interesse – Participantes do questionário