UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

TICIANA SÁ DA JUSTA

O consumo de filmes em cinemas no Brasil: uma análise de florestas aleatórias

São Paulo
2019

TICIANA SÁ DA JUSTA

O consumo de filmes em cinemas no Brasil: uma análise de florestas aleatórias

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação

Área de Concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 04 de janeiro de 2019. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011

Orientadora: Profa. Dra. Marislei Nishijima

São Paulo
2019

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB-8 4936

Justa, Ticiania Sá da

O consumo de filmes em cinemas no Brasil: uma análise de florestas aleatórias / Ticiania Sá da Justa ; orientadora, Marislei Nishijima. – 2019.

113 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, em 2018.

Versão corrigida

1. Mineração de dados. 2. Comportamento do consumidor.
3. Cinemas - Brasil. 4. Exibição de filmes - Brasil. 5. Internet.
I. Nishijima, Marislei, orient. II. Título.

CDD 22.ed.– 006.312

Dissertação de autoria de Ticiane Sá da Justa sob o título "**O consumo de filmes em cinemas no Brasil: uma análise de florestas aleatórias**", apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 04 de janeiro de 2019 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Marislei Nishijima
Universidade de São Paulo
Presidente

Prof. Dr. Alexandre Sartoris Neto
Universidade Estadual Paulista

Profa. Dra. Karina Valdivia Delgado
Universidade Federal de São Paulo

Prof. Dr. Silvio Yoshiro Mizuguchi Miyazaki
Universidade de São Paulo

Dedico esta dissertação à minha mãe Maria Madalena Sá da Justa que sempre me incentivou a estudar e a querer conhecer mais e me tornar uma pessoa melhor, a seus filhos que ajudaram a financiar meu ensino básico e aos meus filhos que me motivam a fazer sempre o melhor e a persistir sempre.

Agradecimentos

Agradeço primeiramente a Deus por ter me dado a vida, um planeta maravilhoso no qual habitamos e saúde plena. Agradeço também por ele ter me encaminhado para o seio de uma mulher acolhedora que me criou como sua filha e me transmitiu a base de minha formação moral e esta base tem sido a responsável pelas minhas melhores conquistas.

Agradeço a seus filhos, Maria José Macêdo e José Maria Macedo Júnior por terem continuado a financiar meus estudos sempre que sua mãe teve alguma dificuldade financeira.

Agradeço ainda por Ele ter me dado a dádiva de ter três crianças lindas, saudáveis e felizes, cujo futuro vislumbro maravilhoso e parte dele é responsabilidade minha ajudar a construir.

Agradeço em especial à minha orientadora, Profa. Dra. Marislei Nishijima por ter me acompanhado nesta jornada e me auxiliado a suplantando todas as dificuldades que surgiram no caminho, tornando – o uma parte essencial do meu desenvolvimento.

Agradeço à Profa. Dra. Sarajane Perez e ao Prof.. Dr. Marcelo Laurretto pelas contribuições à execução deste trabalho.

Por fim agradeço a todos os professores do programa pelas excelentes disciplinas ministradas e experiências proporcionadas por esta convivência, e à Universidade de São Paulo por fazer tão bom uso do dinheiro público.

“O amor nunca desiste, nunca perde a fé, sempre tem esperança e sempre se mantém firme.”

(Apóstolo Paulo – 1 Coríntios 7)

Resumo

JUSTA, Ticiania Sá da. **O consumo de filmes em cinemas no Brasil:** uma análise de florestas aleatórias. 2019. 113 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2019.

Baseado em dois momentos distintos da realidade brasileira, os anos de 2002/03 e 2008/09, este estudo busca avaliar se houve mudança no perfil dos consumidores de filmes em salas de cinema, por meio de variáveis observáveis de indivíduos consumidores e não consumidores. Para esta finalidade são empregadas técnicas de mineração de dados de florestas aleatórias sobre os microdados das Pesquisas de Orçamentos Familiares (POFs) de 2002/2003, quando o uso de banda larga no Brasil era praticamente nulo, e de 2008/2009, quando esta já estava estabelecida no país. Esta diferença no tempo e de acesso à banda larga proporcionam uma janela de oportunidade para nosso objetivo, dado o elevado grau de mudanças tecnológicas do período investigado. Apesar de metodologicamente não ser possível isolar completamente o efeito do acesso à Internet sobre o consumo de cinema, espera-se que quanto maior a velocidade de acesso e maior a evolução das tecnologias de compactação de arquivos, a distribuição de conteúdo aumente na rede e permita maior consumo de filmes de maneira alternativa ao em salas de cinema. Neste contexto, além de identificar se houve alguma mudança de perfil dos consumidores de filmes em salas de cinema no Brasil após a popularização da banda larga no país, que oficialmente pode ser considerada no ano de 2006, também avaliamos se estes consumidores diferem significativamente dos não consumidores em suas características observáveis. Adicionalmente, num segundo estudo, usando um modelo de intervenção de Diferenças em Diferenças, investigamos o efeito da banda larga sobre o consumo de cinema. Os resultados apontam divergência clara entre os perfis de consumidores e não consumidores de filmes de cinema.

.

Palavras-chaves: Filmes. Cinema. Mineração de dados. Florestas aleatórias. Internet banda larga. Diferenças em Diferenças. Análise de consumidores. Perfil de consumidores.

Abstract

Justa, Ticiana Sá da. **Movies' consumption in Brazil: a random forests analysis.** 2019. 113 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2019.

I investigate whether film at theaters' consumers in two distinct time are different in their observable characteristics. My investigation is based on observable variables of consumers and non-consumers data from the Family Budget Survey (POF) of 2002/2003, when the use of broadband in Brazil was almost nil, and from 2008/2009, when it already was established in the country. This change in addition to the high degree of technological changes over the period provides a window of opportunity for our objective. Although methodologically it is not possible to full isolate the effect of Internet access on film consumption, it is expected that the higher the user access speed and the greater the evolution of file compression technologies, the faster and bigger will be the traffic of films files on the web. Accordingly, the greater distribution of content on the network enables the consumption of films in alternative ways to movie theaters. Thus, in addition to identifying possible changes in the profile of movie consumers at the theaters in Brazil since the popularization of broadband in the country, which can officially be considered in the year 2006, the study assess whether these consumers differ significantly from not consumers in their observable characteristics. The results points to several distinctions between cinema consumers and no cinema consumers.

Keywords: Movies. Cinema. Broadband. Consumption. Data mining. Random forests. Consumer profile analysis. Consumer profiling. Differences in Differences.

Sumário

1	Introdução.....	12
1.1	<i>Objetivo geral.....</i>	<i>13</i>
1.2	<i>Objetivos específicos.....</i>	<i>13</i>
1.3	<i>Estrutura.....</i>	<i>14</i>
2	Cinema e avanços tecnológicos no Brasil.....	15
2.1	<i>Parque exibidor no Brasil.....</i>	<i>15</i>
2.2	<i>Evolução do consumo de cinema de 2002 a 2009.....</i>	<i>16</i>
2.3	<i>Evolução do acesso à Internet banda larga de 2002 a 2008.....</i>	<i>17</i>
2.4	<i>Internet e seu impacto sobre os bens de informação.....</i>	<i>18</i>
2.5	<i>Evolução do consumo de filmes em salas de cinema no Brasil.....</i>	<i>19</i>
3	Fundamentação teórica	21
3.1	<i>Aprendizagem de máquina.....</i>	<i>23</i>
3.2	<i>Principais tarefas da mineração de dados.....</i>	<i>24</i>
3.3	<i>Técnicas utilizadas no estudo.....</i>	<i>24</i>
4	Florestas de decisão aleatórias	26
4.1	<i>Árvores de decisão.....</i>	<i>26</i>
4.2	<i>Florestas aleatórias de Breiman e Cutler (2001).....</i>	<i>29</i>
4.3	<i>Avaliação dos classificadores gerados com as florestas de decisão aleatórias.....</i>	<i>31</i>
5	Balanceamento dos conjuntos originais para a execução do treinamento das florestas.....	33
5.1	<i>Racing.....</i>	<i>33</i>
5.2	<i>SMOTE.....</i>	<i>34</i>
6	Dados.....	36
6.1	<i>Pesquisa de Orçamentos Familiares do IBGE</i>	<i>36</i>
6.1.1	<i>Microdados da POF 2002/2003 e POF 2008/2009.....</i>	<i>37</i>
6.2	<i>Obtenção dos dados e extração dos registros.....</i>	<i>38</i>
6.3	<i>Seleção de dados</i>	<i>39</i>
6.4	<i>Formatação</i>	<i>39</i>

6.5	<i>Análise exploratória e sumarização das variáveis</i>	40
7	Treinamento e análise das florestas geradas e suas variáveis preditoras	45
7.1	<i>Florestas da POF2002</i>	46
7.2	<i>Florestas da POF2008</i>	50
7.3	<i>Classificação dos dados da POF2008 a partir do modelo da POF2002</i>	53
7.4	<i>Comparação entre as variáveis de classificação e seleção das variáveis de análise do perfil de consumo.</i>	53
8	Análise das variáveis preditoras das florestas	55
8.1	<i>Análise das 4 variáveis com maior importância relativa total</i>	55
8.1.1	<i>Gastos com recreação</i>	55
8.1.2	<i>Rendimentos totais do indivíduo:</i>	58
8.1.3	<i>Unidade da Federação</i>	61
8.1.4	<i>Idade</i>	63
8.2	<i>Análise das variáveis que mais perderam importância</i>	65
8.3	<i>Análise das variáveis com relação apenas à importância de permutação</i>	66
9	Conclusões do estudo realizado com as florestas aleatórias	70
9.1	<i>Verificar se houver mudança no perfil dos consumidores de cinema antes e após o advento da internet banda larga.</i>	70
9.2	<i>Verificar se o perfil de consumidores de cinema difere do dos não consumidores....</i>	70
9.3	<i>Avaliar o desempenho da aplicação das florestas aleatórias para fins de extração de informações relevantes sobre consumidores.</i>	71
10	Modelo de regressão diferenças em diferenças	73
10.1	<i>Experimentos naturais.</i>	73
10.2	<i>Diferenças em diferenças</i>	74
10.3	<i>Avaliação dos modelos de regressão</i>	75
11	Estudo específico sobre efeitos da banda larga sobre consumo de cinema no Brasil utilizando diferenças em diferenças	77
11.1	<i>Introdução</i>	77
11.2	<i>Dados e Estratégia Empírica</i>	80
11.3	<i>Achados Empíricos e Discussão</i>	81
11.4	<i>Conclusão.</i>	85

12	Discussão e trabalhos futuros.....	86
12.1	<i>Seleção de variáveis de atuação</i>	86
12.2	<i>Verificação dos resultados obtidos</i>	87
12.3	<i>Importância do trabalho para a área de inteligência artificial e trabalhos futuros ..</i>	87
	Referências	88
	Apêndice A - Revisão da Literatura	96

1 Introdução

O cinema é arte e indústria, mas independente da motivação o seu objetivo final ao ser produzido pode ser considerado um típico bem de consumo (HADIDA, 2010). Trata-se de arte no sentido em que se comunica diretamente com as emoções do público, transmitindo mensagens, recontando fatos históricos, influenciando comportamentos e apontando tendências nos mais amplos setores da sociedade. Indústria porque a motivação do produtor é o lucro ou remuneração de sua atividade produtiva.

No caso dessa indústria, a maior parte da produção ocorre em outros países, sendo o Brasil um pequeno produtor. A maioria dos filmes consumidos mundialmente é produzida em Hollywood (EUA), Bollywood (Índia) e uma pequena parte em Hong Kong, o que indica que centros de produção concentrados em uma determinada região geográfica favorecem a produção cinematográfica (CAPLAN;COWEN, 2004). No que se refere ao consumo de filmes em salas de cinema, dada a enorme desigualdade do Brasil, verifica-se um pequeno público consumidor, concentrado nas pessoas mais ricas e mais educadas (NIEUWENHOFF, 2017).

Considerando que o acesso ao consumo de cinema, um bem cultural de informação, representa um ganho social para a população em geral, este estudo busca investigar em uma rica base de dados de indivíduos (POF do IBGE, Pesquisa de Orçamento Familiar) disponíveis nos anos 2002/03 e 2008/09 se houve mudança no perfil do consumidor, por meio de técnicas de florestas aleatórias, decorrente da mudança do padrão de Internet com a entrada da banda larga entre estas duas datas.

A motivação dessa investigação decorre do fato do filme ser um bem de informação, com alto custo de produção e muito baixo custo de reprodução (SHAPIRO; VARIAN, 2000). Sendo assim, o advento da banda larga trouxe uma enorme inovação para os veículos de transmissão de informações, o que possibilitou a redução dos custos de transmissão (OBERHOLZER-GEE;STRUMPF,2007), uma vez que aumenta a velocidade, o processamento e o volume de transporte de dados. Além disso, a mudança tecnológica do filme de película para um padrão digital também implicou uma redução muito significativa no custo de produção, reprodução e distribuição para a indústria cinematográfica (HUSAK,2004).

O comportamento dos consumidores consiste num instrumento analítico possível para descobrir quais políticas de incentivo são mais proveitosas e quais poderiam ser revistas à luz

de informações adicionais, uma vez que o consumo de filmes em salas de cinema consiste em importante indicador do sucesso destas políticas de incentivo.

1.1 Objetivo geral

Esta dissertação busca avaliar a aplicabilidade das florestas aleatórias para a finalidade de identificar as possíveis mudanças comportamentais no consumo e também no perfil dos consumidores brasileiros de filmes em salas de cinema em dois diferentes pontos do tempo. Sendo que no primeiro período o acesso à Internet banda larga era exíguo e no período seguinte mais abundante, sendo concomitante com outras mudanças tecnológicas associadas.

Os momentos investigados são definidos conforme a disponibilidade de dados individuais de consumidores na Pesquisa de Orçamento Familiar (POF) do IBGE, ou seja, 2002/3 e 2008/9. Além disso, o estudo investiga se o perfil dos consumidores de cinema se difere em características observáveis dos não consumidores.

Importante notar que embora não seja possível dizer que mudanças de perfil tenham ocorrido exclusivamente por banda larga e outras mudanças tecnológicas, dada a impossibilidade de isolar perfeitamente a variável causadora, tais elementos são os grandes motivadores do estudo, uma vez que são em parte seus responsáveis.

1.2 Objetivos específicos

- Verificar se houve mudança no perfil dos consumidores de cinema antes e após o advento da internet banda larga.
- Verificar se o perfil de consumidores de cinema difere do dos não consumidores
- Avaliar o desempenho da aplicação das florestas aleatórias para fins de extração de informações relevantes sobre os consumidores.
- Aplicar técnicas de mineração de dados e inteligência artificial sobre os microdados do IBGE.

- Aplicar técnica econométrica de análise de intervenção, modelo de Diferenças em Diferenças, para estudar o potencial efeito da Internet banda larga sobre o consumo de cinema.

1.3 Estrutura

A dissertação apresenta dois estudos. O primeiro investiga se os consumidores de cinema possuem características semelhantes ou diferentes em dois momentos distintos do tempo e se os consumidores de cinema se diferem dos não consumidores com base em aplicações de florestas aleatórias. O segundo estudo verifica o impacto da Internet banda larga sobre o consumo de cinema no país usando modelo de diferença em diferenças.

Assim, em modo mais detalhado para cada estudo, a dissertação está composta do capítulo 2 que apresenta informações sobre consumo de cinema, condições de acesso à banda larga e outras mudanças tecnológicas no Brasil. O capítulo 3 apresenta a fundamentação teórica no qual este trabalho se baseia e os capítulos 4, 5 e 10 detalham as técnicas utilizadas. O capítulo 6 apresenta os dados utilizados nos estudos. Os capítulos 7 e 8 apresentam os treinamentos dos modelos com florestas aleatórias, sua aplicação e análise dos resultados. O capítulo 9 apresenta as conclusões sobre os dados que foram atingidas com o modelo de florestas aleatórias. O capítulo 11 apresenta o estudo realizado com o método de diferenças em diferenças e o capítulo 12 realiza uma breve discussão sobre os achados do trabalho e possibilidades de pesquisas futuras.

2 Cinema e avanços tecnológicos no Brasil

2.1 Parque exibidor no Brasil

A distribuição de filmes em salas de cinema é comumente a primeira forma disponibilização do produto cinematográfico aos consumidores. No Brasil, conforme dados da OCA/ANCINE em 2015 o Parque Exibidor de filmes era composto por: 3005 salas de exibição; 2874 salas de exibição digitais; 1190 salas de exibição 3D; 742 complexos exibidores; e 1 sala para cada 68037 habitantes, distribuídos no território nacional conforme a figura 1.

Figura 1 – Distribuição de salas de cinema no território nacional

Fonte: Ticiania Sá da Justa a partir de dados da OCA / ANCINE, 2018

2.2 Evolução do consumo de cinema de 2002 a 2009

Segmentando estes dados para o período investigado, 2002 a 2009, temos a evolução do número de ingressos totais, figura 2, e do número de ingressos por capita conforme a produção doméstica e estrangeira, figura 3.

Figura 2 – Evolução do número total de ingressos: Brasil 2002 a 2009

Fonte: Ticiania Sá da Justa a partir de dados da OCA/ANCINE, 2018

Figura 3 – Evolução do número de ingressos per capita no Brasil e por produção brasileira ou estrangeira

Fonte: Ticiania Sá da Justa a partir de dados da OCA/ANCINE, 2018

2.3 Evolução do acesso à Internet banda larga de 2002 a 2008

O uso da Internet em banda larga tem aumentado substancialmente no Brasil após 2002 com crescimento no número de acessos¹. A taxa anual média de crescimento foi de 49% entre os anos 2002 e 2008 no acesso fixo. Antes de 2011 não estava oficialmente estabelecida a velocidade mínima de transmissão definidora da banda larga no país². Nas jurisdições de outros países o padrão, até recentemente, era considerar velocidades maiores que 256 Kbps. Atualmente a velocidade mínima necessária para caracterizar a banda larga está sendo revista dada as mudanças estruturais que os avanços tecnológicos tem gerado, possibilitando mudanças nos conteúdos transmitidos.

Verifica-se que de fato a entrada da banda larga no país aconteceu entre 2002 e 2008/2009. A figura 4 foi elaborada a partir de dados da ANATEL – Agência Nacional de Telecomunicações e ilustra a evolução do acesso por velocidade de transmissão, quando se verifica que mesmo a velocidade entre 512Kbps e 2Mbps tem participação pequena até 2003.

Figura 4 – Evolução dos acessos no Brasil conforme a velocidade de transmissão

Fonte: Ticiane Sá da Justa a partir de dados da ANATEL, 2018

¹ <http://www.caminhosdabandalarga.org.br/2012/10/capitulo-1>

² Em 2011, a internet banda larga passa a ser regulada pelo Plano Nacional para Banda Larga sob a administração do Ministro das Comunicações Hélio Costa.

2.4 *Internet e seu impacto sobre os bens de informação*

A Internet de banda larga consiste em uma inovação dos veículos de transmissão de bens de informação, que possibilitou a redução do custo de transmissão (OBERHOLZE-GEE; STRUMPF, 2007). Os bens de informação são aqueles que podem ser codificados, armazenados, transmitidos e reproduzidos com custo de reprodução muito baixo (SHAPIRO; VARIAN, 1999). Os filmes são um exemplo de bens de informação cujos custos de transmissão foram reduzidos pela internet banda larga, conforme Husak (2004).

O advento da Internet que permitiu o compartilhamento em massa de informação digital conjuntamente com novas técnicas de compressão e codificação de arquivos de áudio, reduzindo seus tamanhos a alguns megabytes, mudou radicalmente a indústria fonográfica mundial. Esta foi forçada a otimizar outros meios de rentabilidade para além da venda de CDs de áudio, que foi amplamente impactada pela ação dos piratas virtuais (SHAPIRO; VARIAN, 2000). Tal dinâmica de diminuição de tamanho de arquivos aliada a um aumento no acesso à Internet banda larga se repetiu com relação aos arquivos de vídeos, incentivando a distribuição mediante lucro ou não de cópias não autorizadas, conhecida como pirataria, entretanto, o caso prévio da indústria fonográfica pode ter alertado a indústria de cinema para o fenômeno e permitido uma resposta do setor.

Os arquivos de vídeos possuem diversos formatos de containers. Um container é um pacote contendo os arquivos de áudio e vídeo codificados e agrupados, formando assim um vídeo com som e imagens, diferindo basicamente no tamanho do quadro, na taxa de bits por segundo do vídeo e na resolução para os arquivos de imagem. Além disso, suportam diversos tipos de codificação para os arquivos de áudio. Por exemplo, o uso do software Aimersoft Video Converter transforma um DVD com codificação de vídeo MPEG2 e tamanho 3,6GB em outros containers de vídeo com as seguintes características: MP4 – MPEG4 – atual padrão para codificação de vídeos para a internet: 1,5GB; WMV – Windows Media Video – padrão proprietário da Microsoft: 1GB; 3GP – Container voltado aos dispositivos móveis: 131MB; e 4K MP4 – Padrão para filmes com tecnologia 4k: 14,5 GB.

Deste modo, a compactação se mostrou eficiente no sentido de reduzir o volume em mais de 50%, em especial o 3GP com 131MB, exceto o container de arquivos de imagem de altíssima qualidade, conhecidos como 4K, com aumento significativo de seu tamanho.

Além disso, a mudança tecnológica de padrão de filmes de película para digital implicou também numa redução muito significativa do custo de reprodução e distribuição

para a indústria cinematográfica (HUSAK, 2004). O autor observa também que isso permitiu o lançamento mundial simultâneo de filmes, o acesso do conteúdo nas residências, múltiplas versões digitais com finais alternativos, cortes de diretores, múltiplas legendas em diferentes línguas e multi-censor que podem ser usados para evitar a pirataria.

2.5 Evolução do consumo de filmes em salas de cinema no Brasil

O Brasil também tem passado pelo processo de transição para o cinema digital, com salas de cinema exibindo prioritariamente em versões digitais e com o aumento da produção de filmes nacionais propiciadas pelo barateamento da produção digital (ANCINE, 2014). Sendo já uma versão digital sua capacidade de transmissão em containers fica imediata.

A evolução do número de expectadores em salas de cinema do país pode ser verificada nas Figuras 2 e 3. Observa-se uma tendência de crescimento no período interrompida entre 2005 e 2008. Este fenômeno pode estar relacionado com a evolução para filmes digitais, de modo que as mudanças nas salas de cinema teriam sido mais lentas, exigindo a troca de equipamentos e trabalhadores com diferentes habilidades para a reprodução digital ao invés de reprodução por película. A melhor qualidade do filme transmitido pela Internet pode ter concorrido com a qualidade da reprodução dos filmes em salas de cinema, obrigando-os a se adaptarem ao novo padrão digital.

Interessante observar que o faturamento do serviço de televisão a cabo, que opera supostamente como um concorrente, bem substituto, ao consumo de filmes em salas de cinema com certo atraso, também apresentou queda de faturamento de 2005 até 2006, Figura 5. Os valores monetários foram deflacionados pelo INPC do IBGE para valores de 2003.

A Figura 5 mostra a evolução do faturamento das vendas de serviços de televisão à cabo, que pode ser visto como bem substituto do consumo de filmes em cinemas. O filme no cinema é um produto de melhor qualidade porque está disponível na data de seu lançamento, enquanto qualquer outra forma de consumo de filmes ocorre com certo atraso, exceto pela pirataria. Entretanto, o grau de substituição tende a ser significativo de acordo com a literatura (HUI;PNG, 2002). Verifica-se que o faturamento caiu a partir de 2004 e voltou a crescer após 2006. A queda do faturamento da TV a cabo pode estar associado com a ampliação da banda larga no Brasil, no sentido dos filmes transmitidos pela Internet substituírem o consumo de cinemas e de televisões a cabo.

Figura 5 – Evolução do faturamento da TV à cabo em Reais constantes de 2003 a 2014

Fonte: Ticiania Sá da Justa a partir de dados da ANATEL, 2018

No que se refere à proteção da propriedade intelectual, a pirataria é uma das ameaças à indústria cinematográfica e constitui uma violação aos direitos de propriedade intelectual, ocasionando prejuízos à indústria e ao Governo com a perda de arrecadação de impostos.

3 Fundamentação teórica

Levando em conta a grande mudança tecnológica que o acesso a Internet banda larga e as grandes mudanças tecnológicas associadas à produção digital de filmes trouxeram ao seu consumo, que originaram várias alternativas de consumo em cinema, tais como DVD, Internet, etc., interessa verificar se o perfil do consumidor de filmes no cinema foi alterado. Além disso, busca-se verificar se é possível distinguir o consumidor do não consumidor com base nas informações observáveis. Para essa finalidade, utilizamos florestas aleatórias aplicadas aos indivíduos consumidores dos anos 2002/03 e 2008/9.

Desde o seu surgimento a humanidade vem desenvolvendo métodos de registro e documentação de sua história. Nesse processo, os registros de fatos e conhecimentos adquiridos, sua análise, e sua evolução por diversas mentes diferentes possibilitaram o acúmulo de conhecimentos e tecnologias que possibilitam melhores condições de vida para a humanidade.

Com o surgimento dos sistemas computacionais, as organizações tiveram a oportunidade de registrar e acumular cada vez mais dados, dos mais diversos tipos, tendo sido desenvolvidas tecnologias recentes e complexas como Data Warehouses, bancos de dados, entre outras (SWINIARSKI; et al, 2007)

Uma vez que a análise manual dos dados já não era possível devido ao seu volume, surgiu então a necessidade de novas técnicas que proporcionassem a extração de informações a partir dos dados armazenados. Esta área foi denominada de Mineração de Dados. Os algoritmos de aprendizagem de máquina provêm a sua base tecnológica e são utilizados para extrair informações a partir de dados em bancos de dados, trazendo à tona os padrões implícitos e relações existentes entre eles (WITTEN;FRANK, 2005).

Estes padrões identificados são úteis para poder realizar previsões futuras, classificações e encontrar padrões de associações entre conjuntos de dados.

Algumas das áreas em que a mineração de dados é amplamente utilizada são: Detecção de fraudes, identificação de perfis de clientes, diagnósticos médicos, ou seja, todas as áreas em que a informação esteja disponível e possa ser submetida aos algoritmos de aprendizagem de máquina, que procurarão por padrões existentes.

O método mais utilizado para a transformação de dados em conhecimento é a análise manual destas informações por um especialista, que a sumariza em forma de relatórios. Por causa do grande volume de dados, este método tornou-se inviável (FAYYAD, 1996). O termo

KDD (Knowledge Discovery in Databases) e Mineração de Dados são frequentemente utilizados com sinônimos, e serão tratados como tal neste trabalho.

Fayyad (1996) define KDD como "um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis". A Figura 1 apresenta a esquematização do processo de KDD segundo este autor.

Figura 6 - Processo de KDD

Fonte: Fayyad (1996)

Utilizamos este modelo do processo de KDD como base para desenvolvermos a metodologia de pesquisa deste trabalho, a qual está elencada nas seguintes etapas e detalhada no capítulo 6:

1) Seleção: Coleta dos microdados no site do IBGE e seleção dos registros de consumo de interesse para a pesquisa e também seleção dos registros de pessoas e condições de vida, os consumos de interesse se encontram identificados por códigos numéricos disponibilizados na documentação das pesquisas.

2) Pré – processamento: Unificação das tabelas de condições de vidas e pessoas e consumos de interesse.

3) Formatação: Criação de variáveis de apoio *dummies* e conjuntos de dados necessários para a execução dos treinamentos das florestas aleatórias e do método de diferenças em diferenças.

4) Mineração: Submissão dos conjuntos formatados aos algoritmos de florestas aleatórias e modelo de diferenças em diferenças.

5) Interpretação / Avaliação: Avaliação dos modelos gerados e sua eficiência e ajustes necessários. Interpretação dos resultados obtidos e discussão.

3.1 *Aprendizagem de máquina*

Um agente aprende se melhorar a sua forma de atuação nas tarefas futuras após fazer observações sobre o mundo (RUSSEL; NORVIG, 2010). Do ponto de vista computacional isto se aplica a partir do momento que os algoritmos conseguem melhorar seu desempenho a partir da análise de dados históricos.

Russel e Norvig (2010) elencaram os principais tipos de aprendizagem como sendo os seguintes:

Aprendizagem não-supervisionada, na qual o agente analisa os dados na entrada, sem metainformações sobre os mesmos e nenhum tipo de feedback explícito sobre o resultado do processamento. Neste aspecto, a tarefa mais comum é a de agrupamento, na qual os dados são agrupados em grupos semelhantes sem prévia classificação. Há diversos algoritmos de agrupamento disponíveis, entre os quais destacamos: i) AGNES e DIANA para o agrupamento de forma hierárquica entre os dados e que gera uma estrutura de árvore a partir da medida de similaridade ou dissimilaridade entre os elementos do conjunto, ii) o k – médias que encontra k (fornecido pelo usuário) grupos distintos particionando o conjunto inicial e agrupando os dados mais próximos aos k centróides encontrados, agrupamento por densidade com o DBSCAN e seus variantes, iii) redes neurais artificiais também podem ser utilizadas para o aprendizado não-supervisionado, como exemplo temos a utilização dos mapas auto – organizáveis (SILVA; BOSCARIOLI; PEREZ, 2016).

Aprendizagem por reforço, na qual o agente aprende a partir de recompensas ou punições, sendo alguns exemplos de algoritmos o Q-learning, SARSA, DQN e DDPG (KAELBLING; LITTMAN; MOORE, 1996).

Aprendizagem supervisionada, sendo este tipo fornecido pelos pares de entrada e saída, de modo que o agente mapeia uma função que explique esta relação. Sendo o método normalmente empregado para as tarefas de classificação. Alguns algoritmos que apresentam esta estratégia são o k – NN, Naive Bayes e Regressões lineares e não lineares (SILVA; BOSCARIOLI; PEREZ, 2016).

Aprendizagem semi - supervisionada, que se torna necessária quando o conjunto inicial de dados possui poucos exemplos rotulados e a grande maioria não rotulados. Neste caso, os exemplos rotulados serão utilizados para se obter informações sobre os dados e guiar o algoritmo no processo de aprendizagem a partir dos não rotulados (BRUCE, 2001). Estes algoritmos podem ser utilizados tanto em tarefas de agrupamento como de classificação.

3.2 Principais tarefas da mineração de dados

As tarefas de mineração de dados se dividem em preditivas e descritivas (HAN; KAMBER; PEI ,2011), sendo que as preditivas aprendem, a partir dos dados históricos, a prever o valor futuro da variável em estudo, enquanto as descritivas sintetizam informações das propriedades inerentes aos dados.

As tarefas descritivas são identificadas em:

- Classificação, que analisa o novo dado e fornece a ele um novo rótulo com valores pré-definidos, aprendidos a partir de dados históricos.
- Regressão, que utiliza processo semelhante porém para fins de predição de valores numéricos, que podem obedecer a uma função linear ou não.
- Extração de regras de associação entre itens de um conjunto em estudo.
- Análise de grupos cuja finalidade é identificar os indivíduos com a maior semelhança intra-grupo e maior diferença entre grupos, descobrindo as classes inerentes ao conjunto estudado.
- Análise de *outliers*, que são os indivíduos que diferem muito em suas características com relação aos demais integrantes do conjunto estudado.

3.3 Técnicas utilizadas no estudo

O objetivo deste trabalho é identificar o perfil dos consumidores de cinema e suas alterações ao longo do tempo, sendo a população dividida em duas classes, a de consumidores e não consumidores, deste modo a tarefa utilizada será a de classificação binária utilizando

florestas de decisão aleatórias e o modelo de regressão diferenças em diferenças para o consumo de cinema e suas relações com outros bens de informação concorrentes. Outros modelos de regressão como o *Probit* e o *Logit* também foram empregados apenas para fins de validação dos resultados obtidos.

Para fins de balanceamento dos conjuntos de treinamento utilizamos o algoritmo SMOTE(CHAWLA et. al, 2002) o qual foi escolhido após a aplicação do algoritmo *racing* (POZOLLO et. al, 2013).

4 Florestas de decisão aleatórias

4.1 Árvores de decisão

Árvores de decisão consistem em uma das técnicas mais empregadas na área de mineração de dados, normalmente utilizadas para a tarefa de classificação de dados. É constituída por nós internos e nós folha dispostos em forma hierárquica (SILVA; BOSCARIOLI; PEREZ; 2016).

Uma forma de otimização deste algoritmo genérico é a inclusão de uma função de seleção do atributo que será utilizada para criar um novo nó na árvore em vez da escolha aleatória. Esta função tem por objetivo determinar qual o melhor atributo para se realizar a partição do conjunto atual.

Critérios baseados em medidas de impureza, que se referem à quantidade de classes presentes em uma partição, quanto mais classes diferentes existirem na partição mais impura ela será (SILVA; PERES; BOSCARIOLI, 2016). Dentre os vários critérios de seleção que podem ser utilizados, detalharemos o Índice Gini, que é o aplicado para a seleção da criação das árvores das nossas florestas.

O índice Gini mede as diferenças entre as probabilidades de distribuição dos atributos alvos. Sendo dado por:

$$\text{Gini}(S) = 1 - \sum_{i=1}^m p_i^2 \quad (4.1)$$

Sendo p_i a probabilidade da tupla em S pertencer à classe C_i , sendo calculada através divisão entre a quantidade de elementos de uma classe pertencentes a um conjunto dividida pelo total de itens do conjunto. Neste caso, dividindo o conjunto S em duas partições binárias com base no atributo A , S_1 e S_2 o índice Gini de S dado o particionamento é calculado por:

$$\text{Gini}_A(S) = \frac{|S_1|}{|S|} \text{Gini}(S_1) + \frac{|S_2|}{|S|} \text{Gini}(S_2) \quad (4.2)$$

O Cálculo da redução de impureza proporcionada por um determinado atributo é feito a partir da diferença entre o índice Gini do conjunto original e a soma dos índices das partições ocasionadas pelo atributo conforme a equação (2). Ou seja,

$$\Delta\text{GiniA} = \text{Gini}(S) - \text{Gini}_A(S) \quad (4.3)$$

Para seleção do atributo que será utilizado para o particionamento, seleciona-se o que maximiza a redução de impureza, ou seja, aquele que possui o maior ΔGiniA .

A seguir temos uma tabela com o resumo dos métodos de seleção de atributos e referências para aprofundamento.

Tabela 2 - Caracterização dos critérios de seleção de atributos, aplicáveis na construção de árvores de decisão, baseado em Rockak e Manmon (2008)

Critério	Resumo
Ganho de Informação	Usa a medida de entropia, da Teoria da Informação, como meio para a análise do grau de impurezas das partições geradas a partir da análise dos valores de um atributo descritivo.
Índice Gini	Usa um critério baseado em impurezas para analisar as diferenças entre as distribuições de probabilidade dos valores dos atributos de classe (rótulos).
<i>Likelihood Ration</i> <i>Chi-Squared statistics</i>	Mede a significância estatística do critério de ganho de informação.
DKM	Critério baseado em impureza, especialmente projetado para atributos de classes binárias. Gera árvores menores que o ganho de informação e o índice Gini.
Raio do ganho	Normaliza o critério do ganho de informação. Deve ser usado para selecionar atributos que obtiveram um ganho de informação médio.
Twoing	Alternativa ao índice Gini, melhorando o desempenho quando o atributo de classe tem um domínio grande.
ORT	Critério binário baseado no cálculo do ângulo entre vetores que representam a distribuição de classes na partições geradas
Kolgomorov_Smirnov	Critério binário baseado na distância de Kolgomorov_Smirnov. Possui uma versão estendida para tratar múltiplas classes e dados faltantes.
AUC Splitting	Baseado na área sobre a curva ROC, referente ao uso de cada um dos atributos descritivos.

Fonte: Silva et al.(2016)

O processo de indução das árvores de decisão envolve a escolha dos atributos que serão os novos nós, até onde estes nós irão crescer e o critério de parada da indução da árvore e poda (se houver), quando aplicada (ROCKACH; MAIMON, 2008). O algoritmo 1 apresenta um método genérico top down (raiz à folhas) de indução de árvores de decisão.

Algoritmo 1 – Algoritmo genérico *top down* para a indução de uma árvore de decisão

```

TreeGrowing (S, A, y, Função de seleção, Critérios de Parada)
Onde:
S - Conjunto de Treinamento
A - Conjunto de recursos de entrada
y - Recurso de destino
Função de seleção - o método para avaliar uma determinada
divisão
Critérios de parada - os critérios para parar o processo de
crescimento da árvore
Crie uma nova árvore T com um único nó raiz.
SE o critério de parada for atingido para o conjunto (S)
ENTÃO
    Marque T como uma folha com o valor mais comum de y em S
    como um rótulo.
SENÃO
    Procure no conjunto S o elemento que apresenta melhor
    critério da função de seleção.
    Etiqueta t com um a.
    Para cada elemento vi de a:
        Definir subárvore i = TreeGrowing(a = viS, A, y).
        Conecte o nó raiz de tT à subárvore i com
        um arco que é rotulada como vi
    FIM PARA
    FIM SE
RETURN TreePruning (S, T, y)
TreePruning (S, T, y)
Onde:
Conjunto de Treinamento S

```

```

y - Recurso de destino
T - A árvore a ser podada
FAÇA
    Selecione um nó t em T tal que a poda
    melhore ao máximo alguns critérios de avaliação.
    SE t <> 0 ENTÃO T = poda(T, t)
ATÉ t = ∅
RETORNE T

```

Fonte: Rockach e Maimon (2008)

A tabela 3 foi elaborada a partir de Rockach e Maimon (2008) e apresenta alguns algoritmos para a indução de árvores de decisão:

Tabela 3 - Resumo dos algoritmos de indução de árvores

Algoritmo	Tipo de atributos	Poda	Critério de seleção
ID3	Catégoricos	Não	Ganho de Informação.
C4.5	Catégoricos e numéricos.	Sim	Raio do ganho de informação.
CART	Catégoricos e numéricos	Sim	Twoing.
CHAID	Catégoricos	Não	Teste estatístico (F para atributos contínuos e Pearson qui-quadrado para atributos nominais)
QUEST	Catégoricos ou numéricos	Sim	ANOVA, F-test ou Levene's

Fonte: Rockach e Maimon (2008)

4.2 Florestas aleatórias de Breiman e Cutler (2001)

A fim de melhorar a acurácia de um classificador, uma técnica comum é utilizar a combinação de vários classificadores individuais, utilizando como resultado o voto da maioria deles como resultado final, obtendo assim um alto grau de acurácia (DIETRICH, 2000). Uma

floresta de decisão aleatória é uma combinação de vários preditores do tipo árvore de decisão treinadas aleatoriamente, na qual o resultado final consiste no voto da maioria.

As duas técnicas de treinamento mais usadas para esta finalidade são a *Bagging*, na qual cada árvore é treinada com um subconjunto randômico do conjunto inicial, e a otimização randômica por nó, na qual todo o espaço dos parâmetros é utilizado em parte, sendo os atributos escolhidos aleatoriamente. Estas duas técnicas podem inclusive serem utilizadas conjuntamente.

Criminisi e Shotton (2013) relacionaram as habilidades de predição das florestas com os parâmetros que mais as influenciam, quais sejam: A profundidade máxima permitida das árvores, o nível de aleatorização e o seu tipo, o tamanho da floresta, a escolha do tipo de treinamento e a escolha de características da floresta para as aplicações práticas.

Breiman e Cutler (2001) propuseram um modelo para a criação de florestas aleatórias em que cada árvore é gerada a partir do treinamento com um subconjunto randômico dos dados originais, sendo repostos a cada rodada. Ao mesmo tempo, para um conjunto de M variáveis de atributos, o número de variáveis que serão utilizadas para a criação dos nós também é randomizada, sendo bem menor que o conjunto original. A este subconjunto de atributos selecionados é aplicada o cálculo da função de seleção de atributo para nó da árvore, neste caso o índice Gini. Após a geração da árvore, as quais são criadas em toda a sua extensão, os dados restantes são classificados pela mesma, sendo este processo repetido para todas as árvores criadas. A classificação final do dado será feita pelo voto da maioria. Na fase de treinamento também é calculado o erro para cada árvore criada, utilizando como subconjunto de teste os dados que ficaram de fora da sua geração, aglomerando – os na taxa de erro *Out of Bag* (LIAW; WIENER, 2002).

Por este motivo as florestas aleatórias não necessitam de validação cruzada com um conjunto de testes em separado, uma vez que este teste já é feito internamente e retornado através da estimativa de erro *Out of Bag* - *OOB*.

Ainda também segundo Breiman e Cutler (2001) as florestas aleatórias não estão sujeitas o sobreajustamento aos dados de treinamento, o sobreajuste consiste quando um classificador fica condicionado a apresentar bom desempenho apenas para conjuntos muito semelhantes aos dados de treinamento, falhando na classificação de conjuntos ligeiramente diferentes.

Nossa principal medida de interesse no estudo é a importância das variáveis preditoras para fins de determinação da classe de consumidores de cinema. Nas florestas aleatórias esta medida para uma variável m de atributo é calculada através da permutação

aleatória dos valores possíveis da variável em análise e sua consequente submissão à árvore para classificação. A importância da variável é calculada pela média da diferença entre as classificações corretas feitas com estes dados permutados e as classificações corretas entre os dados não – permutados. Outro critério de importância também fornecido pelo algoritmo é baseado nas diminuições do índice Gini para cada variável individual em todas as árvores da floresta (BREIMAN; CUTLER, 2001).

4.3 Avaliação dos classificadores gerados com as florestas de decisão aleatórias

Para fins de avaliação dos classificadores criados com as florestas de decisão aleatórias, utilizamos as medidas obtidas a partir da matriz de confusão exemplificada na tabela 1.

Tabela 1 - Exemplo genérico de matriz de confusão

	Referência	
	Evento	Não-evento
Predição		
Evento	A	B
Não-evento	C	D

Fonte: Ticiania Justa, 2018

Quanto mais alto o valor destas métricas, melhor o desempenho do classificador gerado.

a. Sensibilidade ou taxa de verdadeiros positivos, que consiste na capacidade do sistema em prever corretamente a condição para casos que realmente pertencem à classe, dada por:

$$A/(A+C) \quad (4.1)$$

b. Especificidade, que consiste na capacidade do sistema em prever o não pertencimento à classe para os casos que realmente não a têm, cuja fórmula é:

$$D/(B+D) \quad (4.2)$$

c. Prevalência, que consiste na percentual de eventos pertencentes à classe positiva do conjunto.

$$(A+C)/(A+B+C+D) \quad (4.3)$$

d. Precisão na detecção de valores positivos.

$$PPV = (sensibilidade * prevalência) / ((sensibilidade * prevalência) + ((1 - especificidade) * (1 - prevalência))) \quad (4.4)$$

e. Precisão na detecção de valores negativos:

$$NPV = (especificidade * (1 - prevalência)) / (((1 - sensibilidade) * prevalência) + ((especificidade) * (1 - prevalência))) \quad (4.5)$$

f. Taxa de detecção

$$A/(A+B+C+D) \quad (4.6)$$

g. Prevalência de detecção:

$$(A+B)/(A+B+C+D) \quad (4.7)$$

h. Acurácia Balanceada, dada por:

$$(sensibilidade + especificidade)/2 \quad (4.8)$$

i. Acurácia, a proporção de predições corretas (positivas e negativas) em relação ao total de dados do conjunto. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema, conforme verificaremos no capítulo 5 e será contornada com o balanceamento dos dados utilizando o algoritmo SMOTE.

$$A + D/(A+B). \quad (4.9)$$

5 Balanceamento dos conjuntos originais para a execução do treinamento das florestas

Conforme veremos no capítulo 6, a análise inicial dos conjuntos revelou que eles se encontram altamente desbalanceados em relação à classe dos não consumidores de cinema, o que ocasionou a criação de modelos de baixo desempenho a partir dos dados originais, o que será explicitado no capítulo 7. Realizamos então a aplicação do algoritmo *racing* (POZOLLO et. al, 2013) para verificar qual método de balanceamento apresentaria o melhor desempenho.

5.1 *Racing*

Pozollo et al. (2013) elencaram em seu trabalho a dificuldade que os algoritmos de aprendizagem de máquina possuem de aprender de conjuntos desbalanceados ficando geralmente tendendo à classe mais frequente. Também devido ao desconhecimento inicial dos dados na maioria dos casos e às várias opções de balanceamento de conjuntos existentes, tais como *Undersampling*, *Oversampling*, *Balance Cascade*, *Easy Ensemble* entre outras e também um subconjunto de combinações de várias técnicas possíveis combinadas. Ficando difícil para o utilizador identificar inicialmente qual o método de balanceamento adequado.

Uma solução natural é aplicar várias técnicas de balanceamento, treinar vários classificadores e depois comparar o desempenho entre eles. Propuseram uma solução para o custo computacional que isto representaria, *racing* consiste em testar em paralelo um conjunto de alternativas de balanceamento e utilizar testes estatísticos para verificar quais alternativas de balanceamento são piores e descartá-las. Nossas execuções do *racing* utilizando o teste de Friedman e a avaliação dos classificadores como a área abaixo da curva ROC (FAWCETT, 2005) para os conjuntos da POF apontaram o SMOTE como a melhor técnica de balanceamento para fins de treinamento dos classificadores de florestas.

5.2 SMOTE

Chawla et al. (2002) Definiram o SMOTE como sendo “uma abordagem de super amostragem em que a classe minoritária é sobre-amostrada criando exemplos sintéticos em vez de super-amostragem da classe minoritária com substituição...”. Estes exemplos sintéticos são criados a partir de operações geradas no espaço de características dos representantes da classe minoritária.

O algoritmo SMOTE seleciona instâncias similares usando k vizinhos mais próximos de forma aleatória, ou seja, caso o aumento necessário seja de 200% e forem utilizados 5 vizinhos, o algoritmo realiza cálculos entre dois escolhidos aleatoriamente. Este cálculo é a diferença entre os vetores de características multiplicado por um número entre 0 e 1, criando um novo vetor de características do exemplo sintético da classe minoritária, repetindo o processo duas vezes para criar dois exemplares, atendendo assim a necessidade de aumento de 200%. O algoritmo 2 explana o funcionamento do SMOTE.

Algoritmo 2 – SMOTE

```

SMOTE (T, N, k)
Entradas:
T - número de exemplos da classe minoritária,
N - Percentual de exemplos sintéticos a serem gerados
k - Número de vizinhos a serem considerados
Saída:
(N/100) * T representantes sintéticos da classe minoritária

Se N < 100 Então
    Randomize a classe minoritária T
    T = (N/100) * T
    N = 100
Senão
    N = (inteiro) N/100
    k = Número de vizinhos próximos
    numattrs = Número de atributos
    Sample[][] : vetor de exemplos da classe minoritária.

```

```

    newindex : guarda o número de exemplos sintéticos
    gerados, iniciado com 0.

    Synthetic[][]: vetor de exemplos sintéticos.
Para i de 1 a T faça
    Encontre os k vizinhos mais próximos para i e guarde seus
    índices no nnarray.
Fim - para
Função popular(N, i, nnarray) (Função que gera os exemplos
sintéticos)
Enquanto N <> 0 faça
    Escolha um número randômico entre 1 e k, chame - o de nn.
    (Este passo escolhe um exemplar randômico dos k vizinhos de
    i.)
    Para attr de 1 até numattrs
        dif = Sample[nnarray[nn]][attr] - Sample[i][attr]
        gap = número randômico entre 0 e 1
        Synthetic[newindex][attr] + gap * dif
    Fim - para
    newindex ++
    N = N - 1
Fim - enquanto
Retorne

```

Fonte: Chawla et. al, 2002

A implementação do SMOTE que utilizamos foi a disponibilizada pelo pacote *unBalanced* do software R, a qual acrescenta ao algoritmo original a possibilidade de diminuição da classe majoritária através da escolha de randômica de seus representantes.

6 Dados

Partindo do processo elencado por Fayyad (1996) e já introduzido no capítulo 3, detalharemos a metodologia de estudo aplicada ao presente trabalho.

6.1 *Pesquisa de Orçamentos Familiares do IBGE*

O objetivo da Pesquisa de Orçamentos Familiares do IBGE é quantificar o consumo, gastos e rendimentos das famílias brasileiras e assim poder elaborar um perfil das condições de vida da população a partir da compreensão de seus orçamentos domésticos (CAVARARO, 2007).

A metodologia de pesquisa da POF previu a coleta de informações durante o período de um ano, utilizando para isso quatro períodos de referência, sendo eles: sete dias, 30 dias, 90 dias e 12 meses, segundo os critérios de periodicidade de aquisição e do montante do valor do gasto. A unidade amostral da pesquisa é o domicílio, caracterizada pelo IBGE como sendo a moradia estruturalmente separada e independente, constituída por um ou mais cômodos, tendo por condição a separação e a independência de acesso para ser definida como tal.

Ainda conforme Cavararo (2007), domicílio particular permanente é aquele que se destina à habitação de uma ou mais pessoas, ligada por laços de parentescos, dependência doméstica ou normas de convivência, configurado como os objetos de pesquisa da POF. A unidade de consumo da POF é a família, sendo esta unidade composta por uma ou mais pessoas no domicílio.

A POF também caracteriza os componentes da família de acordo com diversos critérios pessoais tais como: pessoa de referência (responsável pelas despesas de habitação), relação com a pessoa de referência, sexo, idade, anos de estudo, existência de pessoa com nível superior, cor ou raça, religião, posição na ocupação principal, categoria sócio-profissional principal.

As unidades de orçamento são representadas por moradores que tiveram participação no orçamento da unidade de consumo no período de 12 meses, representada por unidade de orçamento de despesa. Sendo, portanto, o morador cuja participação levou a aquisições monetárias ou não monetárias e/ou unidade de orçamento de rendimento. Neste último caso,

composta pelo morador cuja participação no orçamento representou algum tipo de ganho monetário.

Dentre as diversas apurações de despesas realizadas pelo IBGE, a de nosso interesse enquadra-se conforme a própria classificação do mesmo, em despesa com Recreação e Cultura, que inclui gastos com cinema, e as despesas com cinema especificamente.

Outras despesas correlacionadas também são importantes para nossos objetivos, tais como as relativas ao consumo de internet, DVD, TV a cabo e outros bens relacionados.

Tanto o processo de amostragem da POF 2002/2003 como o da de 2008 /2009 foram feitos a partir da amostragem geográfica e estatística, utilizando como referência os dados do Censo demográfico de 1996 e cobrem somente as regiões metropolitanas das capitais de estados e Distrito Federal.

6.1.1 Microdados da POF 2002/2003 e POF 2008/2009

Microdados são o menor nível de observação dos dados de uma pesquisa, exibindo, sob a forma de códigos numéricos, o conteúdo dos questionários, preservado o sigilo das informações. Os microdados possibilitam aos usuários, com conhecimento de linguagens de programação ou softwares de cálculo, obter estudos com avaliações de padrões de comportamento dos indivíduos.

Os microdados das POF's estão disponibilizados na forma de 14 arquivos de texto para a POF 2002/2016 e 16 para a POF 2008/2009, e são produzidos pelo Instituto Brasileiro de Estatística e Geografia (IBGE). Através da análise da documentação fornecida, identificamos os códigos de consumo referentes às despesas de internet, cinema e demais itens de interesse em ambas as POF's.

6.2 *Obtenção dos dados e extração dos registros*

Os microdados foram obtidos por download e as tabelas de registros foram extraídas utilizando o software Stata12 em conjunto com o pacote DataZoom. Neste processo extraímos os arquivos dos registros separadamente. As POFs possuem registros que incluem as condições dos domicílios, informações das pessoas residentes e suas condições de vida e os demais registros relacionados aos consumos efetivados para os períodos da pesquisa.

O próximo passo foi então cruzar estas informações para obter registros contendo informações das pessoas e suas condições de vida através da criação de chaves de identificação únicas para os domicílios e pessoas: Para fins de identificação dos domicílios as variáveis unidade da federação, número sequencial, dígito verificador e domicílio foram concatenadas. Para identificação dos indivíduos a variável número de ordem foi acrescentada à concatenação, permitindo associar o indivíduo ao domicílio. A escolha destas variáveis ocorreu após o estudo da documentação fornecida pelo IBGE e é comum às duas pesquisas, com algumas pequenas diferenças de denominação.

Foram também unificados os registros de consumo e criados chaves de identificação de pessoas e domicílios nos mesmos.

Em seguida migramos as tabelas extraídas para o banco de dados Mysql com a utilização do software Talend Open Studio for Data Integration³ a fim de selecionarmos os registros necessários e montarmos os conjuntos para a análise.

³ Disponível para download em <https://www.talend.com/products/talend-open-studio/>.

6.3 Seleção de dados

Os dados foram obtidos através de *download* no site do IBGE⁴, juntamente com a documentação dos mesmos. Conforme introduzimos no capítulo 3, para fins de seleção dos registros de consumo da POF2002 selecionamos os que representaram compra de cinema, (código 101, do quadro de consumo 7); acesso à internet (códigos 1601,1602,1603,1604,1605,1606,1607, do quadro de consumo 7); posse de microcomputador (código 3301 do quadro 13); posse e compra de DVD – aparelhos e discos (códigos 3301, 3302, 3303 do quadro 14); e acesso à TV a cabo (códigos 5701 e 5702 do quadro 15) de acordo com a documentação das POF 2002. Da POF2008 selecionamos os mesmos consumos, porém ela apresentou codificação ligeiramente diferente, quais sejam: cinema (código 101 do quadro 28); acesso à internet (códigos 00502, 00503, 00504, 00505, 00506, 01301, 01401, 01501 do quadro 6); posse de computador (códigos 06202 e 007203 do quadro 15); posse e compra de DVD – aparelhos e discos (códigos 02701, 06312, 06311 do quadro 15, código 5701 do quadro 28) e TV a cabo (códigos 00701 e 01001 do quadro 6).

6.4 Formatação

Uma vez identificados e selecionados os registros de interesse, os consumos de cinema foram totalizados por pessoa e as variáveis *dummies* para cinema, dvd, tv a cabo e posse de computador foram criadas e exportadas para o R Studio⁵, onde foram analisadas inicialmente com a análise exploratória estatística básica para fins de conhecimento inicial sobre os dados principais dos conjuntos. Posteriormente com a aplicação das florestas no R *Studio* utilizando os pacotes randomForest, randomForestExplainer, unbalanced e party, para a execução do modelo de diferenças em diferenças foi utilizado o software Stata12.

⁴ https://ww2.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=25

⁵ Disponível para download em <https://www.rstudio.com/>.

6.5 Análise exploratória e sumarização das variáveis

O conjunto de ambas as POF's possui variáveis categóricas e numéricas, as variáveis categóricas foram codificadas numericamente em ambas as POF's, para uma visão detalhada das codificações utilizadas vide documentação distribuída pelo IBGE junto aos microdados.

A variável religião, embora seja categórica foi tratada como numérica devido a grande variedade de categorias. Para a seleção das variáveis que foram analisadas escolhemos as 4 variáveis mais importantes a partir da criação de florestas aleatórias com todas as variáveis preditoras. Extraíu-se, portanto, a importância relativa como sendo a soma de todas as métricas de importância obtidas para cada variável inicial do modelo. Isto após a sua validação como classificador. Abaixo a sumarização dos valores das variáveis utilizadas na criação dos classificadores das POF's.

As tabelas 4 e 6 possuem os valores mínimo, primeiro e terceiro quartis, mediana, média e o valor máximo das variáveis numéricas. Já as tabelas 5 e 7 contém a categoria e o total de observações de cada categoria para as variáveis categóricas.

Tabela 4 - Sumarização dos valores das variáveis numéricas da POF2002 (170.938 observações)

Variável	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Ano	2002	2002	2002	2002	2002	2002
Ano de nascimento	1900	1961	1978	1974	1990	2003
Anos de estudo	0	0	4.0	4.63	8.00	17.00
Rendimentos totais do indivíduo	4.3	457.7	814.6	1566.5	1603.1	385250.7
Idade em anos	0	12	24	28	41	110
Religião	1	16	16	22	16	110
Peso em Kg	2	39.00	56	52.31	70	150
Altura em cm	1	145	158	150	167	202
Número de beneficiários do plano de saúde	0.00	0.00	0.00	0.17	0.00	10.00
Gastos com recreação	1	1	1	395	1	21919
Número de moradores do domicílio	1.0	3.0	4.0	4.7	6.0	20.0
Número de cômodos do domicílio	1.0	5.0	6.0	5.9	7.0	30.0

(continua)

(conclusão)

Variável	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Número de dormitórios	1.0	2.0	2.0	2.3	3.0	9.0
Número de banheiros	0.0	1.0	1.0	1.2	1.0	13.0
Quantidade de unidades de consumo do domicílio	1	1	1	1	1	3

Fonte: Ticiania Sá da Justa, 2018

Tabela 5 - Sumarização dos dados categóricos da POF2002

Categoria : quantidade de representantes							
Unidade da Federação	11: 3886	12: 3566	13: 4338	14: 2132	15: 7054	16: 2482	17: 3326
	21: 9203	22: 8389	23: 7872	24: 5646	25: 8523	26: 6225	27: 10988
	28: 4303	29: 8960	31: 9996	32: 7679	33: 3948	35: 6571	41: 7229
	42: 6305	43: 5503	50: 8101	51: 8089	52: 7525	53: 3099	
Sexo feminino	0: 87867	1: 83071					
Frequente escola	1: 10859	2: 50362	3: 81235	4: 27952	8: 530		
Nível de instrução	0 : 28041	1 : 1824	2 : 4817	3 : 2244	4 : 1834	5 : 90706	6 : 1219
	7 : 1202	8 : 28017	9 : 659	10: 751	11: 368	12: 498	13: 3499
	14: 4208	15: 756	16: 295				
Série que frequenta	0 : 0	1 : 18918	2 : 16770	3 : 35467	4 : 20814	5 : 13839	6 : 8057
	7 : 6949	8 : 10973	9 : 11135	10: 28016	11: 0		
Cor da pele	1: 71759	2: 8976	3: 639	4: 88526	5: 713	8: 325	9: 0
Participação como despesa no orçamento familiar	0: 0	1: 100224	2: 0	3: 66020	4: 0	5: 1781	7: 2913
Contribui com receita para o orçamento familiar	0: 0	1: 86235	2: 0	3: 81732	4: 0	5: 1198	7: 1773
Possui cartão de crédito	0: 147079	1: 23859					

(continua)

(conclusão)

Categoria : quantidade de representantes							
Possui cheque especial	0: 4	1: 14684	2:156250				
Possui plano de saúde	0:154285	1: 16653					
Reside em área urbana	0: 40280	1:130658					
Consumiu cinema	0:167497	1: 3441					
Acesso à internet	0:166581	1: 4357					
Tipo de residência	1:151155	2: 11883	3: 7093	4: 807			
Avaliação do abastecimento de água	1:112760	2: 21683	3: 2425	4: 8911	5: 12331	6: 12828	
Tipo de escoamento de esgoto	1:50328	2:34528	3:58692	4: 4302	5: 4484	6: 1116	7:17488
Condições de ocupação do imóvel	1:120238	2: 7455	3: 20559	4: 7598	5: 12765	6: 2323	
Avaliação da energia elétrica	1:157599	2: 968	3: 1415	4: 10956			

Fonte: Ticiania Sá da Justa, 2018

Tabela 6 - Sumarização dos valores das variáveis numéricas da POF2008
(178.283 observações)

	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Ano	2008	2008	2008	2008	2008	2008
Ano de nascimento	0	1963	1980	1967	1994	2009
Idade em anos	0	14	28	31	45	104
Anos de estudo	0.0	1.0	5.0	5.6	10.0	15.0
Religião	2	112	112	192	115	913
Peso em Kg	0	42	42	41	42	150
Altura em cm	0	154	154	142	154	245
Número de beneficiários do plano de saúde	0.00	0.00	0.00	0.11	0.00	18.00
Rendimentos totais do indivíduo	0	833	1412	2354	2564	117219
Gastos com recreação	1	1	1	2481	1050	21925
Número de moradores do domicílio	1	3	4	4	4	20
Número de cômodos do domicílio	1.0	5.0	6.0	6.1	7.0	27.0

(continua)

(conclusão)

	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Número de dormitórios	1.0	2.0	2.0	2.1	2.0	8.0
Número de banheiros	0.0	1.0	1.0	1.2	1.0	9.0
Quantidade de unidades de consumo do domicílio	1	1	1	1	1	4

Fonte: Ticiane Sá da Justa, 2018

Tabela 7 - Sumarização dos dados categóricos da POF2008

	Categoria: quantidade de representantes						
Unidade da Federação	11: 2840	12: 3036	13: 5230	14: 2320	15: 7200	16: 2675	17: 4274
	21: 9150	22: 7075	23: 6472	24: 4461	25: 5230	26: 7502	27: 9158
	28: 5489	29: 10083	31: 15134	32: 10271	33: 5541	35: 10676	41: 7513
	42: 5902	43: 6283	50: 6745	51: 7065	52: 7882	53: 3076	
Sexo feminino	0: 87382	1: 90901					
Frequente escola	1: 10770	2: 46789	3: 100288	4: 20436	8: 0		
Nível de instrução	0 : 78004	1 : 157	2 : 75	3 : 14	4 : 1469	5 : 21103	6 : 4150
	7 : 2491	8 : 30040	9 : 2640	10: 26898	11: 1827	12: 306	13: 22
	14: 7448	15: 1328	16: 311				
Série que frequenta	0 : 78004	1 : 9359	2 : 9114	3 : 33028	4 : 19947	5 : 8473	6 : 3733
	7 : 3921	8 : 6774	9 : 27	10: 3750	11: 2153		
Cor da pele	1: 70560	2: 13341	3: 762	4: 91662	5: 1371	8: 0	9: 587
Participação como despesa no orçamento familiar	0: 29989	1: 117260	2: 24794	3: 3793	4: 2447	5: 0	7: 0
Contribui com receita para o orçamento familiar	0: 29989	1: 97106	2: 47675	3: 1303	4: 2210	5: 0	7: 0
Possui cartão de crédito	0: 150169	1: 28114					
Possui cheque especial	0: 54783	1: 12508	2: 110992				
Possui plano de saúde	0: 143989	1: 34294					
Reside em área urbana	0: 28774	1: 149509					
Consumiu cinema	0: 174831	1: 3452					
Acesso à internet	0: 131937	1: 46346					

(continua)

(conclusão)

Categoria: quantidade de representantes								
Tipo de residência	1:170103	2: 7474	3: 706	4: 0				
Avaliação do abastecimento de água	1:147665	2: 26837	3: 3781	4: 0	5: 0	6: 0		
Tipo de escoamento de esgoto	1: 11441	2:166842	3: 0	4: 0	5: 0	6: 0	7: 0	
Condições de ocupação do imóvel	1:142390	2: 4242	3: 5115	4: 8687	5: 1046	6: 16803		
Avaliação da energia elétrica	1: 119	2:178164	3: 0	4: 0				

Fonte: Ticianá Sá da Justa, 2018

Nesta fase verificamos que os dados encontram-se altamente desbalanceados para a variável de consumo de cinema nas duas POFs, bem como a heterogeneidade dos dados.

7 Treinamento e análise das florestas geradas e suas variáveis preditoras

As medidas de avaliação dos classificadores treinados foram elencadas no capítulo 4, para fins de avaliação somente das variáveis preditoras utilizaremos as seguintes métricas:

Queda na acurácia - Diminuição da precisão de previsão após a permutação do atributo analisado.

Queda no índice Gini – Aumento da pureza do nó por divisões feitas com base nesta variável.

Profundidade mínima média – Média da profundidade mínima calculada das árvores com a variável.

Número de árvores– Número total de árvores em que uma divisão pela variável ocorre.

Número de nós – Número total de nós que usam a variável para realizar a divisão.

Veze em que foi raiz – Número de vezes que a variável foi utilizada como raiz.

P value – Teste binomial unilateral que diz se o número observado de sucessos (número de nós em que a variável foi usada na divisão) excede o número de sucessos que a variável teria se a escolha fosse aleatória.

Após a extração da importância total (soma de todas as medidas de importância) e ordenamento por importância relativa das variáveis, verificaremos o impacto da variável para fins de definição da classe utilizando a análise de dependência parcial.

A dependência é calculada através da seguinte fórmula:

$$f(x) = \log p_k(x) - \frac{1}{K} \sum_{j=1}^K \log p_j(x), \quad (7.1)$$

Onde K é o número de classes, k é a classe de interesse e p é a proporção de votos para a classe k.

Ela fornece o efeito marginal de contribuição uma variável para fins de definição da probabilidade de um indivíduo pertencer à classe desejada.

7.1 Florestas da POF2002

A floresta da POF2002 foi gerada inicialmente com 500 árvores, utilizando todo o conjunto de observações originais, sendo que cada árvore utilizou 5 variáveis para efetuar a classificação. O resultado apresentou uma taxa de erro OOB estimada em 1,99%, e demais medidas conforme as tabelas seguintes:

Tabela 8 matriz de confusão POF2002 - dados originais

	Referência		Erro da classe
Predição	0	1	
0	167337	160	0.000955241
1	3237	204	0.940714908

Fonte: Ticianá Sá da Justa, 2018

Verificamos que o erro de detecção da classe de consumidores de cinema no modelo gerado a partir dos dados originais chegou a 94%, ou seja o classificador gerado a partir dos dados desbalanceados foi ineficaz. Aplicamos então o algoritmo SMOTE (CHAWLA;NITESH, 2011) para a criação de um novo conjunto de treinamento, com as configurações abaixo e o erro foi praticamente zerado:

Tabela 9 - Configuração dos parâmetros do algoritmo SMOTE aplicado à POF2002

Parâmetro	Valor
Percentual acima (número de instâncias geradas para cada instância rara)	200
Percentual abaixo (percentual de redução da classe majoritária, sendo os exemplares selecionados aleatoriamente para cada exemplo criado com o SMOTE).	2000
Quantidade de vizinhos a serem considerados para a criação do novo exemplo.	5

Fonte: Ticianá Sá da Justa, 2018

Criando um novo conjunto de treinamento com 137.640 representantes da classe negativa e 3.441 representantes da classe positiva. O conjunto original apresentava 167.497 representantes da classe negativa e 3.441 representantes da classe positiva.

Tabela 10 - Matriz de confusão – POF 2002

	Referência		Erro da classe
Predição	0	1	
0	167298	3	0.07180
1	199	3438	0

Fonte: Ticianá Sá da Justa, 2018

Medidas de avaliação do modelo de classificação da POF2002 a partir do modelo treinado com os dados balanceados aplicados ao conjunto original da POF2002:

Tabela 11 - Medidas de avaliação do classificador da POF2002 com dados balanceados

Medida	Valor
Acurácia	0.9986
Sensibilidade	0.99971
Especificidade	0.99859
PPV	0.93580
NPV	0.99999
Prevalência	0.02013
Taxa de detecção	0.02012
Prevalência de detecção	0.02150
Acurácia balanceada	0.99915

Fonte: Ticianá Justa, 2018

A análise da importância das variáveis da POF2002 está elencada na tabela 12:

Tabela 12 - Medidas de importância das variáveis da POF 2002

Variável	Média da profundidade mínima	Número de nós	Diminuição da acurácia	Diminuição do índice Gini	Número de árvores	Veze em que foi raiz	P value
Rendimentos totais do indivíduo	2,514	139375	0,007657124	654,2648466	500	55	0
Unidade da Federação	2,952	134565	0,010653571	816,5833245	500	13	0
Peso em Kg	4,53	119316	0,006637629	408,9364145	500	2	0
Altura em cm	3,692	115064	0,005653522	394,0659893	500	8	0
Idade em anos	3,504	109406	0,014957624	411,1850181	500	1	0
Ano de nascimento	3,58	108799	0,014356069	409,0524228	500	1	0
Número de cômodos do domicílio	4,172	89421	0,003137533	286,1299411	500	15	0
Número de moradores do domicílio	4,874	82376	0,002324818	252,7396744	500	0	0
Série que frequenta	3,952	63795	0,008388125	232,9816066	500	0	0
Anos de estudo	2,588	61047	0,013874734	272,137304	500	62	0
Religião	5,01	58436	0,000924535	196,9283425	500	1	0
Número de dormitórios	5,638	56654	0,001526187	155,1521373	500	0	0
Condições de ocupação do imóvel	5,088	53550	0,001534117	178,6901395	500	0	0
Número de banheiros	3,102	46432	0,003488516	190,4751457	500	44	1
Cor da pele	5,336	40887	0,001442372	116,9835186	500	4	1
Gastos com recreação	3,724	36525	0,001027059	162,0671052	500	13	1
Nível de instrução	2,224	35720	0,013707719	316,2347859	500	71	1
Forma de escoamento do esgoto	3,226	34936	0,005429352	135,6845958	500	37	1
Contribui com receita para o orçamento familiar	4,73	30006	0,002465657	92,62797998	500	0	1

(continua)

(conclusão)

Variável	Média da profundidade mínima	Número de nós	Diminuição da acurácia	Diminuição do índice Gini	Número de árvores	Veze em que foi raiz	P value
Sexo feminino	6,662	29861	0,000826202	69,90648502	500	0	1
Frequenta escola	3,396	29211	0,004745639	140,4354559	500	24	1
Número de beneficiários do plano de saúde	5,854	25904	0,000513824	81,15372599	500	0	1
Participação como despesa no orçamento familiar	2,602	23783	0,008943637	185,4085005	500	20	1
Possui cartão de crédito	5,36	22270	0,001372565	66,29167893	500	11	1
Possui plano de saúde	6,072	19310	0,000785806	53,7571952	500	6	1
Avaliação do abastecimento de água	5,822	18351	0,001280753	49,98038473	500	3	1
Tipo de residência	2,692	16864	0,002791402	126,7677244	500	61	1
Possui cheque especial	4,752	15777	0,001484755	61,95527169	500	37	1
Acesso à internet	4,536	14884	0,000809327	62,94019389	500	11	1
Habita espaço urbano	7,048	9659	0,000452123	24,09673161	500	0	1
Avaliação do fornecimento de energia elétrica	10,595416	1516	0,0001131	4,163123016	482	0	1
Quantidade de unidades de consumo do domicílio	10,451536	1220	1,13165E-05	4,164020721	472	0	1
Ano da pesquisa		0	0	0	0	0	1

Fonte: Ticiania Justa, 2018

7.2 Florestas da POF2008

Inicialmente a floresta da POF2008 também foi gerada com os dados desbalanceados e 500 árvores, sendo utilizadas 5 variáveis para a criação de cada uma delas, resultando em um erro OOB estimado em 1.89%. Mostrando – se também ineficaz devido ao alto erro apresentado na predição de consumidores de cinema. Realizamos, então, a reamostragem do conjunto com a utilização do SMOTE obtendo assim um novo conjunto de treinamento com 3452 representantes da classe positiva e 138.080 representantes da classe negativa e fizemos a validação do classificador utilizando os dados originais, apresentando os resultados nas tabelas seguintes:

Tabela 13 - Matriz de confusão POF2008 com dados originais

	Referência		Erro da classe
Predição	0	1	
0	174672	159	0.0009094497
1	3212	240	0.9304750869

Fonte: Ticiania Justa, 2018

Tabela 14 - Médias de avaliação do classificador da POF2008 com dados balanceados aplicados ao conjunto da original.

Medida	Valor
Acurácia	0.9983
Sensibilidade	0.99971
Especificidade	0.99823
PPV	0.91758
NPV	0.99999
Prevalência	0.01936
Taxa de detecção	0.01936
Prevalência de detecção	0.02110
Acurácia balanceada	0.99897

Fonte: Ticiania Justa, 2018

Tabela 15 - Medidas de importância das variáveis da POF 2008

Variável	Média da profundidade mínima	Número de nós	Diminuição da acurácia	Diminuição do índice Gini	Número de árvores	Veze em que foi raiz	P value
Gastos com recreação	1,944	144624	0,012607096	1082,676	500	66	0
Rendimentos totais do indivíduo	2,988	128001	0,006066462	553,5732	500	31	0
Unidade da Federação	3,228	127504	0,01117995	851,94	500	4	0
Ano de nascimento	3,666	105517	0,01111488	398,5877	500	2	0
Idade em anos	3,592	104849	0,01123941	394,0701	500	3	0
Número de cômodos do domicílio	4,248	84278	0,00302537	273,2202	500	20	0
Número de moradores do domicílio	4,918	75321	0,00202101	228,5414	500	0	0
Religião	4,57	73436	0,00120725	267,9708	500	0	0
Anos de estudo	2,382	55260	0,00859213	277,4749	500	81	0
Altura em cm	4,556	54178	0,0010517	185,721	500	0	0
Número de dormitórios	4,956	54110	0,0017349	150,9461	500	6	0
Condições de ocupação do imóvel	4,846	47152	0,00124207	154,4229	500	3	0
Cor da pele	5,01	44945	0,00138338	140,4899	500	2	1
Série que frequenta	3,862	44399	0,00609930	184,9794	500	0	1
Número de banheiros	3,632	41811	0,00310613	159,7816	500	29	1
Peso em Kg	5,134	36176	0,00065253	122,6014	500	0	1
Nível de instrução	2,818	33205	0,00901559	190,7675	500	48	1
Sexo feminino	6,168	33081	0,00161997	82,85616	500	0	1
Acesso à internet	4,642	25987	0,00199808	83,76735	500	13	1

(continua)

(conclusão)

Variável	Média da profundidade mínima	Número de nós	Diminuição da acurácia	Diminuição do índice Gini	Número de árvores	Veze em que foi raiz	P value
Contribui com receita para o orçamento familiar	4,682	25297	0,00227903	90,38584	500	4	1
Número de beneficiários do plano de saúde	6,208	22683	0,00031942	69,97775	500	1	1
Possui cartão de crédito	4,394	19784	0,00231598	74,4122	500	30	1
Frequenta escola	3,508	18531	0,00331069	109,215	500	8	1
Possui cheque especial	4,44	17523	0,00175831	62,29674	500	22	1
Possui plano de saúde	3,308	17194	0,003436447	109,4821	500	47	1
Participação como despesa no orçamento familiar	3,358	15339	0,002343985	92,46762	500	19	1
Avaliação do abastecimento de água	6,24	14901	0,000546315	39,34185	500	0	1
Tipo de residência	2,732	13631	0,002006426	124,228	500	60	1
Forma de escoamento do esgoto	7,346	13535	0,000173193	35,98146	500	0	1
Habita espaço urbano	6,37	10629	0,000505106	28,18032	500	1	1
Avaliação do fornecimento de energia elétrica	10,5821	875	1,02935E-06	3,929504	431	0	1
Quantidade de unidades de consumo do domicílio	11,7808	810	3,77291E-06	2,721963	408	0	1
Ano da pesquisa	0	0	0	0	0	0	1

Fonte: Ticianá Sá da Justa, 2018

7.3 Classificação dos dados da POF2008 a partir do modelo da POF2002

Realizamos a classificação dos dados da POF2008 a partir do modelo criado com os dados da POF2002 e obtivemos as seguintes medidas de desempenho:

Tabela 18 - Medidas de avaliação do classificador do modelo da POF2002 aplicado aos dados da POF2008

Medida	Valor
Acurácia	0.9751
Sensibilidade	0.9902
Especificidade	0.2086
PPV	0.9845
NPV	0.2964
Prevalência	0.9806
Taxa de detecção	0.9711
Prevalência de detecção	0.9864
Acurácia balanceada	0.5994

Fonte: Ticiania Justa, 2018

Verificamos que as medidas de desempenho dos dados classificados a partir do modelo treinado com a POF2002 se apresentaram piores na tarefa de classificação, tendendo a não identificar corretamente os indivíduos não consumidores de cinema. O que reforça a necessidade de treinamentos temporais do classificador e a sua contínua avaliação.

Ou seja, um modelo treinado em com informações de 2002 tende a não ser eficiente quando aplicado a dados com grande lapso temporal. Esta ineficiência deveu-se à alteração do *ranking* de importância das variáveis preditoras.

7.4 Comparação entre as variáveis de classificação e seleção das variáveis de análise do perfil de consumo.

A tabela 19 apresenta o *ranking* das variáveis de ambas as POF's e o incremento de sua importância para fins de classificação da POF 2002 até a de 2008.

Verificamos que cerca de metade das variáveis mantiveram-se estáveis com variação de colocação entre -2 e 2, mostrando-se estáveis para a tarefa de classificação ao longo do tempo.

Tabela 19 – Incremento da importância das variáveis

Variável	POF2008	POF2002	Incremento
Gastos com recreação	1	16	15
Rendimentos totais do indivíduo	2	1	-1
Unidade da Federação	3	2	-1
Ano de nascimento	4	6	2
Idade em anos	5	5	0
Número de cômodos do domicílio	6	7	1
Número de moradores do domicílio	7	8	1
Religião	8	11	3
Anos de estudo	9	10	1
Altura em cm	10	4	-6
Número de dormitórios	11	12	1
Condições de ocupação do imóvel	12	13	1
Cor da pele	13	15	2
Série que frequenta	14	9	-5
Número de banheiros	15	14	-1
Peso em Kg	16	3	-13
Nível de instrução	17	17	0
Sexo feminino	18	20	2
Acesso à internet	19	29	10
Contribui com receita para o orçamento familiar	20	19	-1
Número de beneficiários do plano de saúde	21	22	1
Possui cartão de crédito	22	24	2
Frequenta escola	23	21	-2
Possui cheque especial	24	28	4
Possui plano de saúde	25	25	0
Participação como despesa no orçamento familiar	26	23	-3
Avaliação do abastecimento de água	27	26	-1
Tipo de residência	28	27	-1
Forma de escoamento do esgoto	29	18	-11
Habita espaço urbano	30	30	0
Avaliação do fornecimento de energia elétrica	31	31	0
Quantidade de unidades de consumo do domicílio	32	32	0
Ano da pesquisa	33	33	0

Fonte: Ticianá Sá da Justa, 2018

8 Análise das variáveis preditoras das florestas

Detalharemos inicialmente as 4 variáveis mais importantes da POF de 2008 com relação à soma das importâncias elencadas, comparando os resultados obtidos com as mesmas variáveis para a POF de 2002.

Em seguida veremos as 5 variáveis que mais perderam importância e na última seção detalharemos as variáveis mais importantes apenas levando em consideração a importância de permutação, a qual é menos sujeita à viés de seleção.

8.1 Análise das 4 variáveis com maior importância relativa total

8.1.1 Gastos com recreação

Os gastos com recreação obtidos pela pesquisa de 2002 tanto da parte dos consumidores como não consumidores de cinema não apresentaram uma grande relevância, possuindo todos os limites inferior e superior e os quartis e a mediana iguais a 1 (talvez por algum erro de imputação).

Figura 10 - Gráfico da contribuição marginal da variável gastos com recreação da POF2002

Fonte: Ticiane Sá da Justa, 2018

Os gastos com recreação declarados pela pesquisa da POF2002 apresentaram uma contribuição variando entre -2 e -10 na contribuição para a definição da classe de consumidores de cinema.

Observamos também que o crescimento da curva é mais suave que a da POF2008, o que indica que a proporcionalidade da contribuição para a definição da classe de consumidores de cinema não cresce na velocidade em que crescem os gastos com recreação, ou seja, os gastos com recreação computados foram para outras formas predominantes de consumo que não o cinema pelos dados da POF2002.

Figura 11 - Representação em Boxplot da distribuição dos gastos com recreação da POF2008

Fonte: Ticiania Sá da Justa, 2018

O gráfico em boxplot dos gastos com recreação da POF2008 evidencia que o valor dispendido entre os consumidores de cinema com recreação e cultura apresentou predominância sobre os não consumidores apresentando um limite inferior de R\$854,00 primeiro quartil iniciando em R\$ 8618.5, mediana de R\$ 14035.0 fim do quarto quartil em 18660.5 e limite superior de 21918.0. Como foi excluído dos gastos com recreação e cultura o gasto com cinema, este resultado sugere, em conjunto com a POF2002, que em 2008 as preferências reveladas por cultura em geral (via dispêndio) são o elemento mais importante para explicar o consumo de cinema.

Figura 12 - Gráfico da contribuição marginal da variável gastos com recreação da POF2008

Fonte: Ticianá Sá da Justa, 2018.

A margem máxima de contribuição da variável gastos com recreação aumentou de -6 a 0 e também a curva ficou mais acentuada, indicando que a probabilidade de um indivíduo ao decidir gastar com recreação e cultura escolha o cinema como opção é maior em na POF2008 que na POF2002.

No geral, o gasto médio com recreação e cultura também aumentou de R\$ 395,01 para R\$ 2.481, em valores não deflacionados. Em valores deflacionados essa diferença diminui de R\$553 e R\$ 2.481 em reais correntes de 2008.

8.1.2 Rendimentos totais do indivíduo:

Figura 13 - Gráfico boxplot dos rendimentos totais da POF2002

Fonte: Ticiania Sá da Justa, 2018

Na POF2002 a maior parte dos não consumidores de cinema se enquadraram na faixa de renda entre R\$ 452,00 a R\$ 1.541,00 com limite inferior de R\$ 4,00 e superior de R\$ 3.175,00 e mediana de R\$ 798,00. Já entre os consumidores de cinema da POF2002 a faixa principal de renda ficou entre R\$ 1.780,00 e R\$ 6.735,00, com limite inferior de R\$ 50 e superior de R\$ 14.162,00 e mediana de R\$ 3.378,00. Ficando claro que renda importa para consumo de cinema.

Figura 14 - Gráfico da contribuição marginal da variável rendimentos totais da POF2002

Fonte: Ticiania Sá da Justa, 2018

O gráfico de dependência parcial evidencia uma relação positiva entre o consumo do cinema e a elevação da renda em 2002.

Figura 15 - Gráfico boxplot dos rendimentos totais da POF2008

Fonte: Ticiania Sá da Justa, 2018

Na POF2008 o padrão de 2002 se repete, de modo que os rendimentos dos não consumidores de cinema ficaram em sua maior parte entre R\$ 824,00 e R\$ 2.500,00 com

mediana de R\$ 1.390 e limite superior de R\$ 5.023. Já entre os consumidores a maior faixa (1º e 3º quartis) ficou entre R\$ 2.242 e R\$ 7.781,05 com média de R\$ 4.085 e limite superior de R\$ 1.6073. Ou seja, houve um aumento na distribuição da faixa de renda dos não consumidores.

Figura 16 - Gráfico da contribuição marginal da variável rendimentos totais da POF2008

Fonte: Ticiania Sá da Justa, 2018

De fato, a renda individual mostrou-se fortemente associada ao consumo de cinema em ambas as POFs, porém na POF de 2008 apresentou um comportamento peculiar de diminuição na contribuição da margem de probabilidade de definição da classe de consumidores. Apresentou uma leve queda da margem para a renda em torno de R\$ 8.000 e 12.000, esta queda pode indicar que consumidores de mais alta renda estão partindo para outros meios de entretenimento. Uma vez que a POF apresenta 348 opções entre os itens de recreação e cultura, um estudo específico seria necessário para determinar quais meios de entretenimento são preferidos por faixa de renda.

8.1.3 Unidade da Federação

Figura 17 - Distribuição dos consumidores de cinema por UF

Fonte: Ticianá Sá da Justa, 2018

Em ambas as POFs, verificamos que a variável Unidade da Federação ocupou a 5ª posição de importância, embora o consumo registrado difira bastante da POF2002 para a POF2008, não tivemos como mensurar se esta mudança se deveu a uma mudança na oferta do parque exibidor, uma vez que a informação de salas de exibição por estado e ano não está disponibilizada pela Ancine na data de 2002.

Figura 18 - Gráfico da contribuição marginal da variável UF da POF2002

Fonte: Tician Justa, 2018

Figura 19 - Gráfico da contribuição marginal da variável UF da POF2008

Fonte: Tician Sá da Justa, 2018

Analisando os gráficos de dependência parcial das POFs verificamos que a contribuição marginal da variável era mais relevante na POF2002, pois o seu mínimo era -5 ao passo que na POF2008 seu mínimo foi de -7. Verificamos também uma competição mais

acirrada entre as margens de contribuição da variável para a definição da classe na POF2002, com amplitude entre os valores maior que o de 2008. Tomando por exemplo o Distrito Federal, a probabilidade de uma pessoa frequentar cinema em 2002 era maior do que em 2008, se considerarmos apenas a análise gráfica sem levar em consideração mudanças populacionais e a oferta, o que pode ser visualizado na frequência absoluta com a queda do consumo.

A diminuição da margem de contribuição da variável de 2002 a 2008 e sua relativa uniformização da amplitude sugerem uma oferta mais generalizada de salas de cinema pelo país, de fato, a oferta de salas subiu de 1635 para 2278.

8.1.4 Idade

A POF2002 teve a maior parte dos seus consumidores de cinema entre a faixa etária de 19 a 36 anos, conforme pode ser visualizado no gráfico boxplot.

Figura 20 – Gráfico boxplot da variável idade da POF2002

Fonte: Ticiania Sá da Justa, 2018

Figura 21 – Gráfico de dependência parcial da variável idade da POF2002

Fonte: Ticianá Sá da Justa, 2018

O gráfico de dependência parcial sugere que para os indivíduos da POF2002 há uma tendência de queda no consumo de cinema a partir da idade aproximada de 16 anos, havendo novamente um acréscimo no público a partir dos 50 anos aproximadamente.

Figura 22 – Gráfico boxplot da variável idade em 2008

Fonte: Ticianá Sá da Justa, 2018

O gráfico boxplot da idade em 2008 indicou que a maioria dos consumidores de cinema ficou entre 20 e 37 anos, ou seja, houve aumento de 1 ano na faixa etária.

Figura 23 – Gráfico da dependência parcial da idade em 2008

Fonte: Ticianá Sá da Justa, 2018

O gráfico de dependência parcial da variável idade também indica um aumento na idade dos consumidores e uma maior frequência de idosos ao cinema, conforme pode ser visto pelas áreas das curvas sob o gráfico.

8.2 Análise das variáveis que mais perderam importância

Relembrando o conceito do índice Gini, quanto maior o grau de pureza dos conjuntos finais significa que a variável é a melhor escolha naquele momento. Ou seja, quando uma variável se torna comum tanto aos representantes como não representantes, ela diminui seu índice Gini e demais medidas de importância elencadas anteriormente.

Verificamos uma queda maciça da importância das seguintes variáveis:

Tabela 20 - Variáveis que mais perderam importância

Variável	Aumento da Importância da variável
Peso em Kg	-13
Forma de escoamento do esgoto	-11
Altura	- 6
Série que frequenta	- 5
Participação como despesa no orçamento familiar	-14

Fonte: Ticiania Justa, 2018

A queda da importância da variável forma de escoamento de esgoto indica que na POF2008 o valor da mesma tendeu a se uniformizar o que foi visto na sumarização dos dados, enquanto na POF 2002 todas as categorias (Rede Geral de Esgoto ou Pluvial, Fossa Séptica, Fossa Rudimentar, Vala, Direto para o Rio/ lago ou mar, Outra Forma e Não Tem) estiveram presentes na coleta, na POF2008 apenas as categorias de Rede Geral de Esgoto ou Pluvial e Fossa Séptica estiveram presentes. Isto sugere uma maior capilaridade da malha de escoamento de esgoto de 2002 a 2008.

A variável de participação como geração de despesa no orçamento familiar apresentou uma maior variabilidade das suas categorias, estando presente em 2008 registros de pessoas que se auto declararam não – participantes como despesas no orçamento familiar, ao passo que em 2002 este valor foi igual a 0.

A variável peso, que junto com a variável altura constituem uma *proxy* para o estado de saúde do indivíduo, perdeu em importância devido à diminuição da sua variância que foi de 4.69 para 2.27 na população estudada. Da mesma forma a variável altura e série que frequenta, que tiveram queda de 42,64 para 26,48 no caso da altura e 9,51 para 7,32 para a série que frequenta.

8.3 Análise das variáveis com relação apenas à importância de permutação

Considerando que o índice Gini é enviesado com relação às variáveis contínuas e com muitas categorias, e por isso adiciona viés de seleção na importância relativa total calculada, resolvemos verificar quais foram as variáveis mais importantes apenas do ponto de vista do erro de permutação (*mean decrease acuracy*).

Com base apenas na importância de permutação, temos que as 10 variáveis mais relevantes para a POF2008 para a definição da classe foram, em ordem decrescente de importância:

Tabela 21 - Variáveis da POF2008 em ordem decrescente de importância de permutação.

Variável POF2008
Gastos com recreação
Idade em anos
Unidade da Federação
Ano de nascimento
Nível de instrução
Anos de estudo
Série que frequenta
Rendimentos totais do indivíduo
Possui plano de saúde
Frequenta escola

Fonte: Ticiania Sá da Justa, 2018

Dentre as variáveis elencadas 4 se relacionam com a escolaridade do indivíduo e variam nas suas formas de representação (categóricas e numéricas) e faixa de valores possíveis, embora apresentem a mesma informação, sugerindo uma forte relação entre a escolaridade do indivíduo e o consumo de cinema. Por este motivo realizamos a análise mais detalhada da variável anos de estudo para as POFs de 2002 e 2008. Comparando os gráficos de dependência parcial e *boxplot* da POF2002 e POF2008 verificamos que a quantidade de anos de estudo está diminuindo de relevância ao longo do tempo.

Figura 24 – Gráfico boxplot de anos de estudo da POF2002

Fonte: Ticiania Sá da Justa, 2018

A análise do gráfico boxplot da variável anos de estudo em 2002 mostra uma dicotomia clara entre a escolaridade de consumidores e não-consumidores de cinema. Sendo que a maior faixa de consumidores se encontra entre os 8 e 14 anos de estudo, com mediana de 11 anos, enquanto os não consumidores enquadram-se em sua maioria entre 0 e 8 anos de estudo, com mediana de 4 anos de estudo. Para os dois conjuntos o limite inferior foi de 0 e superior 17.

Figura 25 – Gráfico dependência parcial da variável anos de estudo da POF2002

Fonte: Ticiania Sá da Justa, 2018

O gráfico de dependência parcial evidencia esta tendência apresentando uma relação fortemente positiva entre os anos de estudo e o aumento da probabilidade de se consumir cinema.

Figura 26 – Gráfico boxplot de anos de estudo da POF2008

Fonte: Ticiania Sá da Justa, 2018

Em 2008, a quantidade de anos de estudo dos não consumidores de cinema ficou entre 0 e 10 anos de estudo, enquanto a escolaridade dos consumidores passou de 10 a 15 com mediana de 11 anos, ou seja, a escolaridade para o público – alvo de consumidores de cinema aumentou.

Figura 27 – Gráfico dependência parcial da variável anos de estudo da POF2002

Fonte: Ticianá Sá da Justa, 2018

9 Conclusões do estudo realizado com as florestas aleatórias

A fim de concluirmos este primeiro estudo, voltaremos aos objetivos específicos do trabalho relacionados à extração de informação e suas conclusões.

9.1 *Verificar se houver mudança no perfil dos consumidores de cinema antes e após o advento da internet banda larga.*

Embora não tenhamos como determinar se a mudança do perfil dos consumidores foi exclusivamente devido à introdução da internet verificamos que sim, houve alteração no perfil dos consumidores estudados entre o período.

Inicialmente esta mudança é vista através da importância das variáveis elencada na tabela comparativa através da mudança de importância das variáveis.

Através da análise das cinco variáveis elencadas para a definição da classe de consumo de cinema da POF2008 obtivemos os principais achados resumimos a seguir.

Verificamos que os consumidores de cinema da POF2008 consumiram mais com recreação e cultura (considerando que os dados da POF2002 estão corretos), pessoas com maior poder aquisitivo tenderam a buscar outros meios de entretenimento que não o cinema pelos dados da POF2008 e a idade dos mesmos teve um pequeno aumento, inclusive com a maior presença de consumidores da terceira idade nos cinemas. O estudo também revelou uma migração de consumidores de cinema entre as unidades da Federação.

9.2 *Verificar se o perfil de consumidores de cinema difere do dos não consumidores*

Conforme as análises elencadas anteriormente, o perfil dos consumidores e não consumidores de cinema difere principalmente em nível de renda e anos de estudo.

9.3 Avaliar o desempenho da aplicação das florestas aleatórias para fins de extração de informações relevantes sobre consumidores.

A importância relativa como sendo a soma total das importâncias computadas não forneceu informações suficientes para realizarmos a dicotomia entre os consumidores e não consumidores de cinema.

Somente com a análise de dependência parcial podemos ver claramente as regiões em que os consumidores de cinema preponderam para cada variável analisada, constituindo um meio claro de se obter informação acessível sobre a importância e o comportamento da variável para fins de determinação da classe e separação entre consumidores e não consumidores.

Strobl (2007) cita em seu trabalho o problema do viés do índice Gini para a escolha e determinação das variáveis mais importantes, o qual tende a escolher variáveis contínuas e com maior quantidade de categorias para realizar as partições. Em vez de utilizar o índice Gini, as árvores das florestas utilizadas por ele utilizam o particionamento com base em análise condicional e testes estatísticos de hipóteses. Embora não seja o objeto deste trabalho, tentamos verificar o impacto do viés do índice Gini para a importância das variáveis da POF2008 criando uma floresta com árvores condicionais, porém não obtivemos sucesso devido às limitações computacionais. Em um computador com 16GB de memória RAM, Processador Core i7 e HD de 500GB SSD após 48h de atividade o processo foi abortado pelo sistema devido à falta de recursos.

Embora não tenhamos conseguido avaliar as importâncias das variáveis entre as florestas condicionais de Hothorn et al. (2006) e as florestas de Breiman e Cutler (2001), verificamos que houve muita semelhança entre os achados da técnica de aplicada e o trabalho de Nieuwenhoff (2017). Que trabalhando com as técnicas de *Support Vector Machine* e *Nayve Bayes* chegou a conclusões semelhantes sobre os determinantes do consumo de cinema referente ao ano de 2018. Apresentou três variáveis semelhantes selecionadas como determinantes da classe, quais sejam: despesas com recreação, anos de estudo e renda per capita, validando assim a importância das variáveis selecionadas para análise.

7.4. Síntese da Conclusão

Concluimos que a metodologia apresentada é satisfatória para a determinação das variáveis de interesse de estudo e seus respectivos impactos na determinação da classe. Os resultados apontam uma mudança no padrão de variáveis mais importantes para classificar os consumidores e não consumidores, quando se verifica que, em 2008, os principais determinantes do consumo são as preferências por bens culturais, oferta de cinemas educação e renda, tendo havido portanto uma mudança relativa a 2002 na posição dos principais determinantes. A capacidade de classificar eficientemente entre consumidor e não consumidor em cada período, permite a distinção clara entre o padrão de consumidores e não consumidores em cada ano.

10 Modelo de regressão diferenças em diferenças

10.1 Experimentos naturais

Os chamados experimentos naturais ou quase experimentos (MAFFIOLI, 2011) acontecem quando algum evento exógeno muda o ambiente no qual os indivíduos, famílias, instituições ou municípios estão inseridos.

No caso da nossa pesquisa, esta mudança foi a política de privatização do setor de telecomunicações, a qual através Plano Geral de Metas para a Universalização do Serviço Telefônico Fixo Comutado (PGMU - previsto na Lei 9.998/00 e implementada entre 2002 e 2009) ocasionou que a rede telefônica do país (mais claramente os fios telefônicos) fossem largamente distribuídos no território nacional. Essa disseminação da rede de fios possibilitou que a tecnologia ADSL (*Asymmetric Digital Subscriber Line*), um dos meios de distribuição de internet de alta velocidade, pudesse ser disponibilizada em massa pelas operadoras sem custo de infraestrutura inicial muito alto, possibilitando a oferta do serviço a um valor acessível à população, iniciando assim o mercado de internet banda larga no qual a competição entre as operadoras envolve velocidade de conexão e preço.

Para avaliar um experimento natural por meio de análise de intervenção (modelos de diferenças em diferenças neste caso) é necessário que se tenham dois grupos distintos observados em dois períodos distintos (antes e depois da intervenção – choque ou mudança de política), o primeiro de controle, que tenha sido submetido à mudança de política e o segundo que não tenha sido afetado pelos os efeitos da mesma.

No caso deste trabalho, os grupos analisados são os indivíduos das POF2002 e POF2008 que consumiam internet e portanto sofreram as consequências do PGMU no quesito acesso à banda larga na passagem entre 2002 e 2008, constituindo o grupo de tratamento, e os não consumidores de internet representando o grupo de controle.

Ainda para fins de estudos quase - experimentais os grupos devem ter características semelhantes, diferindo apenas no recebimento da política. Como no Brasil, em geral, quem consome internet possui características distintas de quem não consome, usamos variáveis de controle como renda e muitas outras, bem como controles por coortes de idade e local de nascimento para garantir a homogeneidade dos grupos.

Contrariamente aos experimentos reais, nos quais os grupos de tratamento e controle são escolhidos aleatoriamente para impedir viés nas estimativas, os grupos em um experimento natural aparecem da maneira com que a mudança é realizada.

Sendo assim, a amostra é dividida em quatro grupos: (1) grupo de controle antes da política; (2) grupo de controle depois da política; (3) grupo de tratamento antes da política e (4) grupo de tratamento depois da política.

10.2 Diferenças em diferenças

O modelo de diferenças em diferenças estima a diferença dentro de cada grupo (tratamento e controle) e a diferença dessas duas médias com a aplicação de testes de hipóteses (KMENTA, 1978).

Tabela 22 – Modelo de diferenças em diferenças

Grupo	Antes	Política	Depois	Diferenças
Tratamento	T_0	x	T_1	$T_1 - T_0$
Controle	C_0		C_1	$C_1 - C_0$
Diferenças	$T_0 - C_0$		$T_1 - C_1$	$(T_1 - T_0) - (C_1 - C_0)$ Ou $(T_1 - C_1) - (T_0 - C_0)$

Fonte: Amaral, 2013

Realizando o cálculo $(T_1 - T_0) - (C_1 - C_0)$, temos a diferença entre as diferenças nos dois períodos em cada grupo. Já fazendo $(T_1 - C_1) - (T_0 - C_0)$ encontramos a diferença entre os grupos em cada um dos períodos.

Matematicamente o método é representado da seguinte forma:

$$DD = (Y_{T1} - Y_{C1}) - (Y_{T0} - Y_{C0}) \quad (10.1)$$

Onde Y representa a media da variável estudada para cada ano e grupo, a letra C representa o grupo controle e a letra T o grupo de tratamento. Representando o método através de uma regressão e criando as variáveis indicadoras (*dummies*). T_i igual a 1 para os indivíduos do grupo de tratamento e 0 para o grupo de controle, t_i igual a um quando o dados

forem do momento pós-política e zero para o momento pré-política. Chega-se então na seguinte expressão (AMARAL, 2013):

$$Y_i = \alpha + \beta T_i + \gamma t_i + \delta(T_i * t_i) + \varepsilon_i \quad (10.2)$$

Onde Y representa a variável estudada, α o valor esperado para Y antes da mudança (termo constante), β é o impacto do grupo de tratamento na variável Y, γ impacto do período posterior sobre Y, δ é o impacto pós-evento do grupo de tratamento em relação ao grupo de controle, sobre Y, ε_i sendo o erro aleatório não observado. Outros fatores de controle são necessários a fim de garantir que δ nos dará o impacto da política externa, isto é feito através da inclusão de variáveis de controle relevantes de acordo com o caso em estudo.

10.3 Avaliação dos modelos de regressão

Os modelos de regressão utilizados neste estudo se enquadram na família dos modelos de regressão lineares (FAVERO et. al, 2009). Os quais possuem a formulação geral dada pela equação:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + u \quad (10.3)$$

Na qual Y é a variável dependente, β_k são os coeficientes de cada variável explicativa X_k e u é o termo de erro.

Uma das medidas do poder explicativo do modelo de regressão é dada pelo coeficiente de ajuste R^2 que mede o quanto o comportamento das variáveis X_i explicam a variação de Y, sendo dado pela equação 3.11.

$$R^2 = 1 - \frac{Var(u)}{Var(Y)} \quad (10.4)$$

O R^2 pode variar de 0 a 1e quanto maior menor o resíduo do modelo e mais explicativo ele se torna.

Encontrando um poder explicativo satisfatório, a importância de cada variável X_i para o resultado final pode ser auferida pelo valor do coeficiente encontrado para ela na regressão.

11 Estudo específico sobre efeitos da banda larga sobre consumo de cinema no Brasil utilizando diferenças em diferenças

11.1 Introdução

Aproveitamos a adoção da banda larga no Brasil, oficialmente em 2006, para estudar o efeito do acesso à Internet sobre o consumo de filmes em salas de cinema. Nós empregamos um conjunto muito grande de microdados com 300.000 indivíduos nos anos de 2002/3 e 2008/9. Explorando a estrutura de um modelo de diferença-em-diferenças (DD) aplicado a um pseudo painel controlando por coortes de idade, gênero e estado de nascimento, realizamos uma análise de intervenção para o mercado brasileiro de filmes. O Brasil é um importante caso a ser estudado porque é extremamente desigual, com grande exclusão digital e cultural. Além disso, somente algo em torno de 10% do que é consumido é produzido domesticamente, sendo um caso bastante representativo da maioria dos países em desenvolvimento ou subdesenvolvido (exceto China e Índia).

O uso da Internet de banda larga aumentou substancialmente no Brasil após 2002, o que pode ser ilustrado pelo número de acessos⁶. Entre 2002 e 2008, a taxa de crescimento médio anual foi de 49% em acesso fixo. Antes de 2011, no entanto, a velocidade mínima de banda larga no país não estava oficialmente estabelecida, enquanto em outros países o padrão na época era de velocidades superiores a 256Kbps. A entrada de banda larga (acesso à Internet com velocidades entre 512Kbps e 2Mbps) no país ocorreu entre nossos dois anos de observações. Segundo a ANATEL (agência reguladora de telefonia brasileira), em 2002 a

⁶ <http://www.caminhosdabandalarga.org.br/2012/10/capitulo-1>

média de acessos mensais de banda larga foi de 776.718, enquanto para 2009 a mesma média foi de 12.199.824 (Figura 1).

A Internet de banda larga trouxe uma enorme inovação para os veículos de transmissão de informações, o que possibilitou a redução dos custos de transmissão (OBERHOLZE-GEE;STRUMPF,2007), uma vez que aumentou a velocidade, o processamento e o volume de transporte de dados.

A mudança tecnológica de um filme em película para um padrão de filme digital também implicou uma redução muito significativa no custo de produção, reprodução e distribuição para a indústria cinematográfica. Tais mudanças tecnológicas permitiram a liberação simultânea mundial de filmes, acesso ao conteúdo em casa, várias versões digitais com finais alternativos, cortes de diretor e várias legendas em diferentes idiomas. Como o filme é um bem de informação, teve reduzido seus custos de transmissão / reprodução pela banda larga da Internet e experimentou um aumento no compartilhamento não pago de seus arquivos (HUSAK,2004).

Rob e Waldfogel (2007) vão além da pirataria para argumentar que a substituição entre a Internet e o cinema depende dos tipos de atividades que os indivíduos envolvidos têm na Internet. Waldfogel (2017) argumenta que a Internet promove novos produtos, ocasionando sua distribuição global - por meio do Netflix, Amazon, Spotify, iTunes e outros - e possivelmente fomenta a convergência do consumo entre as regiões. Além disso, as redes sociais, como o Facebook e o Twitter, por exemplo, são novos produtos de entretenimento que competem com filmes e músicas no tempo gasto.

Figura 29 – Acesso à Internet em banda larga de acordo com a velocidade de transmissão no Brasil

Fonte: Ticiania Justa, 2018

Vários estudos exploram o efeito da pirataria de música devido a mudanças tecnológicas (tendo a Internet como o canal principal) no mercado de música legal. A maioria deles encontra evidências de efeitos negativos (Hui e Png 2003; Bishop 2004; Bounie et al. 2006; Zentner 2006; Rob e Waldfogel 2007; Liebowitz 2008; Waldfogel 2010; Danaher et al. 2014). As exceções são Oberholzer-Gee e Strumpf (2007) e Smith e Telang (2009) que encontram um efeito nulo.

Em relação aos filmes, De Vany e Walls (2007) encontram um efeito negativo dos downloads da Internet em filmes nos cinemas. Rob e Waldfogel (2007) encontraram um forte efeito de substituição entre filmes não remunerados e filmes pagos, com base em uma amostra de 500 estudantes de graduação usando o consumo de banda larga e DVD como variáveis instrumentais. Liebowitz e Zentner (2009) constataam que a Internet reduz o consumo de televisão entre a geração mais jovem que cresceu desde que o computador pessoal foi desenvolvido, mas pouco ou nenhum efeito sobre o mais antigo. Seus resultados, no entanto,

estão ligados à substituição de atividades na Internet, que incluem negócios, atividades de entretenimento e conexões sociais, por exemplo.

11.2 Dados e Estratégia Empírica

A adoção da banda larga foi usada para estudar os efeitos da Internet sobre o consumo nos cinemas como um experimento quase natural, isso porque a adoção da banda larga no Brasil ou em qualquer outro país depende da disponibilidade de tecnologia de países produtores, bem como de decisões políticas, no nosso caso a privatização do setor de telecomunicações e o plano de metas de universalização do acesso.

Empregamos dados de pesquisa de consumo em nível individual referente aos anos de 2002/3 e 2008/9, incluindo todas as áreas metropolitanas do IBGE (Instituto Brasileiro de Estatística e Geografia) sob o nome POF (Pesquisa de Orçamento Familiar), que tem como objetivo verificar os padrões de consumo. Empregando um modelo Diferença-em-Diferenças (DD), nosso grupo de tratados foi composto por indivíduos que têm acesso à Internet, e nosso grupo de controle pelas pessoas sem acesso à Internet. No entanto, como temos um pseudo-painel para observações individuais, seguimos Deaton (1985) e controlamos por coortes de ano de nascimento, estado de residência e sexo.

Nós empilhamos os dados para estimar o modelo DD de consumo de filmes em salas de cinema C_{it} , uma variável binária, pelo indivíduo i no ano t . Empregamos um LPM (linear probability model), e um modelo não-linear Probit (WOOLDRIDGE, 2002).

$$C_{it} = \alpha_0 + \alpha_1 I_{it} + \alpha_2 Post + \alpha_3 I_{it} * Post + \alpha' X_{it} + u_{it} \quad (1)$$

sendo $I=1$ se o i -ésimo indivíduo tem acesso à Internet nos anos $t=2002/3$ ou $2008/9$ e 0 caso contrário. $Post$ é a variável binária que assume valor 1 para ano $=2008/9$ e 0 para ano $=2002/3$, e α_3 corresponde ao nosso coeficiente de interesse. A matriz X_{it} inclui as seguintes variáveis que também podem influenciar a decisão de consumir cinema; sexo, idade, dummy de cor da pele, dummy de religião, anos de escolaridade, renda per capita, número de quartos em casa, propriedade de seguro de saúde privado, propriedade de cartão de crédito como proxy para emprego, dummy de estudante, dummy de área urbana, dummies de estado e coorte de ano de nascimento. Todas as variáveis de valor monetário estão em preços de 2002.

Além disso, como o acesso à Internet e sua velocidade (banda larga) aumentaram ao mesmo tempo, também incluímos a propriedade de computadores pessoais (PC) para controlar o maior acesso nos últimos anos e para outras atividades on-line, conforme proposto por Waldfogel (2017).

11.3 Achados Empíricos e Discussão

As estimativas dos modelos DD sugerem um efeito negativo da banda larga sobre o consumo de filmes nos cinemas, que são robustos às diferentes especificações de modelos e à inclusão do controle de propriedade de PCs (Tabela 23).

O erro padrão da robustez do cluster está entre parênteses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$; X variáveis incluídas) na tabela 23 e 24, as variáveis dummies não foram apresentadas:

Tabela 23 - Estimativas de Diferença-Diferença para consumo de filmes nas salas de cinema

Variável dependente : Cinema	LPM	PROBIT	LPM	PROBIT
	Sem controle de pc		Com controle de pc	
Internet	0.0714*** (0.008)	0.2173*** (0.036)	0.0562*** (0.008)	0.1428*** (0.041)
Internet x Post	-	-	-	-0.1062**
	0.0554*** (0.009)	0.1343*** (0.043)	0.0477*** (0.009)	0.0444 (0.044)
Post	-	-0.0687	-	-0.0922**
	0.0082*** (0.002)	0.0098*** (0.044)	0.0098*** (0.002)	0.0444 (0.044)
Posse de computador			0.0330*** (0.003)	0.1880*** (0.022)
Idade em anos	0.0009 (.)	0.0738 (0.155)	0.0009 (.)	0.0778 (0.152)
Sexo feminino	-	-	-	-
	0.0034*** (0.001)	0.0704*** (0.012)	0.0033*** (0.001)	0.0704*** (0.012)
Raça branco	0.0035* (0.002)	0.0680* (0.035)	0.0024 (0.002)	0.0566 (0.035)
Raça pardo	-0.0005 (0.001)	-0.0038 (0.037)	-0.0006 (0.001)	-0.0054 (0.036)
Raça amarela	0.0243*** (0.007)	0.2594*** (0.094)	0.0224*** (0.007)	0.2490*** (0.094)
Anos de estudo	0.0033*** (0.000)	0.0870*** (0.003)	0.0028*** (0.000)	0.0830*** (0.003)
Estuda atualmente	0.0206*** (0.003)	0.2184*** (0.013)	0.0192*** (0.003)	0.2040*** (0.014)
Religião católica	-0.0011	-	-0.0009	-0.0531**
		0.0555*** (0.001)		0.021 (0.021)
Possui plano de saúde	0.0154*** (0.002)	0.0977*** (0.017)	0.0122*** (0.002)	0.0879*** (0.016)
Log da renda média familiar por pessoa deflacionada	0.0160*** (0.002)	0.3523*** (0.013)	0.0135*** (0.002)	0.3282*** (0.013)
Número de dormitórios	0.0021*** (0.000)	0.0412*** (0.007)	0.0006* (0.000)	0.0279*** (0.006)
Posse de cartão de crédito	0.0063*** (0.001)	0.0949*** (0.018)	0.0046*** (0.001)	0.0882*** (0.018)
Reside em zona urbana	0.0033** (0.001)	0.4696*** (0.053)	0.0022* (0.001)	0.4520*** (0.052)
Constante	-0.1709 (15.439)	-12.5224 (13.465)	-0.1536 (.)	-12.6084 (13.184)
Quantidade de observações	292,392	256,341	292,392	256,341
R quadrado	0.069		0.073	

Fonte: Ticianá Sá da Justa, 2018

As estimativas de gastos com filmes nos cinemas mostram a mesma tendência (Tabela 23). Além disso, as estimativas para o consumo de DVD não indicam mudanças devido à adoção da Internet de banda larga, enquanto o consumo de TV a cabo apresenta um efeito negativo. Estes resultados sugerem que os brasileiros tratados não estavam substituindo filmes nos cinemas por DVD ou TV a cabo. As tendências entre o grupo tratado e controle mostram o mesmo padrão (Figura 2). Vale a pena notar que a opção de substituição, um filme visto em um lugar do que em um teatro e não exatamente no momento do lançamento, é para um produto de menor qualidade. Apesar da baixa qualidade, Hui e Png (2002) argumentam que geralmente o grau de substituição tende a ser significativo.

Tabela 24 - Estimativas do OLS-DD para bens substitutos e para filmes em gastos nos cinemas

Variáveis	DVD	TV a cabo	Gasto com cinema
Internet	-7.535 (6.013)	63.87*** (7.859)	408.2*** (85.00)
Post	0.0984*** (0.00200)	-0.00132*** (0.000218)	-0.0228*** (0.00469)
Internet*Post	0.00379 (0.00300)	-0.0318*** (0.00391)	-0.203*** (0.0423)
Observations	344,532	344,532	344,510
R-squared	0.481	0.182	0.050

Fonte: Ticiania Justa, 2018

Figura 30 - Tendências entre grupos tratados e não tratados antes e depois da banda larga

Fonte: Ticiania Justa, 2018

Esses resultados não são capazes de eliminar a pirataria como possível mecanismo importante para a redução do consumo de filmes nos cinemas, como afirmam Bai e Waldfogel (2012) e Danaher e Smith (2014). Isso porque, além de DVD e TV a cabo, poucas opções legais on-line, postadas por Liebowitz e Zentner (2009) - como Netflix, Amazon, Spotify,

iTunes e outros serviços - não estavam disponíveis entre 2002-2009 no Brasil. Esses serviços de streaming ficam disponíveis somente depois de 2010, devido ao atraso da chegada de novas tecnologias em um país que não é um desenvolvedor de tecnologia. Além disso, há fortes evidências de pirataria de filmes no Brasil, de acordo com o IPEA (2012), cerca de 41% dos usuários brasileiros de Internet cometeram pirataria em 2010 e 2011, baixando arquivos de música e filmes.

11.4 Conclusão

Estimamos o efeito do acesso à Internet no consumo de filmes nos cinemas brasileiros entre 2002 e 2009, quando os serviços de streaming de filmes ainda não estavam disponíveis. Nossos resultados indicam uma redução do consumo de filmes nos cinemas e em gastos relacionados devido à Internet. Além disso, nem o consumo de DVD nem o de TV a cabo, os principais produtos substitutos legais, aumentam. Contribuindo para o controle de atividades on-line alternativas, esses resultados sugerem uma substituição do consumo legal de filmes pelo consumo ilegal em um local onde apenas os mais ricos e instruídos consomem filmes nos cinemas e a maior parte dos filmes é produzida no exterior.

A combinação das técnicas utilizadas de florestas aleatórias com análise da dependência parcial demonstrou-se inicialmente confiável para a extração de conhecimento sobre os dados, através da comparação com medidas estatísticas básicas para fins de validação inicial.

12 Discussão e trabalhos futuros

Este trabalho de pesquisa demonstrou um método de seleção de variáveis de interesse e como as mesmas se comportam utilizando as florestas aleatórias, focando especificamente na importância das variáveis preditoras e sua influência no resultado final da classificação dos consumidores de cinema. Também utilizou o método de regressão de diferenças em diferenças para demonstrar o efeito pontual de uma intervenção analisando grupos de indivíduos antes e depois da mesma.

Utilizadas em conjunto, estas duas ferramentas constituem um método a escolha de variáveis de atuação e a verificação dos resultados obtidos após a atuação.

12.1 Seleção de variáveis de atuação

Para que se possa inicialmente avaliar sobre quais variáveis preditoras de um determinado conjunto se deve atuar para se obter um determinado resultado deve ser efetuada inicialmente a investigação do impacto das variáveis para o resultado final. A combinação de ranking de importância com a análise da margem de contribuição fornece as variáveis mais importantes e como o comportamento delas influencia no resultado final, norteando assim a decisão sobre as políticas de intervenção que possivelmente trarão melhores resultados.

Porém a seleção das variáveis preditoras para fins de atuação deve ser tomada de forma cautelosa. Verificamos que a importância das variáveis preditoras no decorrer do tempo para o mesmo resultado pode variar devido a fatores externos que não podem ser controlados, como exemplificado para as florestas das POFs 2002 e 2008. Caso haja a disponibilidade de dados temporais o ideal é que sejam verificadas as importâncias das variáveis preditoras para períodos determinados para fins de se verificar a estabilidade das mesmas com relação à sua influência na predição, dentre as variáveis mais estáveis pode-se escolher qual será objeto da intervenção e desta forma possivelmente diminuir os riscos de falha da intervenção a ser realizada.

Temos como exemplo a variável preditora idade, que se manteve estável na quinta posição para as duas POFs e foi objeto de intervenção através a aplicação do Estatuto do Idoso (Lei Federal no. 10741/2003) o qual garantiu a meia - entrada às pessoas com 60 anos

ou mais. Verificamos pela análise realizada que esta intervenção combinada a outros fatores externos como o aumento da expectativa de vida, por exemplo, teve um efeito positivo e houve aumento no consumo de cinema entre os integrantes da terceira idade.

12.2 Verificação dos resultados obtidos

Para um estudo específico sobre os efeitos do estatuto poderia ser aplicado o método de diferenças em diferenças, incluindo as variáveis de controle necessárias e verificaríamos então se o aumento do consumo se deveu exclusivamente à lei e em qual grau.

12.3 Importância do trabalho para a área de inteligência artificial e trabalhos futuros

Através da aplicação das florestas aleatórias aos dados das duas POFs verificamos que os modelos de predição são sensíveis às variações de importância das variáveis preditoras, e que essa sensibilidade não pode ser caracterizada como sobreajustamento, uma vez que as florestas aleatórias são imunes a ele conforme seus criadores Breiman e Cutler (2001).

Faz-se necessária então a investigação de possíveis soluções para a criação de algoritmos de predição menos suscetíveis a estas mudanças, abrindo assim um campo de interesse de pesquisa na área de inteligência artificial.

Referências⁷

ALCOCK, S.; NELSON, R. **Measuring the impact of the copyright amendment act on New Zealand residential dsl users.** In: Proceedings of the 2012 Internet Measurement Conference. New York, NY, USA: ACM, 2012. (IMC '12), p. 551–558. ISBN 978-1-4503-1705-4. Disponível em: <http://doi.acm.org/10.1145/2398776.2398833>.

AMARAL, E.F.; **Notas das aulas do curso Técnicas Econométricas para Avaliação de Impacto.** International Policy Centre for Inclusive Growth, Brasília, Brasil. 2013. Disponível em: <http://www.ipc-undp.org/evaluation/>

ASIKUZZAMAN, M. et al. **A blind and robust video watermarking scheme using chrominance embedding.** In: IEEE. Digital Image Computing: Techniques and Applications (DICTA), 2014 International Conference on. [S.l.], 2014. p. 1–6.

BAE, J. K.; KIM, J. **Integration of heterogeneous models to predict consumer behavior.** 2010,[S.I.] , Expert Systems with Applications, v. 37, p. 1821 – 1826. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417409006861>

BAI, J.; WALDFOGEL, J., **Movie piracy and sales displacement in two samples of Chinese consumers.** Information Economics and Policy, China, 2012, v. 24, n. 3, p 187 – 196. Disponível em: <https://EconPapers.repec.org/RePEc:eee:iepoli:v:24:y:2012:i:3:p:187-196>

BAUDRY, S.; CHUPEAU, B.; LEFÈVRE, F. **A framework for video forensics based on local and temporal fingerprints.** In: 2009 16th IEEE International Conference on Image Processing (ICIP). [S.l.: s.n.], 2009. p. 2889–2892. ISSN 1522-4880.

BECKER, M.; DESOKY, A. **A study of the dvd content scrambling system (css)algorithm.** In: IEEE. Signal Processing and Information Technology, 2004. Proceedings of the Fourth IEEE International Symposium on. [S.I.], 2004. p. 353–356.

BISHOP, J, **Who are the Pirates? The Politics of Piracy, Poverty, and Greed in a Globalized Music Market.** Popular Music and Society, Routledge, 2004,[S.I.], v.27, n. 1, p. 101-106. Disponível em: < <https://doi.org/10.1080/0300776042000166639>>

BISHOP, J; CAND. **Who Are the Pirates? The Politics of Piracy, Poverty, and Greed in a Globalized Music Market. Popular Music and Society.** [S.I.], 2004, v.27 p.101–6.

BLOOM, J. A.; POLYZOIS, C. **Watermarking to track motion picture theft.** In:Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004. [S.I.: s.n.], 2004. v. 1, p. 363–367.

⁷ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BOURDON, P. et al. **A metamerism-based method to prevent camcorder movie piracy in digital theaters.** In: 2010 IEEE International Conference on Multimedia and Expo. [S.l.:s.n.], 2010. p. 468–473. ISSN 1945-7871.

BRASIL. LEI Nº 10.741, DE 1º DE OUTUBRO DE 2003. **Dispõe sobre o Estatuto do Idoso e dá outras providências.** , Brasília,DF, junho 2003. Disponível em: http://www.planalto.gov.br/ccivil_03/LEIS/2003/L10.741.htm

BRASIL. DECRETO Nº 4.769, DE 27 DE JUNHO DE 2003. **Plano Geral de Metas para a Universalização do Serviço Telefônico Fixo Comutado** , Brasília,DF, junho 2003. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/2003/D4769.htm

BREIMAN, L. **Random forests.** *Machine Learning*, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <<https://doi.org/10.1023/A:1010933404324>>.

BRUCE, R. **Semi-supervised learning using prior probabilities and em.** In IJCAI Workshop on Text Learning: Beyond Supervision, August 2001,[S.I.] 23.

BRUYÈRE, S.; OECHSEL, V. **Creation of a data observatory enables the uncovering of consumer behavior by client behavioral study, through the use of 1st generation loyalty and stimulation platforms: Conception of an observatory of consumer data, the "Promo Place" project.** 2015, [S.I.], 6th International Conference on Information Systems and Economic Intelligence (SIIE), p. 18 - 24.

BYERS, S. et al. **Exposing digital content piracy: approaches, issues and experiences.** In:Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004. [S.l.: s.n.], 2004. v. 1, p. 377–381.

CAPLAN B.,COWEN T. 2004. **Do We Underestimate the Benefits of Cultural Competition?** *American Economic Review*, [S.I.] , 94 v. 2 p.402-407.

CAPLAN, B.; COWEN, T. **Do we underestimate the benefits of cultural competition?**The *American Economic Review*, American Economic Association,[S.I.], vol. 94, n. 2, p. 402–407, 2004. Disponível em: <<http://www.jstor.org/stable/3592918>>.

CAVARARO, R. **Pesquisa de Orçamentos Familiares 2002 - 2003 - Perfil das despesas no Brasil Indicadores selecionados.** [S.I.]: Instituto Brasileiro de Geografia e Estatística -IBGE, 2007.

CHAWLA, N. V. et al. **SMOTE: Synthetic Minority Over-sampling Technique.** 2002, [S.I.], *Journal of Artificial Intelligence Research*, v. 16, p. 321–35.

CHELLAPPA, R. K.; SHIVENDU, S. **Economics of technology standards: Implications for offline movie piracy in a global context.** In: IEEE. *System Sciences*, 2003. Proceedings of the 36th Annual Hawaii International Conference on. [S.l.], 2003. p. 10–pp.

CHUPEAU, B.; BAUDRY, S.; DOERR, G. **Forensic characterization of pirated movies: digital cinema cam vs. optical disc rip.** In: 2014 IEEE International Workshop on Information Forensics and Security (WIFS). [S.l.: s.n.], 2014. p. 155–160. ISSN 2157-4766.

CIOS K. J. et al. A., **Data Mining: A Knowledge Discovery Approach**. 2007, Springer-Verlag, Berlin, Heidelberg.

CORIA, L. E. et al. **A video watermarking scheme based on the dual-tree complex wavelet transform**. IEEE Transactions on Information Forensics and Security , IEEE, v. 3, n. 3, p.466–474, 2008.

CRIMINISI, A; SHOTTON, J. **Decision Forests for Computer Vision and Medical Image Analysis**. Springer, [S.I], 2013.

DANAHER, B.; SMITH M. **Gone in 60 Seconds: The Impact of the Megaupload Shutdown on Movie Sales**. 2013, SSRN Electronic Journal. Disponível em: <<https://doi.org/10.2139/ssrn.2229349>> Acesso em 01 de julho de 2018.

DANAHER;B., SMITH; M. D.; TELANG; R;CHEN; S. **The Effect of Graduated Response Anti-Piracy Laws on Music Sales: Evidence from an Event Study in France**. The Journal of Industrial Economics, 2014, [S.I], v. 62, p. 541–53.

DEATON, A. **Panel Data from Time Series of Cross-Sections**. Journal of Econometrics 1985, [S.I] , v. 30, p. 109–26.

DIETTERICH T., **Ensemble methods in machine learning**, InFirst International Workshop on Multiple Classifier Systems, J. Kittler, and F. Roll (eds.), Lecture Notes in Computer Science, p. 1–15, Springer-Verlag, 2000, [S.I]

DUBEY, N. K.; KUMAR, S. **A review of watermarking application in digital cinema for piracy deterrence**. In: IEEE. Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on. [S.I.], 2014. p. 626–630.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. , **From data mining to knowledge discovery in databases**. AI Magazine,[SI], v. 17, p. 37–54, 1996.

FÁVERO, et. al. **Análise de dados : modelagem multivariada para a tomada de decisões**. Rio de Janeiro, Elsevier, 2009.

FUJIKAWA, M. et al. **Development of countermeasure systems for content leaks by video recording/camera shooting**. In: International Conference on Information Society (i-Society, [S.I.: s.n.], 2012. p. 76–81.

GAO, J. **Self-clustering analysis of Shanghai's new pragmatic consumers - A new way of CRM's application**. 2011, [S.I].

GAO, Z. et al. **Demo: Dlp based anti-piracy display system**. In: 2014 IEEE Visual Communications and Image Processing Conference. [S.I.: s.n.], 2014. p. 325–325.

HADIDA, A. **Commercial success and artistic recognition of motion picture projects**, Journal of Cultural Economics,2010, [S.I] 34, v. 1, p. 45-80.

HAJJ-AHMAD, A. et al. **Flicker forensics for camcorder piracy**. IEEE Transactions on Information Forensics and Security, v. 12, n. 1, p. 89–100, Jan 2017. ISSN 1556-6013.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 01238147909780123814791.

HE, H.; GARCIA, E. A. **Learning from imbalanced data**. IEEE Transactions on knowledge and data engineering, IEEE, v. 21, n. 9, p. 1263–1284, 2009.

HOTHORN T., et al. **Unbiased Recursive Partitioning: A Conditional Inference Framework**. [S.I.] , 2006, *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

HUSAK, W. **Economic and Other Considerations for Digital Cinema**. Signal Processing: Image Communication, 2004, [S.I], v. 19, p. 921–36.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa de Orçamentos Familiares 2008-2009. Perfil das despesas no Brasil. Indicadores selecionados**. [S.l.]: IBGE, 2012.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA . **Download de músicas e filmes no Brasil: Um perfil dos piratas online**. Comunicados do IPEA, 2012, Rio de Janeiro, n.147.

KAEHLING, L. P.; MICHAEL, L.; MOORE, A. P. **Reinforcement Learning: A Survey**, In Journal of Artificial Intelligence Research, [S.I], v. 4, 1996.

KAI-LUNG, H.; PNG, I. **Piracy and the Legitimate Demand for Recorded Music**. Contributions in Economic Analysis & Policy, 2003, [S.I], v. 2.

KAI-LUNG, H; PNG, I.P.L. **On the Supply of Creative Work: Evidence from the Movies**. The American Economic Review, 2002 ,[S.I], v. 92 p. 217–20.

KAO, P. et al, **Consumer-centered product design specifications using fuzzy multivariate regression analysis**. 1993, [S.I], (2nd) International Symposium on Uncertainty Modeling and Analysis, p. 387 – 392, doi: 10.1109/ISUMA.1993.366740.

KIROVSKI, D.; PETITCOLAS, F. A. **Blind pattern matching attack on watermarking systems**. IEEE Transactions on Signal Processing, IEEE, v. 51, n. 4, p. 1045–1053, 2003.

KMENTA, Jan. **Elementos de Econometria**, 1978, São Paulo, Editora Atlas S. A.

KWONG, T. C. H.; LEE, M. K. O. **Understanding the behavioral intention to digital piracy in virtual communities - a propose model**. In: *e-Technology, e-Commerce and e-Service*, 2004. IEEE '04. 2004 IEEE International Conference on . [S.l.: s.n.], 2004. p.

LANDAU, S.; STRATULATE, R.; TWILLEAGER, D. **Consumers, fans, and control: What the games industry can teach hollywood about drm**. In: *Proceedings of the ACM*

Workshop on Digital Rights Management. New York, NY, USA: ACM, 2006. (DRM '06), p. 1–8. ISBN 1-59593-555-X. Disponível em: <<http://doi.acm.org/10.1145/1179509.1179511>>.

LARABI, M. C.; ROSSELLI, V.; FERNANDEZ-MALOIGNE, C. **Annoyance maximization for digital cinema anti-piracy applications.** In: 2009 Fifth International Conference on Signal Image Technology and Internet Based Systems. [S.l.: s.n.], 2009. p.232–236.

LI, N. J.; CHOI, K. H.; SHIN, D. R. **Rfid-based digital content copy protection system in movie and audio rental agency.** In: IEEE. Advanced Communication Technology (ICACT), 2013 15th International Conference on. [S.l.], 2013. p. 169–172.

LI, S.; XING, Q.; JIN, S. **The Influence Factors of Subscriber Locked in the Era of 4G - A Consumer Behavior Perspective.** 2014, [S.I.], 2014 Seventh International Joint Conference on Computational Sciences and Optimization, p. 636 – 640, doi: 10.1109/CSO.2014.120.

LI, Z. et al. **On Designing Optimal Data Purchasing Strategies for Online Ad Auctions.** 2018, Estocolmo, Suécia, Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Disponível em: <<http://dl.acm.org/citation.cfm?id=3237927>>

LIEBOWITZ, S. J. **File Sharing: Creative Destruction or Just Plain Destruction?** The Journal of Law and Economics, 2006 ,[S.I.], v. 49, p. 1–28.

LIEBOWITZ, S. J.; ZENTNER A. **Clash of the Titans: Does Internet Use Reduce Television Viewing?** The Review of Economics and Statistics. 2012 ,[S.I.], v. 94, p. 234–45.

LIN, E. I. et al. **Advances in digital video content protection.** Proceedings of the IEEE, v. 93, n. 1, p. 171–183, Jan 2005. ISSN 0018-9219.

MACEDO, N.Q. et al, **Typification of load curves for DSM in Brazil for a smart grid environment.** International Journal of Electrical Power & Energy Systems, 2015, [S.I.], v.67, p. 216 – 221. Disponível em <http://www.sciencedirect.com/science/article/pii/S0142061514007200>

MAFFIOLI, Alessandro. 2011. **Non-Experimental Methods:Propensity Score Matching and Difference in Difference.** IDB: Cuernavaca. Disponível em <http://www.impactevaluation2011.org/forum/wpcontent/uploads/2011/06/C5-commissioners-of-evaluationsfile-1.pdf>

MAZĂRE, A. et al. **Intelligent system for determining the consumer profile and generate alarm in case of significant deviations from the profile.** 2017, New York, NY, USA, ICEC '08: Proceedings of the 10th International Conference on Electronic Commerce, p. 361 – 366.

MOREIRA-PÉREZ, J. J. et al. **Exploring color information to characterize camcorder piracy.** In: 2013 IEEE International Workshop on Information Forensics and Security (WIFS). [S.l.: s.n.], 2013. p. 132–137. ISSN 2157-4766.

NAGI, J. et al. **Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines**. 2010, [S.I], IEEE Transactions on Power Delivery, v.25, n.2, p. 1162 – 1171, doi: 10.1109/TPWRD.2009.2030890.

NAKASHIMA, Y.; TACHIBANA, R.; BABAGUCHI, N. **Watermarked movie soundtrack finds the position of the camcorder in a theater**. IEEE Transactions on Multimedia, IEEE, v. 11, n. 3, p. 443–454, 2009.

NIEUWENHOFF, Nathalia. **Uma comparação da aplicação de métodos computacionais de classificação de dados aplicados ao consumo de cinema no Brasil**. 2017, São Paulo, Dissertação (Mestrado em Sistemas de Informação) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2017. doi:10.11606/D.100.2017.tde-01062017-085136. Acesso em: 2018-09-24.

OBERHOLZER-GEE, F.;KOLEMAN, S. **The Effect of File Sharing on Record Sales: An Empirical Analysis**. The Journal of Political Economy, 2007,[S.I], v. 115, p. 1–42.

OZYIRMIDOKUZ, E. K. et al, **A Data Mining Based Approach to a Firm's Marketing Channel**. Procedia Economics and Finance, 2015, v. 27, p. 77 – 84, [S.I]. Disponível em:<http://www.sciencedirect.com/science/article/pii/S2212567115009752>

PETITCOLAS, F. A.; KIROVSKI, D. **The blind pattern matching attack on watermark systems**. In: IEEE. Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. [S.l.], 2002. v. 4, p. IV–3740.

Piracy on the Silver Screen. The Journal of Industrial Economics, 2007, [S.I], v. 55, p. 379–95.

QUINLAN, J. R. **Induction of decision trees**. Machine learning, Springer, v. 1, n. 1, p.81–106, 1986.

REDELBACH, M. ET AL. **Modelling customer choice and market development for future automotive powertrain technologies**. 2013, [SI], 2013 World Electric Vehicle Symposium and Exhibition (EVS27), p. 1-10.

ROOPALAKSHMI, R.; REDDY, G. R. M. **Robust features for accurate spatio-temporal registration of video copies**. In: IEEE. Signal Processing and Communications (SPCOM), 2012 International Conference on. [S.l.], 2012. p. 1–5.

RUDMAN, K. et al. **Toward real-time detection of forensic watermarks to combat piracy by live streaming**. SMPTE Motion Imaging Journal, SMPTE, v. 125, n. 1, p. 34–41, 2016.

RUSSEL, J. S.; NORVIG, P. 2003.[S.I] **Artificial Intelligence: A Modern Approach** (2 ed.). Pearson Education.

SHAPIRO, C.; VARIAN, H. R. **Information Rules: A Strategic Guide to the Network Economy**. Boston, MA, USA: Harvard Business School Press, 2000. ISBN 1578515300.

SHAPIRO, C.; VARIAN, H. R., **Information Rules: A Strategic Guide to the Network Economy**. Boston, MA, USA: Harvard Business School Press, 2000.

SILVA L. A.; PERES S. M. ; BOSCARIOLI, C. B. **Introdução à mineração de dados: Com Aplicações em R.** [S.I.] , Elsevier Brasil, 2017.

SILVER, J.; **Hollywood's Dominance of the Movie Industry: How Did it Arise and how Has it Been Maintained?** Queensland University of Technology, Brisbane, 2007. Disponível em: <<https://books.google.com.br/books?id=UC4RtwAACAAJ>>.

STROBL, C. et al. **Bias in random forest variable importance measures: Illustrations, sources and a solution.** BMC Bioinformatics, v. 8, n. 1, p. 25, Jan 2007. ISSN 1471-2105. Disponível em: <<https://doi.org/10.1186/1471-2105-8-25>>.

THAKUR, P. S.; KUMAR, S. **Technique for estimation of the position of the pirate in-theater piracy.** In: 2015 Second International Conference on Advances in Computing and Communication Engineering. [S.l.: s.n.], 2015. p. 515–518.

VANY, D.;ARTHUR, S. W.; WALLS, D. **Estimating the Effects of Movie Piracy on Box-Office Revenue.** Review of Industrial Organization, 2007, [S.I], v. 30, p. 291–301.

VARIAN, H. R. **Copying and copyright.** The Journal of Economic Perspectives, American Economic Association, v. 19, n. 2, p. 121–138, 2005. ISSN 08953309. Disponível em: <<http://www.jstor.org/stable/4134940>>.

WALDFOGEL, J. **How Digitization Has Created a Golden Age of Music, Movies, Books, and Television.** Journal of Economic Perspectives, 2017 ,[S.I], v. 31, p. 195-214.

WALDFOGEL, J. **Music File Sharing and Sales Displacement in the iTunes Era.** Information Economics and Policy, 2010 ,[S.I], v. 22, p. 306–14.

WANG, M. S.; LEE, R. B. **Architecture for a non-copyable disk (ncdisk) using a secret-protection (sp) soc solution.** In: IEEE. Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on. [S.l.], 2007. p.1999–2003.

WANG, Y.; LV, R.; HONG, Z. **A p2p application demonstration system for resilient overlay networks with intelligent nodes supporting drm.** In: IEEE. Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on . [S.l.], 2009. v. 2, p.336–339.

WHITE, K.-A.; RAO, B.; MINAKAKIS, L. **A qualitative examination of movie piracy behaviors and their impact.** In: IEEE. Technology Management for Global Economic Growth (PICMET),Proceedings of PICMET'10:.. [S.l.], 2010.

YU, S.; YANG, D. **The Role of Big Data Analysis in New Product Development.** 2016, [SI], 2016 International Conference on Network and Information Systems for Computers (ICNISC), p. 279 -283.

ZENTNER, A. **Measuring the Effect of File Sharing on Music Purchases.** The Journal of Law & Economics, 2006 ,[S.I], v. 49, p. 63-90.

ZHAI, G.; WU, X. **Defeating camcorder piracy by temporal psychovisual modulation.** Journal of display technology, IEEE, v. 10, n. 9, p. 754–757, 2014.

Apêndice A - Revisão da Literatura

Considerando o caráter interdisciplinar desta pesquisa, investigação de um fenômeno social e econômico usando ferramentas computacionais em dados de indivíduos, a revisão sistemática da literatura para a pesquisa foi elaborada em três etapas distintas, cada uma delas com objetivos e método de pesquisa e extração de informações conduzidas de maneira independente. A primeira focada no contexto social e econômico relacionados à indústria de cinema e suas influências da evolução tecnológica e de compartilhamento de filmes.

A segunda buscou levantar trabalhos científicos que relatassem a utilização de aprendizagem de máquina para fins de análise de perfil de clientes afim de compreender as principais metodologias e tecnologias envolvidas no processo, comparativos existentes entre análises realizadas utilizando técnicas estatísticas e de aprendizagem de máquina e uma visão abrangente das contribuições da área de aprendizagem de máquina para a área de segmentação e descoberta de perfis de clientes.

Protocolo para revisão sistemática da primeira etapa

Objetivos:

Identificar e analisar os impactos da pirataria de filmes na cadeia produtiva de cinema e como este problema vem sendo combatido por meios tecnológicos e legais.

Questões de pesquisa:

1. Qual o impacto da pirataria de filmes na indústria cinematográfica?
2. Como a tecnologia pode ser empregada para evitar a pirataria de filmes?
3. Quais as principais abordagens de enfrentamento da pirataria de filmes utilizadas?

Intervenção: métodos e técnicas de avaliação

Controle: artigos de surveys obtidos com a orientadora, teses e dissertações sobre a problemática obtidas na Internet.

População: estudos relativos aos impactos da pirataria de filmes no mercado cinematográfico e relacionados a avanços científicos, legais e tecnológicos ocasionados pelo seu enfrentamento.

Resultados: visão abrangente do estado atual do problema, como ele vem sendo enfrentado e por consequência da relevância do projeto de mestrado.

Seleção de fontes:

As fontes deverão estar disponíveis via web, preferencialmente em bases de dados científicas das áreas de economia e computação.

Palavras-chaves:

“broadband, film, movie, piracy;”

Fontes consultadas:

- Biblioteca Digital do IEEE (<http://ieeexplore.ieee.org/Xplore/>)
- Biblioteca Digital da ACM (<http://portal.acm.org/>)
- Web of Science (<https://login.webofknowledge.com>)
- JSTOR (<https://www.jstor.org/>)
- NBER (<https://www.nber.org/>)

Idioma(s) dos artigos:

Inglês.

Critérios de inclusão e exclusão dos trabalhos:

Critérios de inclusão:

1. Serão incluídos trabalhos publicados e disponíveis integralmente em bases de dados científicas ou em versões impressas.
2. Artigos que tratem de métodos de combate à pirataria de filmes.
3. Artigos que abordem os efeitos econômicos da pirataria de filmes.
4. Artigos que tratem da pirataria de filmes pela internet.

Critérios de exclusão:

1. Artigos anteriores a 2000.
2. Artigos em língua diversa do Inglês.

Estratégia de extração da informação:

Para cada artigo aceito foi preenchida uma ficha com as seguintes informações:

1. Abordagem empregada para o combate à pirataria.
2. Conclusões sobre os impactos da pirataria de filmes.

3. Relevância dada à internet no processo de pirataria.

4. Dados bibliográficos básicos.

Strings de busca utilizadas no formato exigido por cada motor:

Observação: As strings de busca variaram de acordo com os resultados iniciais para a string contendo todas as palavras-chave.

ACM : +movie +piracy +broadband

IEE: (((movie*) OR film*) AND piracy)

Web of Science: TS =(((movie* OR film*) AND piracy) AND broadband)

JSTOR: ((movie AND piracy) AND (broadband))

NBER: film movie piracy

Resultados:

A busca primária nas bases consultadas retornou 107 artigos, dos quais 44 foram selecionados para leitura de acordo com os critérios de inclusão e exclusão explicitados.

Tabela 24 – Relação dos trabalhos selecionados para a revisão sistemática

Ref.	Título	Ano
A1	<i>The blind pattern matching attack on watermark systems</i>	2002
A2	<i>Economics of technology standards: implications for offline movie piracy in a global context</i>	2003
A3	<i>Understanding the behavioral intention to digital piracy in virtual communities - a propose model</i>	2004
A4	<i>Watermarking to track motion picture theft</i>	2004
A5	<i>Randomizing the replacement attack</i>	2004
A6	<i>Exposing digital content piracy: approaches, issues and experiences</i>	2004
A7	<i>Managing Piracy: Pricing and Sampling Strategies for Digital Experience Goods in Vertically Segmented Markets</i>	2005
A8	<i>A Comment on Chinese Legal Environment of Online Copyright Protection</i>	2005
A9	<i>Advances in Digital Video Content Protection</i>	2005
A10	<i>Price Discrimination, Copyright Law, and Technological Innovation: Evidence from the Introduction of DVDs</i>	2005
A11	<i>Fear and Norms and Rock & Roll: What Jambands Can Teach Us About Persuading People to Obey Copyright Law</i>	2006
A12	<i>Consumers, Fans, and Control: What the Games Industry Can Teach Hollywood About DRM</i>	2006
A13	<i>Enforcement and Control of Piracy, Copying, and Sharing in the Movie Industry</i>	2007
A14	<i>Piracy on the Silver Screen</i>	2007
A15	<i>An Analysis of the Impact of Economic Wealth and National Culture on the Rise and Fall of Software Piracy Rates</i>	2008
A16	<i>Factors That Influence the Intention to Pirate Software and Media</i>	2008

(continua)

(conclusão)

Ref.	Título	Ano
A17	<i>Competing with Free: The Impact of Movie Broadcasts on DVD Sales and Internet Piracy</i>	2009
A18	<i>Annoyance Maximization for Digital Cinema Anti-piracy Applications</i>	2009
A19	<i>A framework for video forensics based on local and temporal fingerprints</i>	2009
A20	<i>A P2P Application Demonstration System for Resilient Overlay Networks with Intelligent Nodes Supporting DRM</i>	2009
A21	<i>Competing with free: the impact of movie broadcasts on DVD sales and internet piracy I</i>	2009
A22	<i>A qualitative examination of movie piracy behaviors and their impact</i>	2010
A23	<i>A metamerism-based method to prevent camcorder movie piracy in digital theaters</i>	2010
A24	<i>File sharing and copyright</i>	2010
A25	<i>Piracy or promotion? The impact of broadband Internet penetration on DVD sales</i>	2010
A26	<i>An Empirical Analysis of the Impact of Pre-Release Movie Piracy on Box-Office Revenue</i>	2011
A27	<i>Determinants of Digital Piracy: A Re-examination of Results</i>	2012
A28	<i>Measuring the Impact of the Copyright Amendment Act on New Zealand Residential DSL Users</i>	2012
A29	<i>Development of countermeasure systems for content leaks by video recording/camera shooting</i>	2012
A30	<i>News Briefs</i>	2013
A31	<i>Exploring color information to characterize camcorder piracy</i>	2013
A32	<i>Impact of heterogeneity and socioeconomic factors on individual behavior in decentralized sharing ecosystems</i>	2014
A33	<i>A Review of Watermarking Application in Digital Cinema for Piracy Deterrence</i>	2014
A34	<i>A Blind and Robust Video Watermarking Scheme Using Chrominance Embedding</i>	2014
A35	<i>Defeating Camcorder Piracy by Temporal Psychovisual Modulation</i>	2014
A36	<i>Demo: DLP based anti-piracy display system</i>	2014
A37	<i>Forensic characterization of pirated movies: digital cinema cam vs. optical disc rip</i>	2014
A38	<i>An empirical analysis of the impact of pre-release movie piracy on box office revenue</i>	2014
A39	<i>Technique for Estimation of the Position of the Pirate In-Theater Piracy</i>	2015
A40	<i>A blind and robust video watermarking scheme in the DT CWT and SVD domain</i>	2015
A41	<i>Toward Real-Time Detection of Forensic Watermarks to Combat Piracy by Live Streaming</i>	2016
A42	<i>The ecosystem for accessing tv series and films in Spain: an outline of the situation following the intellectual property act 2015</i>	2016
A43	<i>Flicker Forensics for Camcorder Piracy</i>	2017

Fonte: Ticiania Justa, 2018

Em seguida os artigos foram classificados de acordo com a principal abordagem de enfrentamento da pirataria.

Tabela 25 – Classificação dos trabalhos selecionados quanto à sua abordagem

Ref.	Tecnológica	Legal	Comportamental	Econômica
A1	X			
A2				X
A3			X	
A4	X			
A5	X			
A6	X			
A7		X		
A8		X		
A9	X			
A10	X			
A11	X	X	X	X
A12		X	X	
A13		X	X	
A14		X	X	
A15			X	X
A16				X
A17			X	
A18				X
A19	X			
A20	X			
A21	X			
A22		X		
A23			X	X
A24	X			
A25		X	X	
A26				X
A27				X
A28			X	
A29		X	X	
A30				
A31		X		
A32	X			
A33			X	
A34	X			
A35	X			
A36	X			
A37	X			
A38	X			
A39	X			
A40	X			
A41	X			
A42		X	X	
A43	X			

Fonte: Ticiane Sá da Justa, 2018

Tabela 26 – Sumarização das classificações

Tecnológica	Legal	Comportamental	Análise de impacto econômico
21 48%	10 25%	14 32%	9 20%

Fonte: Ticianá Sá da Justa, 2018

Figura 31 – Produção de artigos por ano de publicação

Fonte: Ticianá Sá da Justa, 2018

A análise dos resultados apresentados mostra que a abordagem tecnológica tem sido a medida mais popular de enfrentamento do problema da pirataria de filmes uma vez que pode impossibilitá-la com um custo menor de recursos legais, pois está focada na prevenção e pouco desgaste da imagem dos estúdios.

O endurecimento legal embora tenha sido tentado em alguns países como China e Nova Zelândia mesmo com a existência das leis estas se tornaram ineficazes pois os produtores relutam em exercer as medidas penalizatórias previstas. A seguir apresentamos uma discussão mais detalhada sobre o cenário revelado pela revisão.

Sob a hipótese de que o consumo de bens de informação no mercado legal possa ser afetado por ações de pirataria, portanto mercado ilegal, e por propaganda e referências boca a boca, a literatura implicada envolve questões aos direitos de propriedade, tais como as questões discutidas em Varian (2005). Questões estas vistas do ponto de vista da dificuldade de garantir direito de propriedade aos bens de informação a partir de suas características de bem público.

Com base em 38 países, entre os anos 1990–2000, Hui and Png (2002) investigam o impacto de incentivos econômicos sobre a oferta internacional de filmes em cinema e em particular o impacto da mudança de lei de propriedade nos Estados Unidos. Para isso, usam variáveis de posse de TV e vídeo, renda pessoal e população em um modelo de demanda e de oferta nacional de cinema. Seus resultados sugerem ‘fortemente’ que a oferta de trabalhos criativos responde a incentivos econômicos, mas não encontram efeito da lei de propriedade sobre a produção de cinema. Também encontram evidências de que o efeito da posse de TV afeta negativamente a oferta de cinema.

Walls (2005) usa dados de 2004 para de 26 países para relacionar pirataria com variáveis sócio-econômicas. Usa um índice de coletivismo, custo de impor direito de propriedade como percentual do PIB, renda per capita e o número de usuários de Internet por 1000 pessoas. Seus resultados sugerem que a pirataria cresce com o coletivismo e reduz com a renda, mas quando se inclui o custo de impor a lei de propriedade e o uso de Internet a renda como variável explicativa deixa de ser estatisticamente significativa.

Vany e Walls (2007) desenvolveram e estimaram um modelo estatístico do efeito da pirataria sobre o desempenho da bilheteria de lançamentos de filmes amplamente divulgados. Seu modelo mostra que a pirataria na Internet diminui a receita da bilheteria destes filmes amplamente divulgados. De acordo com os autores:

“The model overcomes a major weakness of counterfactual or “but for piracy” methods widely used to estimate damages. These counterfactual methods violate the “nobody knows” principle because they forecast what the movie would have earned in the absence piracy. The model we present does not violate this basic principle of motion picture uncertainty. We estimate that pre-release and contemporaneous Internet downloads of a major studio movie accelerated its box-office revenue decline and caused the picture to lose about \$40 million in revenue.” (VANY e WALLS, 2007, pg. 291)

Alguns estudos se dedicam a estimar a demanda por cinema, todavia sem o controle de downloads e propagandas. Entre eles pode-se destacar Dewenter and Westermann (2005), que estimam a demanda por cinema da Alemanha, usando dados de séries de tempo entre 1950 e 2002. Por meio de técnicas como OLS, 2SLS e SUR identificam a relação entre número de assentos, preços médios reais e a demanda por filmes per capita. Também testam a relação de longo prazo entre demanda, preços, renda real e elasticidade da demanda estimada. Encontram elasticidade-preço da demanda entre -2.40 e -2.76 e elasticidade-renda da demanda menor que um em módulo.

MacKenzie (2010) faz um levantamento exaustivo sobre estudos econômicos de cinema e entre as suas classificações por assuntos inclui uma seção sobre pirataria e direitos de propriedade. Nesse levantamento descreve os trabalhos Hui and Png (2002), Vany e Walls (2007) e Walls (2005a, b). Seu levantamento sugere que a literatura sobre o tema ainda é pequena e que contribuições podem abrir um flanco para pesquisa acadêmica na área de economia, num setor muito dinâmico.

Existem vários artigos que exploram o efeito do compartilhamento de músicas pela Internet e o mercado legal de músicas. Autores como Bishop (2004) associam a pirataria de bens de informação, com base em compartilhamento de músicas via Napster pela Internet. Também Oberholzer-Gee e Strumpf (2007) estudam o efeito de compartilhamento de músicas pela Internet sobre o mercado Norte Americano. Usando dados de um grande número de álbuns de músicas compartilhados extensivamente na Internet, testam se estes compartilhamento afetam o mercado legal de músicas. Os autores usam como instrumento para os downloads o número de feriados escolares, refletindo o fato do compartilhamento desta nova tecnologia acontece principalmente entre jovens. Seus resultados apontam para efeito nulo da pirataria sobre as vendas de músicas no mercado legal.

Zentner (2006) estima o efeito de downloads de música no mercado legal usando dados cross-section de indivíduos europeus referente ao ano de 2001. De acordo com o autor, existe uma correlação negativa entre vendas de músicas e aumento de compartilhamento grátis de músicas. O autor usa o acesso a banda larga como instrumento para identificar o efeito de downloads sobre a probabilidade de comprar música. Seus resultados apontam para uma redução na probabilidade de venda de 30% decorrente do efeito boca-a-boca (peer-to-peer) e que sem os downloads do ano de 2002, as vendas teriam sido maiores em 7.8 % maior.

A evolução da Internet proporcionou um ambiente propício para que conteúdos protegidos por direitos autorais sejam adquiridos e compartilhados de uma maneira "livre". Entre os principais modos de execução desta atividade estão as comunidades on-line e redes

ponto-a-ponto. Neste âmbito, temos como relevante os trabalhos de Kwong e Lee (2004) que propõe um modelo inicial de análise do comportamento de indivíduos em comunidades de compartilhamento, utilizando a teoria do comportamento planejado como metodologias. White, Rao e Minakakis (2010) também utilizam o mesmo método para estudar os fatores que levam a uma maior ou menor propensão à pirataria e com base em entrevistas estruturadas busca indicar perfis mais ou menos propensos à prática de pirataria, além de discutir vantagens e desvantagens percebidas pelos participantes. Os resultados apontam que os jovens são mais propensos à prática e que os adultos, embora percebam melhor o componente ilegal intrínseco da atividade, também praticam usando uma justificativa financeira e a da impossibilidade de controlar o compartilhamento ponto-a-ponto tal como propõem Wang, Lv e Hong (2009).

De acordo com Byers et al. (2004), os principais agentes que permitem o vazamento de filmes são os próprios funcionários da indústria de cinema, mais frequentemente na fase de distribuição, expondo uma controvérsia entre sua atitude e a proteção do seu mercado de trabalho. Como forma de conter a pirataria durante a fase de distribuição do conteúdo, a tecnologia mais citadas na literatura foi a encriptação dos dados.

O aumento da pirataria na Internet e em geral ensejou uma corrida legal e tecnológica com a finalidade de sua contenção. Landau, Stratulate e Twilleager (2006) propõem uma abordagem na forma de jogo na qual o telespectador participasse ativamente dos filmes de forma parecida com os jogos iterativos como uma maneira de diminuir a pirataria. Uma alternativa que não se popularizou, partia do endurecimento legal, foi o caso da Nova Zelândia, que com base no The Copyright Amendment Act de 2011 prevê inclusive pena de prisão em caso de violação de direitos autorais. De acordo com Alcock e Nelson (2012) como resposta a este caso houve queda de quase metade de usuários DSL em redes ponto-a-ponto de compartilhamento, e, ao mesmo tempo, um pequeno aumento na utilização de métodos de compartilhamento anônimos, como os codificados.

Interessante que mesmo amparados legalmente, os detentores de direitos via de regra não reclamam estes direitos contra os usuários domésticos, continuando a mesma linha de exposição de fraquezas dos sistemas de marca d'água e buscas de formas possíveis de evitá-las, em sua maior parte através de aplicação de técnicas mais avançadas de modulação de sinais (BOURDON et al., 2010; ZHAI e WU, 2014; ROOPALAKSHMI e REDDY, 2012; PETITCOLAS e KIROVSKI, 2002; ASIKUZZAMAN, ALAME e PICKERING, 2015; KIROVSKI e PETITCOLAS, 2003). Larabi, Rosselli e Fernandez Maloigne (2009) propõem uma técnica de adição de ruído com fins de evitar a gravação de filmes em salas de cinema.

Um novo campo de estudo emergiu com esta problemática, composto pela análise forense de vídeos pirateados, buscando a partir da avaliação de detalhes técnicos dos sinais luminosos dos vídeos extrair meios de identificar os agentes que o piratearam, frequentemente pessoas que o gravaram em salas de cinema (RUDMAN et al., 2016; BAUDRY, CHUPEAU e LEFÈBVRE, 2009; HAJJ-AHMAD et al., 2017; THAKUR e KUMAR, 2015; MOREIRA-PÉREZ et al., 2013; ROOPALAKSHMI e REDDY, 2012; CHUPEAU, BAUDRY e DOERR, 2014).

Referente à questão econômica e regional que envolve a pirataria Chellappa e Shivendu (2003) propõem que sejam mantidos os padrões de codificação de DVDs com divisão entre regiões (ex. um aparelho de DVD americano apenas reproduz filmes gravados para esta região), ao mesmo tempo que fornece um método para cálculo de preço de venda com base na qualidade e região de venda do DVD, como um meio de desencorajar a pirataria. Especificamente relacionado ao rendimento da indústria cinematográfica, Rob e Waldfogel (2007) concluem que embora o consumo não-pago diminua a predisposição inicial dos indivíduos em pagar para assistir filmes, a indústria cinematográfica teria obtido um aumento significativo em seus lucros no período estudado, diferentemente do que aconteceu com a indústria fonográfica.

Em resumo, a literatura levantada não é conclusiva sobre os efeitos da pirataria, compartilhamento não autorizado de filmes ou outros bens de informação, sobre o consumo de cinema. Isso porque se por um lado o compartilhamento oferece alternativas de consumo de filmes em outras plataformas que não as telas de cinemas, o que teria efeito negativo sobre consumo, por outro lado oferece propaganda dos filmes, uma vez que os filmes são bens de experiência, quando se sugere um efeito positivo sobre consumo.

Segunda Etapa

Protocolo para revisão sistemática

Objetivos:

Adquirir uma visão abrangente dos principais métodos envolvidos e algoritmos de inteligência artificial empregados no processo de descobrimento de características comuns a perfis de consumidores e elencar trabalhos que comparem a aplicação de técnicas econométricas a algoritmos de aprendizagem de máquina.

Questões de pesquisa:

1. Que benefícios a aplicação de inteligência artificial vem proporcionando ao mercado em termos de proporcionar informações relevantes para o entendimento do comportamento do consumidor, suas decisões de consumo e possibilidades emergentes do uso destas informações.
2. Quais técnicas de inteligência artificial são mais usadas para a identificação de perfis de consumidores?
2. Existem estudos prévios que confrontem a aplicação de florestas aleatórias a técnicas econométricas de identificação de perfis de consumo?
3. Quais as dificuldades operacionais envolvidas nos procedimentos de análise de dados executados.

Intervenção: métodos e técnicas de avaliação

Controle: bibliografia recomendada pela a orientadora, teses e dissertações sobre a problemática obtidas na Internet.

População: estudos relativos à utilização inteligência artificial para fins de descoberta de perfis de consumidores, especificamente os que utilizam algoritmos de aprendizagem de máquina, independente do meio de publicação, desde que o mesmo esteja bem classificado no meio acadêmico.

Resultados: conhecimento amplo de como o campo da inteligência artificial pode ser utilizado para a identificação dos perfis de consumidores, principais métodos e algoritmos envolvidos.

Seleção de fontes:

As fontes deverão estar disponíveis via web, preferencialmente em bases de dados científicas das áreas de economia e computação.

Palavras-chaves:

“artificial intelligence, consumer feature, consumer profile”

Fontes consultadas:

- Biblioteca Digital do IEEE (<http://ieeexplore.ieee.org/Xplore/>)
- Biblioteca Digital da ACM (<http://portal.acm.org/>)
- JSTOR (<https://www.jstor.org/>)
- NBER (<https://www.nber.org/>)

- Science Direct (<https://www.sciencedirect.com/>)

Idioma(s) dos artigos:

Inglês.

Critérios de inclusão e exclusão dos trabalhos:

Critérios de inclusão:

1. Serão incluídos trabalhos publicados e disponíveis integralmente em bases de dados científicas ou em versões impressas.
2. Artigos que tratem da aplicação de aprendizagem de máquina à análise do perfil de consumidores e seu comportamento de consumo.

Critérios de exclusão:

1. Artigos anteriores a 1990.
2. Artigos em língua diversa do Inglês.
3. Artigos aos quais a USP não provê acesso.

Estratégia de extração da informação:

Para cada artigo aceito foi preenchida uma ficha com as seguintes informações:

1. Algoritmo empregado para a análise do perfil dos consumidores.
2. Fontes dos dados estudados.
3. Dados bibliográficos básicos.

Strings de busca utilizadas no formato exigido por cada motor:

Observação: As strings de busca variaram de acordo com os resultados iniciais para a string contendo todas as palavras-chave.

IEE: (((“artificial intelligence”) AND “consumer profile”) OR “consumer analysis”)

JSTOR: (((“artificial intelligence”) AND “consumer profile”) OR “consumer analysis”)

ACM: (+artificial +intelligence +consumer +analysis +consumer +profile)

Science Direct: “artificial intelligence” and “consumer profile”

Resultados:

A busca primária nas bases consultadas retornou 140 artigos, dos quais 17 foram selecionados para estudo de acordo com os critérios de inclusão e exclusão explicitados.

Tabela 27 – Artigos selecionados na segunda etapa da revisão de literatura

Ref.	Título	Ano
A1	Consumer-Centered Product Design Specifications Using Fuzzy Multivariate Regression Analysis	1993
A2	Trust Perception in Internet Shopping: Comparative Study of Customers in Japan and South Korea	2008
A3	Profiling blood donors in Egypt: A neural network analysis	2009
A4	Integration of heterogeneous models to predict consumer behavior	2010
A5	Self-clustering analysis of Shanghai's new pragmatic consumers	2011
A6	Electrical Consumers Data Clustering Through Optimum-Path Forest	2011
A7	Modelling customer choice and market development for future automotive powertrain technologies	2013
A8	Citizens as consumers: Profiling e-government services' users in Egypt via data mining techniques	2013
A9	The Influence Factors of Subscriber Locked in the Era of 4G, a Perspective of Consumer Behavior	2014
A10	A Data Mining Based Approach to a Firm's Marketing Channel	2015
A11	Typification of load curves for DSM in Brazil for a smart grid environment	2015
A12	Creation of a Data Observatory enables the uncovering of consumer behavior by client behavioral study, through the use of 1st generation loyalty and stimulation platforms	2015
A13	The Role of Big Data Analysis in New Product Development	2016
A14	AI-based methodology of integrating affective design, engineering, and marketing for defining design specifications of new products	2016
A15	Intelligent system for determining the consumer profile and generate alarm in case of significant deviations from the profile	2017
A16	Active demand side management for households in smart grids using optimization and artificial intelligence	2018
A17	On Designing Optimal Data Purchasing Strategies for Online Ad Auctions	2018
A18	* Social media analytics for enterprises: Typology, methods, and processes	2018

Fonte: Ticiane Sá da Justa, 2018

A partir da leitura dos trabalhos selecionados, foram identificados os seguintes algoritmos de aprendizagem de máquina utilizados para a identificação de perfis de clientes, destacando-se a aplicação de redes neurais, conforme vemos na tabela 5.

Tabela 28 – Resumo dos algoritmos usados nos trabalhos selecionados

Ref.	Algoritmo usado
A1	Análise de regressão (inespecificada) combinada a conjuntos fuzzy.
A2	Utiliza um programa chamado two step cluster porém não especifica qual algoritmo implementou.
A3	REDES NEURAIIS (Multilayer Perceptron e Probabilistic Neural Network) comparados a Análise de Discriminantes Lineares (LDA).
A4	Combinação de várias técnicas de mineração de dados, incluindo extração de regras com algoritmo Apriori, análise de matriz de frequências e várias técnicas de árvores de decisão, CHAID (Kass, 1980), QUEST (Loh & Shih, 1997), CART (Breiman, Friedman, Olshen, & Stone, 1984) e C5.0 (Quinlan, 1993) a fim de obter as melhores regras para sistemas de recomendação.
A5	Análise de componentes principais.
A6	Algoritmo de clustering Optimum-Path Forest (OPF)
A8	Três modelos de redes neurais artificiais (rede neural perceptron multicamada [MLP], rede neural probabilística [PNN] e rede neural auto-organizável de mapas [SOM]) e três técnicas de aprendizado de máquina (árvores de classificação e regressão [CART], multivariada splines de regressão adaptativa [MARS] e máquinas de vetores de suporte [SVM]) são comparadas à (análise discriminante linear [LDA]).
A9	Análise de componentes principais.
A10	Konohem network - um tipo de SOM para a redução de dimensionalidade do vetor de atributos e árvore de decisão C 5.0 (Quinlan, 1993) para a classificação e extração das regras.
A11	K means
A14	Regressão com conjuntos fuzzy e algoritmo de otimização do caos.
A15	Redes neurais.
A16	Redes neurais combinadas a algoritmos genéticos.

Fonte: Ticianá Sá da Justa, 2018

Com a evolução e popularização da internet, presenciamos o aumento da vida online dos indivíduos, através do uso massivo da rede para fins diversos, tais como compras online, redes sociais, consumo de notícias e publicidade em geral. Este aumento da presença online das pessoas proporciona uma rica fonte de informações sobre suas preferências e hábitos de consumo. Redes sociais, lojas online, provedores dos mais diversos conteúdos pagos ou gratuitos se beneficiam das técnicas de mineração de dados aliadas aos algoritmos de aprendizagem de máquina para a extração de informações relevantes sobre conjuntos de dados dos mais diversos tipos providos pelos usuários ao navegar.

Os hábitos de consumo e comportamentos dos consumidores são uma matéria importante para a tomada de decisões estratégicas das empresas que proveem estes produtos e serviços pois influenciam a tomada de decisões nos mais diversos níveis empresariais. Kwong, Jian e Luo (2016) propuseram um modelo que integrou a visão afetiva dos consumidores ao design e engenharia de novos produtos, já Shidon e Dongsheng (2016) apresentaram um modelo de análise de BIG DATA aplicado à área de marketing e desenvolvimento de produtos.

A análise do perfil dos clientes é amplamente utilizada para fins de definição de campanhas que visem alcançar um grupo específico e assim melhorar a alocação dos recursos publicitários e resultados alcançados. Podemos observar claramente a utilização da segmentação de clientes em sistemas de recomendação para o consumo e compra de bens online, nos quais os sites sugerem produtos e serviços de acordo com os perfis de clientes segmentados e suas atividades online.

Do ponto de vista da aquisição das informações para a execução dos estudos, o questionário direto foi a técnica mais utilizada, estando presente em 8 dos trabalhos analisados, dados de consumo obtidos diretamente dos fornecedores ficaram em segundo lugar ao passo que a utilização de dados não – estruturados não apareceu em nenhum estudo analisado, evidenciando talvez uma maior dificuldade em lidar com informações deste tipo.

Os trabalhos da área de energia evidenciaram o alto grau de interligação entre as disciplinas de estatística, aprendizagem de máquina e pesquisa operacional, as quais conjuntamente aplicadas buscaram chegar a níveis ótimos de fornecimento de energia além da identificação de fraudes e ajuste da demanda energética. A área de automação, medição e análise em tempo real do consumo também esteve fortemente presente nesta área, evidenciando um esforço global no desenvolvimento de uma rede inteligente de fornecimento energético.

Do ponto de vista computacional, a técnica de aprendizagem de máquina mais utilizada foram as redes neurais, seguidas pela análise de componentes principais. Devido à facilidade de extração das regras as árvores de decisão também foram bastante utilizadas.

Apenas dois estudos compararam métodos de aprendizagem de máquina com a técnica proveniente da área estatística de análise de discriminantes lineares (Fischer, 1938). Mohamed (2009) aplicou as técnicas de redes neurais (MultiLayer Perceptron e Redes Neurais Probabilísticas) para fins de classificação de possíveis doadores de sangue comparando –as com o método de Fischer, obtendo uma acurácia de 100% com as redes neurais, enquanto obteve apenas 83% com a análise de discriminantes lineares. Atribui o sucesso ao fato de as

redes neurais serem inerentemente não – lineares, o que as torna ideais para identificar estes tipos de relações entre os dados. Outro trabalho desta vez realizado por Mohamed e Ahmed realizaram o levantamento do perfil dos usuários de serviços públicos no Egito, comparando os resultados obtidos através da aplicação redes neurais (Multilayer Perceptron, Redes neurais probabilísticas e Mapas auto – organizáveis), árvores de decisão CART, splines de regressão adaptativa multivariada (MARS) e máquinas de vetores de suporte, também comparando estas abordagens com a análise de discriminantes lineares. Neste trabalho, a análise de discriminantes apresentou 94.3% de acurácia na classificação, Multilayer perceptron apresentou 98.3%, redes neurais probabilísticas apresentaram acurácia de 100%, árvores de decisão CART apresentaram acurácia de 99,48%, MARS apresentou 99.10% e SVM obteve uma taxa de 100%. Ou seja em ambos os trabalhos apresentados para conjuntos de dados diferentes, as técnicas de aprendizagem de máquina se mostraram mais confiáveis que a análise de discriminantes lineares.