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

NATHALIA NIEUWENHOFF

**Uma comparação da aplicação de métodos computacionais de classificação de dados
aplicados ao consumo de cinema no Brasil**

São Paulo

2017

NATHALIA NIEUWENHOFF

**Uma comparação da aplicação de métodos computacionais de classificação de dados
aplicados ao consumo de cinema no Brasil**

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 13 de
Abril de 2017. 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

2017

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)

Nieuwenhoff, Nathalia

Uma comparação da aplicação de métodos computacionais de classificação de dados aplicados ao consumo de cinema no Brasil / Nathalia Nieuwenhoff ; orientadora, Marislei Nishijima. – São Paulo, 2017.

130 p. : il

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

Versão corrigida

1. Mineração de dados. 2. Cinema - Brasil - Consumo - Simulação computacional. I. Nishijima, Marislei, orient. II. Título

CDD 22.ed.– 006.312

Dissertação de autoria de Nathalia Nieuwenhoff, sob o título "**Uma comparação da aplicação de métodos computacionais de classificação de dados aplicados ao consumo de cinema no Brasil**", 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 13 de Abril de 2017 pela comissão julgadora constituída pelos doutores:

Profª. Dra. Marislei Nishijima
Universidade de São Paulo
Presidente

Profª. Dra. Flavia Mori Sarti
Universidade de São Paulo

Prof. Dr. Ricardo Pires
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

Prof. Dr. Emerson Fernandes Marçal
Fundação Getúlio Vargas

Agradecimentos

Agradeço primeiramente a Deus Pai que, com a Sua infinita graça e misericórdia, possibilitou a realização deste trabalho, a partir dos dons concedidos por Ele.

Agradeço a toda a minha família, principalmente, meus pais e meu irmão, por todo apoio, amor incondicional e por serem meus melhores professores.

Agradeço enormemente a minha orientadora Marislei por todos os ensinamentos proporcionados, ajuda, disposição e compreensão durante todo o trabalho.

Resumo

NIEUWENHOFF, Nathalia. **Uma comparação da aplicação de métodos computacionais de classificação de dados aplicados ao consumo de cinema no Brasil**. 2017. 130 p. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2017.

As técnicas computacionais de aprendizagem de máquina para classificação ou categorização de dados estão sendo cada vez mais utilizadas no contexto de extração de informações ou padrões em bases de dados volumosas em variadas áreas de aplicação. Em paralelo, a aplicação destes métodos computacionais para identificação de padrões, bem como a classificação de dados relacionados ao consumo dos bens de informação é considerada uma tarefa complexa, visto que tais padrões de decisão do consumo estão relacionados com as preferências dos indivíduos e dependem de uma composição de características individuais, variáveis culturais, econômicas e sociais segregadas e agrupadas, além de ser um tópico pouco explorado no mercado brasileiro. Neste contexto, este trabalho realizou o estudo experimental a partir da aplicação do processo de Descoberta do conhecimento (KDD), o que inclui as etapas de seleção e Mineração de Dados, para um problema de classificação binária, indivíduos brasileiros que consomem e não consomem um bem de informação, filmes em salas de cinema, a partir dos dados obtidos na Pesquisa de Orçamento Familiar (POF) 2008-2009, pelo Instituto Brasileiro de Geografia e Estatística (IBGE). O estudo experimental resultou em uma análise comparativa da aplicação de duas técnicas de aprendizagem de máquina para classificação de dados, baseadas em aprendizado supervisionado, sendo estas *Naïve Bayes* (NB) e *Support Vector Machine* (SVM). Inicialmente, a revisão sistemática realizada com o objetivo de identificar estudos relacionados a aplicação de técnicas computacionais de aprendizado de máquina para classificação e identificação de padrões de consumo indica que a utilização destas técnicas neste contexto não é um tópico de pesquisa maduro e desenvolvido, visto que não foi abordado em nenhum dos trabalhos estudados. Os resultados obtidos a partir da análise comparativa realizada entre os algoritmos sugerem que a escolha dos algoritmos de aprendizagem de máquina para Classificação de Dados está diretamente relacionada a fatores como: (i) importância das classes para o problema a ser estudado; (ii) balanceamento entre as classes; (iii) universo de atributos a serem considerados em relação a quantidade e grau de importância destes para o classificador. Adicionalmente, os atributos selecionados pelo algoritmo de seleção de variáveis *Information Gain* sugerem que a decisão de consumo de cultura, mais especificamente do bem de informação, filmes em cinema, está fortemente relacionada a aspectos dos indivíduos relacionados a renda, nível de educação, bem como suas preferências por bens culturais.

Palavras-chave: Algoritmos de classificação. Bens de Informação. Consumo. *Support Vector Machine*. SVM. *Naïve Bayes*. Reconhecimento de Padrões.

Abstract

NIEUWENHOFF, Nathalia. **A comparison of the application of data classification computational methods to the consumption of film at theaters in Brazil**. 2017. 130 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2007.

Machine learning techniques for data classification or categorization are increasingly being used for extracting information or patterns from voluminous databases in various application areas. Simultaneously, the application of these computational methods to identify patterns, as well as data classification related to the consumption of information goods is considered a complex task, since such decision consumption patterns are related to the preferences of individuals and depend on a composition of individual characteristics, cultural, economic and social variables segregated and grouped, as well as being not a topic explored in the Brazilian market. In this context, this study performed an experimental study of application of the Knowledge Discovery (KDD) process, which includes data selection and data mining steps, for a binary classification problem, Brazilian individuals who consume and do not consume a information good, film at theaters in Brazil, from the microdata obtained from the Brazilian Household Budget Survey (POF), 2008-2009, performed by the Brazilian Institute of Geography and Statistics (IBGE). The experimental study resulted in a comparative analysis of the application of two machine-learning techniques for data classification, based on supervised learning, such as Naïve Bayes (NB) and Support Vector Machine (SVM). Initially, a systematic review with the objective of identifying studies related to the application of computational techniques of machine learning to classification and identification of consumption patterns indicates that the use of these techniques in this context is not a mature and developed research topic, since was not studied in any of the papers analyzed. The results obtained from the comparative analysis performed between the algorithms suggest that the choice of the machine learning algorithms for data classification is directly related to factors such as: (i) importance of the classes for the problem to be studied; (ii) balancing between classes; (iii) universe of attributes to be considered in relation to the quantity and degree of importance of these to the classifiers. In addition, the attributes selected by the Information Gain variable selection algorithm suggest that the decision to consume culture, more specifically information good, film at theaters, is directly related to aspects of individuals regarding income, educational level, as well as preferences for cultural goods.

Keywords: Classification Algorithm. Information goods. Consumption. *Support Vector Machine*. SVM. *Naïve Bayes*. Pattern recognition.

Lista de figuras

Figura 1 – Ilustração das Etapas do processo de KDD	22
Figura 2 – Ilustração do <i>Naïve Bayes</i>	27
Figura 3 – Ilustração do classificador binário SVM.	28
Figura 4 – Ilustração da demarcação das linhas que separam as classes através dos vetores- suporte do classificador SVM	29
Figura 5 – Passos seguidos no processo de seleção de estudos.....	33
Figura 6 – Distribuição da quantidade de anos estudos da revisão sistemática por ano.....	36
Figura 7 – Tela da aba de pré-processamento do Weka após importação do banco de dados.	49
Figura 8 – Funcionalidade de pré-processamento do banco de dados do Weka Explorer, aba “Preprocess”	54
Figura 9 – Funcionalidade de execução dos algoritmos de classificação no Weka Explorer, aba “Classify”	57
Figura 10 – Diagrama de 10-fold Cross-validation.	58
Figura 11 – Execução 1: Propriedades do classificador NB no Weka	61
Figura 12 – Execução 1: Propriedades do classificador SVM no Weka	62
Figura 13 – Execução 2: Propriedades do seletor de variáveis <i>Information Gain</i> no Weka ...	63
Figura 14 – Execução 2: Funcionalidade de seleção de variáveis do Weka, aba <i>Select attributes</i>	64
Figura 15 – Execução 2: Resultado parcial do ranking de atributos gerados pelo algoritmo <i>Information Gain</i> no Weka.....	65
Figura 16 – Execução 3: Propriedades do classificador <i>CostSensitive</i> associado ao NB no Weka	67
Figura 17 – Execução 3: Parametrização da matriz de custo do <i>Cost Sensitive classifier</i> no Weka	67
Figura 18 – Execução 3: Propriedades do classificador <i>Cost Sensitive</i> associado ao SVM no Weka	68
Figura 19 – Execução 1: Ilustração da curva ROC do NB gerada pelo Weka.....	80
Figura 20 – Execução 1: Ilustração da curva ROC do SVM gerada pelo Weka.....	80
Figura 21 – Execução 2: Ilustração da curva ROC do NB gerada pelo Weka.....	86
Figura 22 – Execução 2: Ilustração da curva ROC do SVM gerada pelo Weka.....	86
Figura 23 – Execução 3: Ilustração da curva ROC do NB gerada pelo Weka.....	91
Figura 24– Execução 3: Ilustração da curva ROC do SVM gerada pelo Weka.....	92

Figura 25 – Execução 4: Ilustração da curva ROC do NB gerada pelo Weka.....	97
Figura 26 – Execução 4: Ilustração da curva ROC do SVM gerada pelo Weka.....	98
Figura 27 – Aba <i>Main</i> do <i>software</i> Stata para configuração dos parâmetros do comando merge	120
Figura 28 – Aba <i>Options</i> do <i>software</i> Stata para configuração dos parâmetros do comando merge	121
Figura 29 – Aba <i>Results</i> do <i>software</i> Stata para configuração dos parâmetros do comando merge	122
Figura 30 – Tela inicial para acesso aos microdados da POF do pacote Data Zoom no Stata	123
Figura 31 – Tela do pacote Data Zoom para configuração da cesta de despesas para geração dos valores agregados no Stata.....	124

Lista de tabelas

Tabela 1 – <i>String</i> de busca genérica para condução da revisão sistemática	32
Tabela 2 – Relação dos trabalhos selecionados para revisão sistemática	33
Tabela 3 – Lista dos estudos selecionados para a revisão sistemática.....	37
Tabela 4 – Lista das áreas de aplicação dos estudos selecionados na revisão sistemática	38
Tabela 5 – Tabela de distribuição dos tipos de aplicação das técnicas de aprendizagem de máquina dos estudos primários da revisão sistemática.....	40
Tabela 6 – Locais de aquisição da POF (2008-2009) considerados para o consumo de cinema no estudo experimental	47
Tabela 7 – Exemplo de matriz de confusão gerada pelo programa Weka	70
Tabela 8 – Exemplo de matriz de confusão “invertida” gerada pelo programa Weka	71
Tabela 9 – Tabela de interpretação dos valores de concordância da estatística Kappa	72
Tabela 10 – Execução 1: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste	75
Tabela 11 – Execução 1: Tabela de tempo de execução dos algoritmos por tipo de teste	76
Tabela 12 – Execução 1: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka	78
Tabela 13 – Execução 1: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação no Weka.....	78
Tabela 14 – Execução 1: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka	79
Tabela 15 – Execução 2: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste	82
Tabela 16 – Execução 2: Tabela de tempo de execução dos algoritmos por tipo de teste	83
Tabela 17 – Execução 2: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka	85
Tabela 18 – Execução 2: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação	85
Tabela 19 – Execução 2: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka	85
Tabela 20 – Execução 3: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste	88
Tabela 21 – Execução 3: Tabela de tempo de execução dos algoritmos por tipo de teste	88

Tabela 22 – Execução 3: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka	89
Tabela 23 – Execução 3: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação	90
Tabela 24 – Execução 3: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka	91
Tabela 25 – Execução 4: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste	93
Tabela 26 – Execução 4: Tabela de tempo de execução dos algoritmos por tipo de teste	94
Tabela 27 – Execução 4: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka	95
Tabela 28 – Execução 4: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação	95
Tabela 29 – Execução 4: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka	97
Tabela 30 – Tabela de percentuais de eventos corretamente e incorretamente classificados para as quatro execuções deste estudo experimental	99
Tabela 31 – Tabela das métricas de acuracidade para a classe “Filme” referente às quatro execuções deste estudo experimental	100
Tabela 32 – Tabela das métricas de acuracidade para a classe “Não Filme” referente às quatro execuções deste estudo experimental	102
Tabela 33 – Tabela das métricas de estatística Kappa e área ROC referente às quatro execuções deste estudo experimental	103

Sumário

1	Introdução.....	13
1.1	<i>Justificativa para a pesquisa</i>	<i>14</i>
1.2	<i>Objetivos.....</i>	<i>19</i>
1.3	<i>Estrutura do documento</i>	<i>20</i>
2	Fundamentação teórica	21
2.1	<i>Descoberta do conhecimento.....</i>	<i>21</i>
2.2	<i>Mineração de Dados</i>	<i>22</i>
2.3	<i>Principais tarefas de Mineração de Dados.....</i>	<i>23</i>
2.4	<i>Técnicas de Mineração de Dados.....</i>	<i>24</i>
2.4.1	<i>Aprendizado de máquina</i>	<i>24</i>
2.4.2	<i>Naïve Bayes</i>	<i>26</i>
2.4.3	<i>Support Vector Machine</i>	<i>27</i>
3	Revisão Sistemática.....	30
3.1	<i>Questões de pesquisa</i>	<i>30</i>
3.2	<i>Critérios de inclusão e exclusão.....</i>	<i>31</i>
3.3	<i>Fonte de dados e estratégia de busca para seleção de estudos</i>	<i>32</i>
3.4	<i>Resultados gerais da revisão sistemática.....</i>	<i>33</i>
3.5	<i>Discussão dos resultados da revisão sistemática</i>	<i>41</i>
4	Estudo experimental	43
4.1	<i>Banco de dados.....</i>	<i>43</i>
4.2	<i>Software utilizado no estudo experimental</i>	<i>48</i>
4.2.1	<i>Stata</i>	<i>48</i>
4.2.2	<i>Weka</i>	<i>48</i>
4.3	<i>Aplicação das técnicas de KDD</i>	<i>50</i>
4.4	<i>Limpeza e integração dos dados.....</i>	<i>50</i>
4.5	<i>Seleção dos dados</i>	<i>50</i>
4.6	<i>Transformação dos dados</i>	<i>53</i>
4.7	<i>Mineração de Dados</i>	<i>53</i>
4.7.1	<i>Classes desbalanceadas.....</i>	<i>53</i>
4.7.2	<i>Seleção adicional de atributos por meio de algoritmos de seleção de variáveis ..</i>	<i>55</i>

4.7.3	Execução dos algoritmos de classificação	56
4.8	<i>Avaliação ou Pós-processamento</i>	69
4.9	<i>Visualização dos Resultados</i>	73
5	Apresentação e análise dos resultados obtidos	74
5.1	<i>Execução 1: Banco de dados completo (51 atributos)</i>	74
5.2	<i>Execução 2: Banco de dados após a seleção de variáveis por meio do algoritmo Information Gain</i>	81
5.3	<i>Execução 3: Banco de dados completo (51 atributos), considerando o algoritmo de classificação Cost Sensitive</i>	87
5.4	<i>Execução 4: Banco de dados após a seleção de variáveis por meio do algoritmo Information Gain considerando o classificador Cost Sensitive</i>	93
5.5	<i>Comparação geral entre as execuções</i>	99
6	Conclusão	104
	Referências	110
	Apêndice A – Detalhamento dos procedimentos de limpeza e integração dos dados..	116
	Apêndice B – Detalhamento dos procedimentos de transformação dos dados.....	125

1 Introdução

A evolução tecnológica constante resultou na mudança da lógica de consumo de alguns bens. Devido à facilidade de acesso à informação, as pessoas não necessariamente precisam se deslocar ou pagar diretamente para terem acesso a determinados bens de informação, como o cinema, visto que estes se encontram disponíveis na internet ou, também, através de recursos ilegais, como a pirataria, por exemplo.

Jehle e Reny (2000) discutem, a partir da teoria do consumidor, que o consumo é gerado por meio de uma tomada de decisão do consumidor, que, de forma geral, possui como dependência fatores como suas preferências individuais, orçamento, demanda e a oferta do produto. Para a teoria do consumidor, as escolhas de consumo são vistas a partir da capacidade de pagamento do consumidor, que se encontra diretamente relacionada à sua restrição orçamentária, gerando, assim, a necessidade de se fazer escolhas. A partir disso, as características pessoais dos indivíduos, bem como influências psicológicas, sociais e culturais, afetam tal tomada de decisão.

Ao mesmo tempo, Diniz (2009) afirma que o consumo de cultura, a partir dos bens de informação, é capaz de exprimir a situação de capital econômico, social e cultural das pessoas, de suas respectivas famílias e dos grupos nos quais estão inseridos de forma relacionada. Dessa forma, os padrões de decisão do consumo estão relacionados com as preferências dos indivíduos e dependem de uma composição de características individuais, variáveis culturais, econômicas e sociais segregadas e agrupadas. Identificar tais padrões se trata de uma tarefa desafiadora, devido à complexidade de tais fatores e sua heterogeneidade. A partir destes aspectos, também encontrados em diversas áreas de conhecimento, a necessidade de identificar padrões precisos e úteis nos dados tem exigido um aprofundamento contínuo do estudo de técnicas computacionais que otimizem e automatizem tais atividades.

Fayyad, Shapiro e Smyth (1996) propuseram um processo para análise de dados e geração de conhecimento denominado como Descoberta do conhecimento ou *Knowledge Discovery in Databases* (KDD). O KDD possui atividades iniciais como coleta, integração, seleção e transformação prévia dos dados para que o processo de Mineração de Dados (*data mining*) ocorra, bem como envolve atividades posteriores para análise e visualização dos resultados para geração de conhecimento às partes interessadas.

Como parte do processo de Mineração de Dados, que se trata de uma etapa do processo de KDD, a tarefa de Classificação de Dados se aplica com uma alternativa para predição da decisão de consumo dos bens de informação, visto que é capaz de examinar o

conjunto de registros marcados com seus atributos correspondentes e identificar as características que definem tais classes, sendo capaz, assim, de prever as classes que os registros futuros, ainda não classificados, pertencem (LORENA; CARVALHO, 2003). Adicionalmente, a escolha da abordagem e da técnica de Classificação de Dados e Reconhecimento de Padrões ou até mesmo uma combinação de uma ou mais técnicas está diretamente relacionada com o problema, ao contexto em que este se aplica e como os dados estão estruturados.

Neste contexto, neste trabalho será aplicado todo processo de KDD para um problema de classificação binária relacionado à decisão de consumo de um bem de informação (filmes em salas de cinema), classificando, assim, os indivíduos brasileiros que consumiram e não consumiram o produto em questão. O banco de dados deste trabalho se refere aos microdados da Pesquisa de Orçamento Familiar (POF), realizada em um período de 12 meses entre os anos 2008 e 2009 e disponibilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), contendo informações de consumo de cinema para um período de referência de 30 dias.

Adicionalmente, como parte do processo de KDD, a etapa de seleção dos dados será conduzida a partir de aspectos gerais da teoria econômica, após entendimento e análise do banco de dados objeto deste estudo, bem como a partir da aplicação de um algoritmo computacional para uma seleção adicional de atributos relevantes, conhecido como *Information Gain*. A etapa de Mineração de Dados será conduzida através da execução de dois algoritmos de aprendizagem de máquina, sendo estes *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), para treinamento e teste dos modelos para a tarefa de Classificação de Dados, resultando, assim, na análise comparativa dos resultados obtidos a partir destes algoritmos, considerando medidas relacionadas à acuracidade das técnicas e seus desempenhos computacionais.

1.1 Justificativa para a pesquisa

Segundo Canclini (1999), o consumo consiste no “conjunto de processos socioculturais nos quais se realizam a apropriação e o uso dos produtos”. A partir disso, o consumo possui a capacidade de agrupar as pessoas, bem como distingui-las e classificá-las, afinal, através dele é possível indicar características de quem consome, identificar diferenças sociais, culturais e econômicas dos indivíduos e como elas estão relacionadas. O Ministério Brasileiro da Cultura (2007) reforça esta ideia, afirmando que “o consumo cria ordem,

classifica as pessoas e as associa aos bens; enfim, o consumo ordena informações e organiza significados sobre as estruturas sociais”.

Neste cenário, o consumo está diretamente relacionado à cidadania e ao direito, em suas variadas esferas. No aspecto social, está conexo ao direito de acesso aos bens e serviços. Dessa forma, o consumo cultural diz respeito sobre a organização de uma sociedade e é um meio de se estabelecer condições para o exercício da cidadania, formação de indivíduos, bem como a participação destes no contexto político e social. A partir disso, os padrões de decisão do consumo são parte de uma composição de variáveis culturais, econômicas e sociais e identificar tais padrões, constitui uma tarefa desafiadora devido às múltiplas variáveis envolvidas e à necessidade de se mapear como tais variáveis se relacionam (DINIZ, 2009).

Segundo Shapiro e Varian (1999), bem de informação ou *good information* é um bem que pode ser digitalizado ou codificado como um fluxo de bits, que independente da fonte particular de valor, as pessoas estão dispostas a pagar para obtê-lo ou consumi-lo, a partir disso, filmes, músicas, páginas na web são alguns dos exemplos de bens de informação.

Dessa forma, o consumo de cultura aplicado aos bens de informação possui como desafio direto a evolução tecnológica constante. Afinal, os meios de reprodução de tais bens de informação são amplamente variados e vêm sofrendo uma mudança significativa na lógica de seu consumo.

Diniz (2009) afirma que o consumo cultural é afetado diretamente pelas variáveis socioeconômicas, educacionais e sócio demográficas de quem consome e, também, pelas variáveis geográficas que indicam não somente aspectos culturais, mas também variações de oferta dos produtos em cada região. Adicionalmente, as experiências sociais às quais os consumidores estão sujeitos influenciam suas preferências, dessa forma, estas são afetadas diretamente pelo capital pessoal, ou seja, experiências pessoais e individuais dos consumidores, bem como pelo capital social, experiências e influências proporcionadas pelo meio social em que os consumidores estão inseridos. A partir disso, o estudo dos determinantes de consumo de cultura requer o mapeamento das restrições e influências que o consumidor e os grupos nos quais este está inserido sofrem, resultando, assim, no entendimento aprofundado do público alvo de consumo.

Reforçando esta ideia, Throsby (1999) afirma que os produtos culturais, o que inclui os bens de informação, possuem valores de uso que são compostos pelos produtos culturais materializados e os que possuem valor simbólico. Dessa forma, nota-se que a atribuição de valor a eles não é uma tarefa trivial, afinal, é preciso conhecer mais detalhadamente características especiais destes bens, bem como as de quem os consome.

Aplicando este contexto ao cenário brasileiro, o Brasil além de ser um país conceitualmente dividido em cinco grandes regiões, é um país multicultural, o que torna ainda mais complexa a tarefa de mapeamento e identificação dos padrões de consumo da sua população. Afinal, além de sua cultura estar sujeita a um conjunto de incentivos financeiros com recursos públicos disputados que são aplicados, muitas vezes, sem nenhum grau de conhecimento do padrão de consumo, conforme afirmado por Hofstede (2010), as diferenças culturais entre as regiões do país são significativas, tornando ainda mais complexa a tarefa de definição de padrão de consumo do brasileiro de forma geral.

Dessa forma, visto que o consumo de cultura aborda a noção de cidadania impactada pelo acesso, classificar os indivíduos e identificar padrões de consumo de cultura, através dos bens de informação, torna-se cada vez mais necessário para os recursos públicos serem capazes de estabelecer políticas públicas voltadas ao setor eficazes e, ao mesmo tempo, reduzir o consumo ilegal dos bens de informação, combatendo iniciativas como a pirataria, por exemplo.

Adicionalmente, de acordo com Silva (2008), a administração pública do Brasil não possui dados consistentes e precisos sobre as práticas culturais dos brasileiros e como estas se refletem nas diversas regiões do país, bem como sobre os hábitos de consumo cultural da sua população, ocasionando, assim, a elaboração de políticas públicas culturais ineficazes e que não atendem às necessidades para um consumo cultural efetivo e amplamente acessível à população brasileira.

Em paralelo, é de alto interesse das empresas responsáveis pela produção e reprodução de bens de informação a identificação de padrões de consumo no contexto de auxiliar as estratégias de venda e divulgação destes para um público alvo mais preciso, bem como identificar e prever cenários de evolução tecnológica, possíveis concorrências e meios para veiculação de tais bens, maximizando lucros e aumentando a satisfação de seus clientes.

Além do contexto cultural e social, a relevância do consumo de cultura dos brasileiros em relação ao bem de informação filmes em salas de cinema, é reforçada no aspecto econômico, afinal, de acordo com o Ministério Brasileiro da Cultura (2007), o consumo cultural da população é parte não somente de investimentos públicos e privados, bem como da geração do Produto Interno Bruto (PIB). Segundo o estudo de valor adicional do setor de audiovisual efetuado pela Agência Nacional de Cinema (ANCINE), agência reguladora e fiscalizadora do mercado brasileiro de cinema e do audiovisual, as atividades econômicas do setor de audiovisual, que “compreende a indústria cinematográfica e videofonográfica do país”, isto é: os agentes de produção, distribuição e exibição dos segmentos de cinema

(salas de exibição), TV paga (comunicação eletrônica de massa por assinatura), TV aberta (radiodifusão de sons e imagens), vídeo doméstico, vídeo por demanda e mídias móveis, representaram uma renda de R\$22,9 bilhões na economia brasileira, bem como vem apresentando evolução do tamanho do setor entre os anos de 2007 a 2013. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), valor adicionado “refere-se ao valor que a atividade acrescenta aos bens e serviços consumidos no seu processo produtivo”.

Adicionalmente, segundo o Informe Anual de Acompanhamento de Mercado emitido pela ANCINE (2009), o ano de 2009, período dos dados a serem utilizados neste trabalho, registrou um público total de 112.683.383 espectadores, que corresponde a uma renda no valor de R\$ 969.783.735,77. Estes números representam o maior público dos últimos cinco anos, bem como um crescimento de 25,26% e 32,93%, no público e renda, respectivamente, referente ao ano anterior, 2008.

A partir disso, o maior conhecimento permite que se avalie mais precisamente os recursos envolvidos para garantir sua otimização neste setor, que é relevante para o país. Porém, o cenário de consumo cultural, conforme detalhado anteriormente, possui um conjunto complexo de variáveis, bem como requer a manipulação de bases de dados volumosas, devido não só ao tamanho da população brasileira e disponibilização deste tipo de informação em microdados, mas também aos fatores sociais, culturais e econômicos que estão relacionados, a maioria dos estudos sobre tais padrões de consumo que envolvem os bens de informação se baseia em análises qualitativas sem o uso sistemático de informações quantitativas, bem como a combinação destas.

Em paralelo, devido à quantidade e complexidade de tais dados, a exploração manual destes para geração de conhecimento se torna inviável, necessitando, assim, da aplicação de técnicas automatizadas para tal fim. Com o objetivo de reduzir a complexidade do problema relacionado à manipulação e a identificação do relacionamento entre tais dados para auxílio nas tomadas de decisões de forma que se atinja uma precisão ótima nos padrões identificados e, em paralelo, não comprometa o desempenho computacional das aplicações, visto que, nos dias atuais, estas estão cada vez mais exigentes, tais técnicas estão sendo cada vez mais estudadas, aplicadas e desenvolvidas.

Dessa forma, o interesse na área de Reconhecimento de Padrões, que segundo Schalkoff (1991), trata-se do estudo de descrição e classificação de medições, tem crescido muito nos últimos anos e está em constante desenvolvimento, visto que se trata de uma tarefa de classificação ou categorização de dados, cujo objetivo é discriminar amostras de objetos e classificar corretamente as amostras futuras.

Neste contexto, tais fatores conduziram à exploração e desenvolvimento do processo de extração de conhecimento em banco de dados volumosos, não-triviais e dinâmicos, conhecido como KDD para geração de conhecimento a partir de padrões de dados válidos, complexos e potencialmente úteis (Fayyad, Shapiro e Smyth, 1996). As etapas de seleção e Mineração de Dados fazem parte do processo em questão e estão diretamente relacionadas com o Reconhecimento de Padrões das bases de dados para identificação de atributos relevantes, bem como classificação destes.

Mais especificamente, a aplicação da Mineração de Dados requer, inicialmente, que dois fatores sejam considerados, sendo eles a tarefa e a técnica de mineração. A tarefa consiste na definição e detalhamento do objetivo da exploração dos dados, ou seja, qual(is) é(são) a(s) categoria(s) de padrões que se deseja obter. Já a técnica se trata do método a ser utilizado para obtenção de tais padrões de interesse. Atualmente, há diversas técnicas de mineração, dentre elas as estatísticas e de aprendizado de máquina, por exemplo. Em relação às tarefas, estas se subdividem, basicamente, em análise de regras de associação, análise de padrões sequenciais, classificação e predição, análise de agrupamentos e análise de *outliers*.

A função da tarefa de Classificação de Dados, conforme apresentado por Lorena e Carvalho (2013), é examinar o conjunto de registros marcados com seus atributos correspondentes e identificar as características (variáveis) que definem tais classes, predizendo, assim, as classes às quais os registros ainda não classificados pertencem. A partir disso, a aplicação desta técnica possibilita classificar, de forma automatizada, a partir do relacionamento de dados quantitativos e qualitativos, dados relacionados a decisão de consumo, o que inclui o consumo dos bens de informação.

A contribuição deste trabalho é focada na aplicação e análise comparativa de dois algoritmos de aprendizagem de máquina para classificação de dados popularmente utilizados, sendo eles *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), para classificação binária da decisão do consumo do bem de informação filme (cinema) pelos brasileiros, bem como as vantagens e desvantagens da utilização de cada um destes algoritmos, a partir das métricas relacionadas à acurácia e seus desempenhos computacionais, a partir do grande volume de dados disponível da Pesquisa de Orçamento Familiar (POF) realizada pelo Instituto Brasileiro de Geografia e Estatística.

Assim, sendo o Brasil um país em desenvolvimento, com extrema desigualdade de renda, a tarefa de classificação de consumidores de um bem de informação cultural, que também requer uma seleção prévia dos atributos relevantes para esta decisão de consumo, contribui não apenas para reduzir a falha na literatura, bem como para auxiliar os recursos

públicos na definição de políticas públicas que viabilizam o acesso legal e amplo a estes bens, bem como auxiliar recursos privados nas estratégias de vendas e produção.

1.2 Objetivos

Considerando como premissa que a Mineração de Dados, através da aplicação de algoritmos de aprendizagem de máquina para a tarefa de Classificação de Dados, está, nos dias atuais, sendo amplamente utilizada de forma multidisciplinar nos mais variados contextos, esta também pode ser aplicada para classificação da decisão de consumo de um bem de informação.

Dessa forma, o principal objetivo deste trabalho é a análise comparativa das vantagens e desvantagens relacionadas às métricas de acurácia e desempenho computacional, da aplicação do processo de KDD, especificamente, da etapa de Mineração de Dados, a partir de dois algoritmos computacionais, sendo estes *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), para a tarefa classificação binária da decisão do consumo do bem de informação filme (cinema) pelos brasileiros, a partir dos microdados da POF realizada pelo IBGE entre os anos de 2008 e 2009.

Para atingir o objetivo geral deste trabalho, os seguintes objetivos específicos foram estabelecidos:

- Efetuar uma revisão bibliográfica sobre as técnicas de Mineração de Dados e Reconhecimento de Padrões e seus respectivos algoritmos de aplicação.
- Efetuar uma revisão sistemática sobre a aplicação das técnicas de Mineração de Dados para a tarefa de classificação de dados relacionados ao consumo de forma geral.
- Aplicar todo o processo de KDD, considerando aspectos da teoria econômica para seleção geral dos atributos, após entendimento e análise do banco de dados objeto deste estudo, bem como a partir da aplicação de um algoritmo computacional para uma seleção adicional de atributos relevantes, conhecido como *Information Gain*. A etapa de Mineração de Dados será efetuada a partir da execução de algoritmos *Naïve Bayes* (NB) e *Support Vector Machine* (SVM) com o objetivo de classificar os registros referentes ao consumo e não consumo de cinema pelos brasileiros entrevistados na POF.

- Apresentar uma análise comparativa dos resultados obtidos a partir da aplicação de tais técnicas de Mineração de Dados para identificação dos padrões de consumo do bem de informação em questão, conforme as medidas de acuracidade das técnicas, bem como seu respectivo desempenho computacional, apresentando, neste contexto, uma análise de tais técnicas para bases de dados e problemas de classificação da mesma natureza deste estudo.

Adicionalmente, busca-se com este trabalho oferecer como resultado aos pesquisadores da área a partir da aplicação das técnicas em um objeto pouco explorado neste aspecto, bem como às empresas do setor e governo que podem utilizar os resultados obtidos para otimização de seus processos e especificação de novos projetos.

1.3 Estrutura do documento

Esta dissertação é composta por seis capítulos, sendo o primeiro deles a introdução ao tema, subdivididos da seguinte forma:

- O capítulo 2 contém a fundamentação teórica do trabalho, a partir da contextualização e a revisão bibliográfica, apresentando os principais conceitos relacionados a Knowledge Discovery in Databases (KDD), Mineração de Dados e suas respectivas tarefas e técnicas, com ênfase nos algoritmos de aprendizagem de máquina para classificação de dados utilizados no estudo experimental, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM).
- O capítulo 3 apresenta a condução e os resultados da revisão sistemática referente aos trabalhos de aplicação dos algoritmos de aprendizagem de máquina para Reconhecimento de Padrões de consumo.
- O capítulo 4 apresenta o detalhamento da metodologia de pesquisa para realização do estudo experimental que envolveu a aplicação de todo o processo de KDD com o objetivo de classificar os registros referentes ao consumo e não consumo de cinema pelos brasileiros entrevistados na POF.
- O capítulo 5 apresenta os resultados, bem como a análise comparativa das técnicas de classificação de dados aplicadas para identificação de padrões de consumo de filmes em salas de cinema no Brasil.
- O capítulo 6 apresenta a conclusão do trabalho, destacando as suas principais contribuições, suas limitações e trabalhos futuros.

2 Fundamentação teórica

2.1 Descoberta do conhecimento

O modelo tradicional que efetua a transformação dos dados em conhecimento consiste no processamento dos dados por especialistas que apresentam estes de forma que possam ser analisados. Porém, nos dias atuais, devido à automação dos processos e, principalmente, à elevada quantidade de dados relacionados, que podem se encontrar até em fontes distintas, tal prática se torna inviável. A partir disso, foi apresentado o conceito de Descoberta de Conhecimento das bases de dados, também denominado *Knowledge Discovery in Databases (KDD)* por Fayyad, Shapiro e Smyth (1996).

De acordo com Amo e Rocha (2003), o conceito de KDD e Mineração de Dados (*Data mining*) é considerado sinônimo por algumas fontes, porém, o que é comumente considerado é a Mineração de Dados como uma etapa do processo de KDD, conforme detalhamento das etapas do processo em questão:

1. Limpeza dos dados: consiste na eliminação de dados inconsistentes, incorretos ou incompletos.
2. Integração dos dados: consiste na junção e combinação dos dados apresentados de forma fragmentada, consolidando em apenas uma fonte de dados centralizada.
3. Seleção dos dados: consiste na definição dos atributos e variáveis que são relevantes para aplicação das técnicas em questão, a partir do conhecimento do tema.
4. Transformação dos dados: consiste na adaptação dos dados para execução da etapa de mineração. Segundo Camilo e Silva (2009), não há um critério específico para execução desta etapa, bem como a técnica a ser utilizada depende dos objetivos pretendidos, como por exemplo, a normalização, que consiste na inserção de variáveis em uma mesma escala, agrupamento, que se trata da junção de valores, ou até mesmo a criação de novos atributos, a partir de combinações dos já existentes, e a conversão do formato dos atributos, podendo ser requerida dependendo da técnica de Mineração de Dados a ser utilizada.
5. Mineração: consiste na realização da análise dos dados e na aplicação de tarefas e técnicas para geração de conhecimento. O detalhamento desta etapa se encontra descrito na seção 2.2 deste capítulo.

6. Avaliação ou Pós-processamento: consiste na análise dos resultados obtidos na etapa anterior para geração de conhecimento a partir do objetivo inicial proposto.
7. Visualização dos Resultados: consiste na apresentação do conhecimento obtido a partir da base de dados utilizada aos usuários ou partes interessadas.

A figura 1 ilustra todas as etapas que compõem o processo de KDD.

Figura 1 – Ilustração das Etapas do processo de KDD

Fonte: Amo e Rocha, 2003

2.2 Mineração de Dados

Segundo Fayyad, Shapiro e Smyth, 1996, o termo Mineração de Dados ou *Data Mining* pode ser definido como um "passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados". Segundo Camilo e Silva (2009), apesar de o processo de Mineração de Dados suportar a descoberta do conhecimento de um volume elevado de dados, este processo não é puramente automático. A partir disso, existe uma série de ferramentas para execução dos algoritmos de mineração, bem como a análise dos resultados ainda precisa ser efetuada por seres humanos.

No contexto de Mineração de Dados, Bueno e Viana (2012) destacam a importância de se considerar dois fatores para sua execução, sendo eles a tarefa e a técnica de mineração. A tarefa se refere à especificação do objetivo de busca do dado, ou seja, o tipo de categoria e informação que se deseja obter. Já a técnica se refere ao método que vai ser utilizado para

executar a tarefa definida, ou seja, o que será feito para descobrir o padrão que interessa. Esta pode se dividir em diversos tipos, sendo por exemplo, estatística, aprendizado de máquina, crescimento poda-validação, entre outros.

2.3 Principais tarefas de Mineração de Dados

Camilo e Silva (2009) afirmam que o processo de Mineração de Dados possui diversos tipos de tarefas para suportá-los, sendo as mais comuns:

- Classificação, que consiste na análise das características dos dados para atribuição a uma categoria específica previamente definida. Para isso, cada registro possui os atributos de predição, bem como o alvo. Podendo ser utilizada para identificar pessoas, classificar objetos e diagnosticar doenças, por exemplo.
- Regressão, é considerada similar à categorização, porém, consiste na tarefa de identificação do registro a partir de um valor numérico e não categórico. Dessa forma, a regressão possui técnicas conhecidas como: (i) regressão linear, que se refere à técnica em que a relação da variável preditora e a predição possuem comportamentos lineares, o que possibilita relacionar uma variável preditora a uma ou mais predições; e (ii) regressão não-linear, que consiste na tarefa de regressão quando a relação da variável preditora e da predição não possuem comportamentos lineares.
- Agrupamento (*Clustering*), consiste na tarefa de aproximação/agrupamento dos registro com base nas similaridades entre eles, segmentando os bancos de dados em grupos ou subconjuntos, porém, não requerendo que estes estejam previamente categorizados, visto que o próprio algoritmo é capaz de descobrir as classes.
- Associação, consiste na tarefa de identificação dos relacionamento dos atributos, no formato de “SE...ENTÃO”. A partir disso, o objetivo da tarefa em questão é identificar, por exemplo, se o atributo x implica em y . Sua implementação geralmente ocorre a partir de técnicas como *Apriori*, *sampling*, DHP, entre outras (PIMENTEL e OMAR, 2006).
- Descrição, consiste na tarefa de descrever os padrões revelados pelos dados, sendo muito utilizada para análise exploratória de tais dados.

Adicionalmente, nota-se que as tarefas de Mineração de Dados possuem funções distintas e precisam ser escolhidas de acordo com o problema e com o tipo de conhecimento que precisa ser gerado e, para cada tarefa, há diversos tipos de técnicas que podem ser

executadas de forma isolada ou combinadas para a realização das atividades de Mineração de Dados. A decisão por tais combinações depende de fatores relacionados ao objetivo do estudo, desempenho computacional e grau de precisão de cada técnica aplicada a cada situação-problema.

2.4 Técnicas de Mineração de Dados

A partir do objetivo geral deste trabalho, que consiste na aplicação de algoritmos computacionais para classificação dos indivíduos em relação à sua tomada de decisão de consumo de um bem de informação cultural, filmes em salas de cinema, a tarefa considerada neste estudo será a classificação. Como a aplicação das técnicas de mineração depende diretamente da tarefa definida, foram consideradas neste estudo duas técnicas de aprendizado de máquina, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM).

2.4.1 Aprendizado de máquina

O Aprendizado de Máquina (AM) é uma área da Inteligência Computacional que estuda o desenvolvimento de métodos que possibilitam a extração de conhecimento a partir de dados. Com isso, é avaliada a capacidade de um programa de computador aprender a melhorar seu desempenho para executar determinada tarefa, a partir de experiências passadas (Mitchell, 1997). Adicionalmente, de acordo com Han e Kramer (2006), o processo de desenvolvimento das técnicas de aprendizado de máquina foi baseado em diversas áreas do conhecimento, como por exemplo, sistemas biológicos, o que gerou as redes neurais e algoritmos genéticos, que possuem como associação o aprendizado humano (cérebro) e a evolução genética, respectivamente, bem como aprendizado simbólico, representado pelas árvores de decisão, ou também, nas teorias estatísticas, o que resultou em técnicas como *Support Vector Machine* (SVM).

Segundo Han e Kramer (2006), o Aprendizado de Máquina é aplicado no processo de indução, que consiste em um conjunto de treinamento de um classificador para previsão das classes do domínio para o qual foi treinado. As técnicas de AM são comumente classificadas em dois tipos de paradigmas, Aprendizado Supervisionado e Aprendizado Não-Supervisionado (HAYKIN 2009), e a escolha destes paradigmas determina como tal aprendizado ocorrerá a partir de uma base de dados. Estas classificações se distinguem a partir dos seguintes aspectos:

- Aprendizado Supervisionado, consiste no treinamento a partir de uma pré-categorização dos dados, ou seja, exemplos que são compostos pelo objeto de entrada e o valor de saída esperado, tendo, assim, a figura de um “professor-externo”. A partir disso, treinamento do algoritmo acontece pela análise dos dados de treinamento para produção de uma saída inferida já apresentada, podendo, posteriormente, ser aplicado para classificação de outros dados de entrada do mesmo domínio. Árvores de decisão, *k-Nearest Neighbour* (k-NN), *Naïve Bayes* e *Support Vector Machine* (SVM) são exemplos de algoritmos que tratam este tipo de aprendizado de máquina.
- Aprendizado não-supervisionado consiste em identificar uma estrutura não conhecida sem uma saída já definida, ou seja, não há essa figura de “professor”. O algoritmo aprende a agrupar as entradas a partir das relações, padrões, categorias, entre outros, que identifica, codificando a saída, com base em uma medida de similaridade entre os atributos. *Clustering*, Estatística, k-means são exemplos de algoritmos que tratam este tipo de aprendizado de máquina.

Para a tarefa de classificação, além das técnicas de aprendizagem de máquina escopo do estudo experimental a ser realizado neste trabalho, sendo elas *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), a serem detalhadas nas seções seguintes, 2.4.2 e 2.4.3, respectivamente, segundo Camilo e Silva (2009), as seguintes técnicas de aprendizagem de máquina também são comumente utilizadas:

- Árvores de decisão (*decision trees*), que consistem na representação do conhecimento a partir de um número finito de classes. Ou seja, a partir da variável de avaliação definida, o algoritmo retorna as características associadas a ela a partir dos “ramos da árvore.” Os nós correspondem aos atributos, as ligações entre os nós representam os valores para tais atributos e as folhas representam diferentes classes. A partir disso, após a definição da área para realização de um novo registro, é necessário seguir a árvore, cujo início é a raiz.
- Classificação baseada em regras, que é similar à Associação, visto que possui o seguinte formato, SE condição ENTÃO conclusão, cujo objeto é criar associações, pares de registros que possuem similaridade.
- Redes neurais (Neural Network), que são baseadas no conceito de neurônio artificial análogo ao neurônio natural, envolvendo estruturas matemáticas que possuem habilidade de aprendizado, dessa forma, cada unidade de entrada e saída é interligada às outras por conexões que possuem pesos particulares, variando entre -1 e 1, que

correspondem à intensidade da conexão. Apesar de a técnica em questão ter a capacidade de identificar padrões para os quais não foi treinada, ela requer um longo período para treinamento, ajustes detalhados em seus parâmetros, bem como é complexa para utilização.

- Classificação baseada em regras de associação, que se trata da aplicação das regras de associação para classificação de dados é uma abordagem recente, cujo objetivo é basicamente analisar os dados de treinamento para que se obtenham os itens frequentes, e a partir destes são geradas as regras para Classificação de Dados.
- Algoritmos genéticos, que consiste na técnica baseada na teoria da evolução, a partir disso, a população inicial é aleatória e a seguinte é originada a partir da evolução da anterior, em que seus atributos passem por um processo de mutação, até que os atributos atinjam o objetivo proposto. Uma das diferenças desta técnica em relação às demais é que esta trabalha com uma população e não somente um único ponto, utilizando regras de transição probabilísticas e não determinísticas.

2.4.2 Naïve Bayes

Os métodos de aprendizagem bayesianos utilizam um modelo probabilístico baseado no conhecimento prévio do problema, o qual é combinado com exemplos de treinamento para determinar a probabilidade final de classificação (MITCHELL, 1997).

O algoritmo Naïve Bayes (NB), como sugerido pelo seu próprio nome, é baseado no Teorema de Bayes. Consiste em uma técnica baseada em probabilidades, ou seja, é baseado na probabilidade de um evento ocorrer dada a probabilidade de outro evento que já ocorreu.

A equação 1 exemplifica a probabilidade condicional de A dado B, isto é, a probabilidade de observar a hipótese A, dado que B é verdadeiro, onde A e B são eventos e $P(B) \neq 0$ (Segaram, 2007):

$$P(A | B) = (P(B|A)P(A))/P(B) \quad (1)$$

De acordo com Gallagher, Madden e D'Arcy (2015), o classificador é chamado *naïve*, ou seja, ingênuo, porque assume que os atributos do conjunto de dados são independentes, o que significa que cada atributo x_i influencia a classe y , mas não influencia outros atributos, x_1 a x_n . A figura 2 ilustra este conceito.

Figura 2 – Ilustração do *Naïve Bayes*

Fonte: Gallagher, Madden and D'Arcy (2015)

Em seguida, o classificador NB executa o cálculo de probabilidades para a classificação de dados realizando a leitura dos dados de treinamento somente uma vez para estimar as probabilidades. Além disso, o NB pode ser executado a partir de dois modelos estatísticos: binário (0 para não ocorrência de algum atributo, 1 para pelo menos uma ocorrência) ou multinomial (vetor contendo o número de vezes que cada atributo ocorreu).

2.4.3 *Support Vector Machine*

De acordo com Vapnik (1995), o algoritmo de aprendizagem de máquina *Support Vector Machine* (SVM) é uma técnica de aprendizagem de máquina que utiliza a teoria de aprendizado estatístico e teoria da otimização para classificação binária ou seja, classificação que envolve apenas duas classes. A forma original do SVM, que consiste na classificação binária e linear, será considerada objeto deste estudo.

Dessa forma, o SVM original consiste em, dado um conjunto de treinamento como entrada, separar o conjunto em duas amostras, estas podendo ser denominadas genericamente como amostras positivas e negativas, gerando, assim, estas duas saídas de classificação, sendo, assim, considerado um classificador linear binário não probabilístico.

Conforme apresentado por Kinto (2011), o SVM analisado do ponto de vista geométrico consiste em um método que busca a melhor superfície no conjunto das superfícies disponíveis do espaço r -dimensional, que se trata de uma superfície de decisão, afinal, efetua a separação entre os eventos buscando maximizar a distância entre as fronteiras de tais superfícies. A partir disso, o SVM é um classificador linear cuja função f discriminante geral é dada pela equação 2.

$$f(x, w, b) = \text{sgn}(w \times x + b) \quad (2)$$

Onde, $sgn(q) = +1$ se $q \geq 0$ e $sgn(q) = -1$ se $q < 0$ e b é o *bias*, parâmetro que permite que o hiperplano separador se posicione no local correto para separação das amostras. A figura 3 apresenta a ilustração de um SVM linear em que os eixos ordenados representam as dimensões das amostras e estas são identificadas como +1 os círculos da parte superior, representando as amostras positivas, e -1 os quadrados da parte inferior, representando as amostras negativas (KINTO, 2011).

A partir disso, a capacidade de generalização do SVM é dada pela melhor função dentre as possíveis do hiperplano que melhor discrimina as classes. Considerando a existência de várias retas, conforme apresentado na figura 4, nota-se que estas poderiam estar no lugar da reta apresentada figura 3, dessa forma, o objetivo do SVM é encontrar a maior margem de separação entre as classes. Assim, o conjunto de características selecionadas para descrever um padrão é um vetor, e os vetores suporte são a fronteira dessa maximização, que separa as categorias, sendo estes ilustrados na figura 4 através de um círculo ao seu redor. Para conseguir uma boa capacidade de generalização, o objetivo do algoritmo é escolher a melhor função discriminante das classes entre as possíveis.

Figura 3 – Ilustração do classificador binário SVM.

Fonte: Kinto (2011)

Figura 4 – Ilustração da demarcação das linhas que separam as classes através dos vetores-suporte do classificador SVM

Fonte: Kinto (2011)

Neste trabalho, a técnica SVM utilizará a implementação *Sequential minimal optimization* (SMO) apresentada por Platt (1998) para a formação de um vetor de suporte classificador, que define dois pontos por iteração para fazer a otimização. O SMO é considerado um dos algoritmos de implementação mais rápidos e fáceis para o caso geral de SVM.

3 Revisão Sistemática

Este capítulo apresenta o detalhamento da revisão sistemática conduzida como parte deste estudo. O objetivo da revisão sistemática foi identificar os métodos e técnicas de aprendizado de máquina para classificação de dados e Reconhecimento de Padrões comumente utilizados no estado na arte, bem como observar os de experimentos relacionados à aplicação destes, preferencialmente, no contexto de classificação de dados de decisão de consumo.

3.1 Questões de pesquisa

Foram elaboradas uma questão principal de pesquisa e uma questão secundária para atender aos objetivos propostos desta Revisão Sistemática, sendo estas:

QP.1: *Quais tipos de técnicas, ou seja, algoritmos de aprendizagem de máquina estão sendo comumente aplicados para classificação de dados e Reconhecimento de Padrões relacionados à decisão de consumo?*

Para responder esta pergunta foram consideradas os nomes dos algoritmos e suas respectivas técnicas de classificação de dados e Reconhecimento de Padrões, estas podendo ser redes neurais, *Naïve Bayes* (NB), árvores de decisão e *Support Vector Machine* (SVM), por exemplo. Adicionalmente, também foram consideradas as seguintes classificações relacionadas aos tipos de aprendizado de tais algoritmos, estes podendo ser Aprendizado Supervisionado e Aprendizado não-supervisionado.

QP.2: *Quais são as formas de aplicação das técnicas de aprendizagem de máquina para Classificação de Dados e Reconhecimento de Padrões?*

Para responder esta pergunta foram considerados os tipos e formas que tais técnicas podem ser aplicadas, estes podendo ser, por exemplo, relacionados à implementação de uma nova técnica ou abordagem, estudo experimental, análise comparativa entre técnicas, aplicação da técnica para um contexto ou base de dados não explorado anteriormente.

3.2 *Crterios de incluso e excluso*

Para a seleo de trabalhos que estejam diretamente relacionados ao objetivo desta reviso sistemtica, foram considerados os seguintes crterios de incluso e excluso, respectivamente:

Crterios de incluso

CI.1: o artigo trata essencialmente de aprendizagem de mquina aplicada classificao de dados e Reconhecimento de Padres atravs de um estudo experimental envolvendo a aplicao de tcnicas e anlise comparativa entre estas, preferencialmente, para o contexto de deciso de consumo;

CI.2: o artigo trata essencialmente de aprendizagem de mquina aplicada classificao de dados e Reconhecimento de Padres atravs de uma proposta de uma nova implementao ou abordagem de uma tcnica de classificao de dados e Reconhecimento de Padres, bem como uma combinao destas, apresentando seus benefcios e vantagens em relao s tcnicas j existentes, preferencialmente, para o contexto de deciso de consumo;

CI.3: o artigo trata essencialmente de aprendizagem de mquina aplicada classificao de dados e Reconhecimento de Padres atravs da aplicao de uma ou mais tcnicas para um problema de classificao de dados ainda no explorado, preferencialmente, para o contexto de deciso de consumo.

Crterios de excluso

CE.1: o artigo no se encontra disponvel eletronicamente na web;

CE.2: o artigo no est inteiramente escrito no idioma ingls;

CE.3: o artigo no apresenta os mtodos de estudo e tcnicas de classificao de dados e Reconhecimento de Padres utilizadas para conduo do trabalho;

CE.4: o artigo realiza estudos experimentais envolvendo aplicao de determinada(s) tcnica(s) de classificao de dados e Reconhecimento de Padres sem apresentar a contribuio do trabalho, bem como a anlise das principais vantagens e desvantagens das tcnicas utilizadas no trabalho;

CE.5: o artigo realiza uma proposta de uma nova implementao ou abordagem de uma tcnica de classificao de dados e Reconhecimento de Padres, bem como uma combinao destas, sem apresentar a contribuio do trabalho, bem como a anlise das principais vantagens e desvantagens das tcnicas utilizadas no trabalho;

CE.6: o artigo não se refere a um estudo primário, ou seja, apenas apresenta resultados de outros trabalhos, como, por exemplo, uma revisão sistemática ou bibliográfica.

3.3 Fonte de dados e estratégia de busca para seleção de estudos

Esta seção apresenta o detalhamento das fontes de dados utilizadas e estratégia de busca dos trabalhos para seleção de estudos. A busca dos estudos considerou como fonte de dados as seguintes bibliotecas digitais: (i) *IEEEExplore*; (ii) *Scopus*; e (iii) *Web of Science (Web of Knowledge)*. As fontes em questão foram escolhidas a partir de análise exploratória em relação ao tema geral de pesquisa, visto que são considerados repositórios de busca de trabalhos muito reconhecidos nesta área, aprendizagem de máquina. Adicionalmente, foram considerados somente trabalhos no idioma inglês, visto que é a principal língua para redação de trabalhos científicos conhecida e aceita internacionalmente.

A *string* de busca considerou as seguintes palavras-chave: “*classification algorithm*”, “*data classification*”, “*pattern recognition*”, “*machine learning*” e “*learning model*”. A tabela 1 apresenta a *string* de busca formada a partir de tais palavras-chave.

As palavras-chave foram definidas com base na análise exploratória efetuada a respeito do assunto e, também, com base nas questões de pesquisa, cujo objetivo é verificar de forma ampla as técnicas de aprendizagem de máquina aplicadas para classificação de dados relacionados a decisão de consumo, sem restrição a uma técnica ou modelo específico de aprendizagem de máquina, bem como não restrito ao tema de consumo. Inicialmente, havia sido considerada também a palavra “*consumption*” na *string* de busca combinada com a conjunção “*and*”, porém, não foram retornados resultados em todas as fontes de dados, e quando combinada com a conjunção “*or*”, retornou um volume elevado trabalhos relacionados ao tema, mas que não envolviam a área de aprendizagem de máquina. A partir disso, optou-se por retirá-la da *string*, visto que o objetivo deste estudo é, também, conhecer as diversas áreas em que tais algoritmos vêm sendo comumente aplicados.

Tabela 1 – *String* de busca genérica para condução da revisão sistemática

“*classification algorithm*” OR “*data classification*” OR “*pattern recognition*” OR (“*variable selection*” or “*machine learning*” or “*learning model*”)

3.4 Resultados gerais da revisão sistemática

Esta seção apresenta os resultados obtidos na condução desta revisão sistemática. Inicialmente, a *string* criada foi submetida a cada uma das máquinas de busca, sendo estas: (i) *IEEEExplore*; (ii) *Scopus*; e (iii) *Web of Science (Web of Knowledge)*, retornando 137, 46 e 86 trabalhos, respectivamente, o que totalizou em 269 trabalhos.

Todos os artigos foram submetidos ao processo de seleção de estudos, que consistiu na leitura do título, local de publicação, ano de publicação e resumo dos artigos, bem como aplicação dos critérios de inclusão e exclusão, o que resultou em 35 artigos selecionados para este estudo, cujas etapas se encontram detalhadas na figura 5.

Figura 5 – Passos seguidos no processo de seleção de estudos

Fonte: Nathalia Nieuwenhoff, 2017

A tabela 2 apresenta a relação dos 35 trabalhos selecionados. Os dados mostram que 19 trabalhos se referem a publicações em periódicos (*Journals & Magazines*) e 16 em conferências (*Conference Publications*), bem como tais trabalhos se referem a países e anos variados.

Tabela 2 – Relação dos trabalhos selecionados para revisão sistemática

Ref.	Título	Ano	País	Tipo
------	--------	-----	------	------

Ref.	Título	Ano	País	Tipo
A1	<i>Data classification with radial basis function networks based on a novel kernel density estimation algorithm</i>	2005	China	<i>Journals & Magazines</i>
A2	<i>Lidar detection of underwater objects using a neuro-SVM-based architecture</i>	2006	EUA	<i>Journals & Magazines</i>
A3	<i>Adaptive schemes applied to online SVM for BCI data classification</i>	2009	UK	<i>Conference Publications</i>
A4	<i>A Gait Recognition Method Based on KFDA and SVM</i>	2009	China	<i>Conference Publications</i>
A5	<i>Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction</i>	2004	EUA	<i>Journals & Magazines</i>
A6	<i>QP Representable Mahalanobis Ellipsoidal Learning Machine for Imbalanced Data Classification</i>	2012	China	<i>Conference Publications</i>
A7	<i>Determination of food age using neural network</i>	2014	Nigéria	<i>Conference Publications</i>
A8	<i>Semi-supervised method for gene expression data classification with Gaussian fields and harmonic functions</i>	2008	China	<i>Conference Publications</i>
A9	<i>Ensemble fusion methods for medical data classification</i>	2012	Polônia	<i>Conference Publications</i>
A10	<i>Clustering Support Vector Machines for unlabeled data classification</i>	2009	China	<i>Conference Publications</i>
A11	<i>A New Method Based on KFDA and SVM for Gait Identification</i>	2009	China	<i>Conference Publications</i>
A12	<i>A new fuzzy support vectors machine for biomedical data classification</i>	2008	Polônia	<i>Conference Publications</i>
A13	<i>Sleep Versus Wake Classification From Heart Rate Variability Using Computational Intelligence: Consideration of Rejection in Classification Models</i>	2008	Alemanha	<i>Journals & Magazines</i>
A14	<i>Object-oriented change detection for multi-source images using multi-feature fusion</i>	2016	China	<i>Conference Publications</i>
A15	<i>K-means clustering based SVM ensemble methods for imbalanced data problem</i>	2014	Coréia	<i>Conference Publications</i>
A16	<i>A Novel Classification-Rejection Sphere SVMs for Multi-class Classification Problems</i>	2007	China	<i>Conference Publications</i>
A17	<i>Fuzzy rule selection using Iterative Rule Learning for speech data classification</i>	2008	Cingapura	<i>Conference Publications</i>
A18	<i>Fuzzy Logic-Based Scenario Recognition from Video Sequences</i>	2013	Turquia	<i>Journals & Magazines</i>
A19	<i>Pattern Classification of Decomposed Wavelet Information using ART2 Networks for echoes Analysis</i>	2008	México	<i>Journals & Magazines</i>

Ref.	Título	Ano	País	Tipo
A20	<i>Damage detection in a benchmark structure using AR-ARX models and statistical pattern recognition</i>	2007	Brasil	<i>Journals & Magazines</i>
A21	<i>The use of artificial neural network in the classification of pulse-echo and TOFD ultrasonic signals</i>	2005	Brasil	<i>Journals & Magazines</i>
A22	<i>Parallel pattern classification utilizing GPU-based kernelized Slackmin algorithm</i>	2012	Grécia	<i>Journals & Magazines</i>
A23	<i>Object-Oriented Change Detection for Multi-Source Images Using Multi-Feature Fusion</i>	2016	China	<i>Conference Publications</i>
A24	<i>Multi-class LSTMSVM based on optimal directed acyclic graph and shuffled frog leaping algorithm</i>	2016	China	<i>Journals & Magazines</i>
A25	<i>Physical human activity recognition using wearable sensors</i>	2015	Suíça	<i>Journals & Magazines</i>
A26	<i>Gene selection for microarray data classification using a novel ant colony optimization</i>	2015	Irã	<i>Journals & Magazines</i>
A27	<i>Confidentiality based file attributes and data classification using TsF-KNN</i>	2015	Malásia	<i>Conference Publications</i>
A28	<i>Evaluation of three state-of-the-art classifiers for recognition of activities of daily living from smart home ambient data</i>	2015	Suíça	<i>Journals & Magazines</i>
A29	<i>An Extension of Fuzzy L-R Data Classification with Fuzzy OWA Distance</i>	2015	Turquia	<i>Journals & Magazines</i>
A30	<i>An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus</i>	2015	China	<i>Journals & Magazines</i>
A31	<i>Solving the imbalanced data classification problem with the particle swarm optimization based Support Vector Machine</i>	2014	Japão	<i>Journals & Magazines</i>
A32	<i>Vicinal support vector classifier using supervised kernel-based clustering</i>	2014	Cingapura	<i>Journals & Magazines</i>
A33	<i>Leaf species classification based on a botanical shape sub-classifier strategy</i>	2014	França	<i>Conference Publications</i>
A34	<i>Data weighting method on the basis of binary encoded output to solve multi-class pattern classification problems</i>	2013	Turquia	<i>Journals & Magazines</i>
A35	<i>A novel K-NN classification algorithm for privacy preserving in cloud computing</i>	2012	China	<i>Journals & Magazines</i>

Fonte: Nathalia Nieuwenhoff, 2017

Adicionalmente, a figura 6 apresenta o gráfico de distribuição dos trabalhos por ano. Tais dados mostram que os estudos foram publicados de forma distribuída entre os anos e,

adicionalmente, os anos de 2015, 2014 e 2008 apresentam maior concentração dos estudos publicados.

Figura 6 – Distribuição da quantidade de anos estudos da revisão sistemática por ano.

Fonte: Nathalia Nieuwenhoff, 2017

QP.1: *Quais tipos de técnicas, ou seja, algoritmos de aprendizagem de máquina estão sendo comumente aplicados para classificação de dados e Reconhecimento de Padrões relacionados à decisão de consumo?*

A tabela 3 apresenta o resumo dos resultados obtidos para responder esta pergunta. Nota-se que, dentre os 35 estudos, foram aplicados, basicamente, oito tipos de algoritmos para classificação de dados, bem como diversos trabalhos utilizaram mais de um algoritmo. Tais algoritmos se referem às técnicas de redes neurais, árvores de decisão (J48 e *Random Forest*), *Fuzzy c-means*, *K-means*, *k-Nearest Neighbor* (k-NN), *Naïve Bayes* (NB) e *Support Vector Machine* (SVM). Tais técnicas se referem a algoritmos de aprendizagem supervisionada e não supervisionada, sendo a maioria parte do primeiro grupo, visto que somente *Fuzzy c-means* e *K-means* se referem ao segundo, e as redes neurais que podem fazer parte de ambos.

Adicionalmente, é importante ressaltar que, ao analisar detalhadamente cada um dos estudos, notou-se que muitos deles utilizaram algumas variações de tais algoritmos relacionados, porém, para apresentação dos resultados optamos por apresentar somente o algoritmo ou a técnica geral.

Tais resultados mostram que o SVM e suas variações foi o algoritmo mais utilizado nos estudos escopo desta revisão, sendo utilizado em 20 estudos primários que correspondem

a 57% do total, seguido pelos algoritmos de redes neurais, utilizados em 26% dos trabalhos, e k-NN, em 20%.

Tabela 3 – Lista dos estudos selecionados para a revisão sistemática

Ref. Artigo	<i>Redes neurais</i>	<i>SVM</i>	<i>k-NN</i>	<i>Naïve Bayes</i>	<i>J48</i>	<i>Random Forest</i>	<i>K-means</i>	<i>Fuzzy c- means</i>
A1		X						
A2	X	X		X				
A3		X						
A4		X						
A5				X				
A6		X						
A7	X							
A8			X		X	X		
A9	X							
A10		X					X	
A11	X	X						
A12		X						
A13	X	X						
A14		X						
A15		X					X	
A16		X						
A17	X							
A18	X							
A19	X							
A20								X
A21	X							
A22		X						
A23		X						
A24		X						
A25		X	X			X	X	
A26		X			X			
A27			X					
A28		X		X		X		
A29			X					
A30			X					
A31		X						
A32		X						
A33						X		
A34			X					
A35			X					
Total	9	20	7	3	2	4	3	1

A tabela 4 apresenta a visão geral das áreas em que tais estudos foram aplicados, o que envolve os tipos de bases de dados utilizadas, bem como o contexto em que os problemas de classificação de dados foram estudados e analisados. Os dados demonstram que a maioria dos estudos foram realizados na área de medicina, e ao analisar tais trabalhos, foi notado que as aplicações se referem a variadas áreas, como, por exemplo, identificação de doenças, classificação e tratamento de pacientes e genética.

Nota-se também que sete estudos utilizaram bases de dados variadas a partir de repositórios de dados disponíveis para a área de aprendizagem de máquina. Ao analisar tais estudos, foi observado que dados variados foram utilizados para comparação dos algoritmos, não havendo um contexto específico para utilização das técnicas de Classificação de Dados, visto que o objetivo de tais trabalhos era validar a performance destes. Adicionalmente, nota-se que dentre os 35 estudos selecionados, nenhum deles foi aplicado ao contexto de decisão de consumo de produtos, o que inclui os bens de informação.

Tabela 4 – Lista das áreas de aplicação dos estudos selecionados na revisão sistemática

Área de aplicação	Quantidade de estudos
Medicina	10
Dados variados - Repositório (ex: UCI, BCI)	7
Física (Energia; Sons)	3
Vídeo, processamento de imagens ou gráficos	3
Alimentação	2
Educação	2
Geografia	2
Aeroespacial	1
Biologia	1
Biomedicina	1
<i>Cloud computing</i>	1
<i>Healthcare/Homecare</i>	1
Interação humano-computador	1
Segurança da informação	1
Telefonia	1
Total	35

Fonte: Nathalia Nieuwenhoff, 2017

QP.2: *Quais são as formas de aplicação das técnicas de aprendizagem de máquina para classificação de dados?*

A tabela 5 apresenta a divisão das três classificações gerais referentes às formas de aplicação das técnicas de aprendizagem de máquina para classificação de dados. Sendo estas:

1. Novas técnicas, abordagens ou implementação: Esta classificação se refere aos estudos que realizaram implementação de uma nova técnica ou abordagem, bem uma variação desta, podendo envolver a combinação de uma ou mais técnicas computacionais já existentes, utilização de variadas funções *kernel* para uma mesma técnica e a combinação de técnicas computacionais com técnicas estatísticas ou matemáticas, por exemplo.
2. Comparação entre técnicas: Esta classificação se refere aos estudos que realizaram a análise comparativa entre técnicas já existentes, não contemplando a modificação destas ou novas propostas de implantação.
3. Aplicação das técnicas: Esta classificação se refere aos estudos que realizaram a aplicação de uma ou mais técnicas combinadas já existentes, cuja contribuição está focada no contexto em que se aplicam, não contemplando a modificação destas ou novas propostas de implantação ou análise comparativa entre técnicas distintas.

Tabela 5 – Tabela de distribuição dos tipos de aplicação das técnicas de aprendizagem de máquina dos estudos primários da revisão sistemática

Ref. Artigo	Novas técnicas, abordagens ou implementação	Comparação entre técnicas	Aplicação das técnicas
A1	X		
A2	X		
A3		X	
A4	X		
A5	X		
A6		X	
A7	X		
A8			X
A9			X
A10		X	
A11	X		
A12	X		
A13	X		
A14		X	
A15	X		
A16	X		
A17	X		
A18			X
A19			X
A20			X
A21			X
A22	X		
A23	X		
A24	X		
A25		X	
A26		X	
A27	X		
A28		X	
A29			X
A30			X
A31	X		
A32	X		
A33	X		
A34			X
A35			X
Total	18	7	10

Fonte: Nathalia Nieuwenhoff, 2017

A partir disso, nota-se que a maioria dos estudos apresentaram novas técnicas, abordagens ou implementação destas como principal contribuição, totalizando em 18 estudos, o que representa 51% do total de estudos primários selecionados, seguido pelos estudos cuja principal contribuição é a aplicação das técnicas de aprendizagem de máquina para classificação de dados em contextos ou banco de dados específicos ou pouco explorados, representando 29% dos estudos selecionados, e os estudos referentes a comparação entre técnicas de aprendizagem de máquina para classificação de dados, representando os 20% restantes.

3.5 *Discussão dos resultados da revisão sistemática*

Esta revisão sistemática teve como objetivo principal a identificação das técnicas de aprendizado de máquina para classificação de dados e Reconhecimento de Padrões e, também, a observação dos tipos de experimentos relacionados à aplicação destes, preferencialmente, no contexto de classificação de dados de decisão de consumo.

A partir da revisão efetuada, inicialmente, observou-se que a utilização das técnicas de aprendizado de máquina para classificação de dados e Reconhecimento de Padrões não estão fortemente concentradas em anos e locais específicos, o que sugere ser um tema de pesquisa de interesse de vários países a partir dos últimos 15 anos. Após o mapeamento das técnicas utilizadas nos estudos, estas podendo ser de aprendizado supervisionado e não-supervisionado, resumem-se em oito tipos de técnicas para classificação de dados e Reconhecimento de Padrões, bem como diversos trabalhos utilizaram mais de uma delas. E, adicionalmente, ao analisar detalhadamente cada um dos estudos, notou-se que muitos deles utilizaram variações de tais algoritmos relacionados.

Observou-se que o SVM e suas variações foi o algoritmo mais utilizado nos estudos escopo desta revisão, sendo utilizado em 57% destes, seguido pelos algoritmos de redes neurais, utilizados em 26% dos trabalhos, e k-NN, em 20%. Porém, ao mesmo tempo, nota-se que não há uma concentração específica em um único algoritmo apenas, visto que os estudos apresentaram variadas combinações entre algoritmos e técnicas, bem como análises comparativas entre estas. Adicionalmente, notou-se na análise detalhada de tais estudos que o desempenho dos algoritmos e técnicas em relação à precisão das classificações de dados e custo computacional dependem diretamente do contexto em são aplicados, envolvendo as características das bases de dados utilizadas, como, por exemplo, se tratam de dados balanceados ou desbalanceado e a quantidade de registros e atributos, bem como as

combinações entre algoritmos e técnicas realizadas, o que apresentou resultados variados, não havendo uma conclusão específica sobre a superioridade de determinado algoritmo ou técnica ou determinada combinação entre eles.

A análise dos estudos possibilitou observar que a utilização das técnicas de aprendizagem de máquina pode ocorrer de formas distintas, estas podendo estar relacionadas a apresentação de novas técnicas, abordagens ou implementação destas como principal contribuição, que representaram 51% dos estudos selecionados, bem como podem ter como principal contribuição a aplicação das técnicas de aprendizagem de máquina para classificação de dados em contextos ou banco de dados específicos ou pouco explorados, que representaram 29% dos estudos selecionados, e, também, relacionadas a análise comparativa entre técnicas já existentes, o que representou os 20% restantes da relação de estudos analisados.

A análise dos estudos também sugere que tais técnicas podem ser aplicadas nos mais diversos contextos e áreas. Afinal, nota-se que, apesar da área de Medicina possuir a maior quantidade de estudos objeto desta revisão sistemática, os estudos efetuados não estão concentrados em uma única área específica, visto que tais artigos foram aplicados em áreas e contextos distintos, podendo ser saúde, física, alimentação, educação, computação, telefonia, entre outros, o que reforça a aplicabilidade da aprendizagem de máquina nos mais variados segmentos e situações-problema.

Por fim, a revisão efetuada indica que a utilização das técnicas de aprendizado de máquina para classificação de dados e Reconhecimento de Padrões para classificação de dados relacionados às decisões de consumo, mais especificamente, bens de informação, não é um tópico de pesquisa maduro e desenvolvido, visto que não foi abordado em nenhum dos trabalhos estudados.

4 Estudo experimental

Este trabalho realizou o estudo experimental da execução de dois algoritmos de classificação de dados, sendo estes: *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), com o objetivo de classificar os microdados contidos na Pesquisa de Orçamento Familiar (POF) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), disponível para download no próprio *Website* da instituição, de modo a identificar as características dos indivíduos consumidores e não consumidores de cinema no período investigado (IBGE, 2009). A POF foi realizada entre os anos de 2008 e 2009 e o indivíduo responde se consumiu cinema nos últimos 30 dias de sua entrevista pelo IBGE.

Mais especificamente, a execução de tais algoritmos visou ao treinamento e teste de tais modelos para classificação binária dos indivíduos em relação ao consumo de cinema, ou seja, se estes consumiram ou não consumiram cinema, a partir de um conjunto de atributos. Tais atributos são relacionados a fatores sociais, características pessoais, socioeconômicos e sócio demográficos tanto dos indivíduos como dos domicílios aos quais pertencem.

As seções seguintes apresentam o detalhamento do entendimento da base de dados utilizada, bem como dos materiais e métodos que suportaram a realização deste estudo experimental.

4.1 Banco de dados

O banco de dados utilizado neste trabalho será apresentado nesta seção. A base de dados para a aplicação das técnicas de Classificação de Dados foi a Pesquisa de Orçamento Familiar (POF) realizada durante os 12 meses entre os anos de 2008-2009 pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

De acordo com o Manual do Agente de Pesquisa da POF (2008-2009), as informações contidas na Pesquisa de Orçamento Familiar são utilizadas para “estudar a evolução dos hábitos de consumo das famílias a POF e atualizar as estruturas de ponderação, necessárias para a produção dos Índices de Preços ao Consumidor (índices, calculados e publicados mensalmente pelo IBGE, que indicam a variação média ocorrida nos preços do conjunto de produtos adquiridos e de serviços utilizados pela população)”.

A partir disso, os dados coletados pela POF foram obtidos por meio de entrevistas diretas nos domicílios particulares selecionados por amostragem, através de entrevistas com

os moradores durante um período de nove dias consecutivos, cuja duração é de 12 meses de coleta das informações no período de 19 de maio de 2008 até 18 de maio de 2009.

Dessa forma, a POF possibilita mensurar os gastos, recebimentos e poupança da população através dos indivíduos e das famílias, bem como demonstrá-los em relação às regiões do país, tipos de produtos e serviços, e como tais informações se apresentam durante o período de um ano, contemplando todas as suas épocas.

De acordo com o IBGE, a amostragem considerada na POF é em relação aos domicílios particulares permanentes. Em tais domicílios, identifica-se cada unidade básica da pesquisa, correspondendo a cada morador do domicílio que compartilha da mesma fonte de alimentação ou despesas com moradia. Os domínios para os quais se pode gerar resultados são: “Municípios das Capitais, situação urbana, e área urbana das Regiões Metropolitanas, estas últimas incluindo o Município da Capital”. Por fim, a POF 2008-2009 considerou como Regiões Metropolitanas as seguintes cidades: Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre. Adicionalmente, as amostras de domicílios foram selecionadas de forma aleatória e foram distribuídas entre os estados brasileiros, totalizando em 55.970 domicílios entrevistados.

De acordo com o *layout* da POF disponibilizado pelo IBGE, os dados da pesquisa se encontram estruturados através dos seguintes questionários:

- POF 1 – Características do domicílio e dos moradores, contendo informações de identificação do domicílio, suas características e composição, bem como as características dos moradores. A partir disso, de acordo com o *layout* da POF (2008-2009) disponibilizado pelo IBGE, a POF 1 é composta pelas seguintes tabelas: 01 – domicílio; 02 – pessoas; 03 – pessoas – imputação; e 10 – aluguel estimado.
- POF 2 – Questionário de aquisição coletiva, preenchido pela Pessoa de Referência e cada membro da Unidade de Consumo que tenha realizado qualquer aquisição de produto ou serviço, pesquisado no questionário. Sendo assim, de acordo com o *layout* da POF (2008-2009), a POF 2 é composta pelas seguintes tabelas: 05 – inventário de bens duráveis; 06 – despesas de 90 dias; 07 – despesas de 12 meses; 08 – outras despesas; e 09 – despesas com serviços domésticos.
- POF 3 – Caderneta de Aquisição coletiva, contendo registros referentes a aquisições de produtos alimentícios, higiene pessoal, uso doméstico, e demais produtos cuja aquisição seja frequente e consumidos por todos os moradores do domicílio, seu preenchimento ocorreu diariamente por cada Unidade de Consumo

do domicílio, durante sete dias consecutivos. De acordo com o *layout* da POF (2008-2009), a POF 3 é composta somente pela tabela 11 – caderneta de despesa.

- POF 4 – Questionário de aquisição individual, contendo as informações de aquisição de produtos e serviços geralmente de uso pessoal não abordadas nas pesquisas anteriores, podendo estes ser transportes, alimentação fora do domicílio, vestuário, lazer e cultura, veículos e etc., por cada Unidade de Consumo identificada como uma Unidade de Orçamento-Despesa na POF 1. De acordo com o *layout* da POF (2008-2009), a POF 4 é composta pelas seguintes tabelas: 12 – despesa individual e 13 – despesas com veículos.
- POF 5 – Questionário de trabalho e rendimento individual, contendo as informações de rendimento e vínculo empregatício em relação a cada a membro da Unidade de Consumo considerado como Unidade de Orçamento-Trabalho e/ou Rendimento. De acordo com o *layout* da POF (2008-2009), a POF 5 é composta pelas seguintes tabelas: 14 – rendimentos e deduções e 15 – outros rendimentos.
- POF 6 – Avaliação das condições de vida, contendo as informações relacionadas a opinião/avaliação do entrevistado sobre suas condições de vida, sendo respondido pela pessoa responsável pela administração da Unidade de Consumo. De acordo com o *layout* da POF (2008-2009), a POF 5 é composta pela tabela 04 – condições de vida.
- POF 7 – Bloco de consumo alimentar pessoal, contendo as informações de consumo de produtos alimentícios dos indivíduos moradores com 10 anos ou mais de idade, pertencentes à(s) Unidades(s) de Consumo existente(s) no domicílio. Adicionalmente, este questionário foi aplicado para uma subamostra dos domicílios presentes na amostra da POF. E de acordo com o *layout* da POF (2008-2009), a POF 7 é composta pela tabela 16 – consumo alimentar.

A partir do entendimento da disponibilização dos dados contidos na POF, notou-se que tais dados estão fragmentados em 16 tabelas e as informações estão dispostas entre elas de formas distintas, estas podendo ser, por exemplo, por Unidade de consumo, Unidade de Orçamento-Trabalho e/ou Rendimento e Unidade de Orçamento-Despesa.

Conforme detalhado no Manual do Agente de Pesquisa da POF (2008-2009), a POF considera como Unidade de consumo, a unidade composta por um morador ou um conjunto deles que compartilham aquisições alimentares, ou seja, possuem uma mesma fonte de alimentação, bem como despesas com moradia. Além disso, a POF considera como Unidade

de Orçamento das moradias, todos os indivíduos maiores de 10 anos, que são moradores presentes, subdividindo-se em: Unidade de Orçamento-Trabalho e/ou Rendimento, que se refere aos indivíduos que possuem qualquer tipo de rendimento ou executado qualquer tipo de trabalho durante o período de referência, sendo este de 12 meses, e Unidade de Orçamento-Despesa, que são os indivíduos que possuem hábitos de, bem como despesas para si ou para outros membros do domicílio, mesmo que não tenham fonte de renda própria.

Dessa forma, a partir do objetivo deste estudo referente a análise do consumo de cinema pelos indivíduos, a tabela 12 – despesa individual pertencente a POF 4 foi utilizada como base de dados origem para composição do banco de dados utilizado na execução dos algoritmos de classificação, visto que contém as informações relativas às despesas com cultura, o que inclui os filmes, por indivíduo. Após o entendimento dos campos e estrutura da tabela em questão, foram observados os seguintes aspectos:

- A tabela apresenta informações de consumo por Unidade de Orçamento-Despesa.
- Cada linha da tabela é composta por um ou mais itens consumidos, pois de acordo com o Manual do Agente de Pesquisa da POF (2008-2009), quando um tipo de produto ou serviço foi adquirido de formas distintas durante o período de referência, as informações referentes a este foram ser registradas em mais de uma linha. Dessa forma, notou-se a existência de mais de uma linha de registro de consumo por Unidade de Orçamento-Despesa, sendo que tais registros podem ser referentes a um mesmo item consumido por diferentes formas de pagamento, bem como em locais de aquisição distintos ou, também, por diferentes itens de consumo.
- A POF 4 é um questionário estruturado em 31 Quadros. A partir disso, foram analisadas as descrições dos quadros em questão e notou-se que o Quadro 28 contém as informações referentes ao consumo de cultura, conforme a sua descrição no Manual da POF: “Quadro 28: Diversões, Aquisição de Ingressos para Eventos Esportivos e Culturais, Diversões e Uso de Celulares no Período de Referência de 30 Dias”. Dessa forma, nota-se que apesar da POF ser uma pesquisa anual, o consumo de cultura tem seu período de referência de 30 dias.
- De acordo com a tabela “Cadastro de produtos da POF 2008-2009” da POF, disponível no site do IBGE, os produtos são identificados através do “QUADRO”, código de 2 dígitos que se refere ao quadro das tabelas da POF a que este pertence, “GRUPO”, campo composto por 3 dígitos gerados pela POF e 2 dígitos referente

ao sequencial do código, bem como pelo “CÓDIGO”, 2 dígitos referentes aos códigos da caderneta de despesas coletivas presentes na POF 3. Dessa forma, identificou-se através do campo de descrição do produto na tabela em questão, denominado “PRODUTO”, que o produto “CINEMA (INGRESSO)” é identificado através do valor dos campos numéricos: “QUADRO” igual a 28, “GRUPO” em branco e “CÓDIGO” igual a “00101”.

- Ao analisar a tabela “Cadastro de local de aquisição da POF 2008-2009”, disponível no site do IBGE, contendo o detalhamento dos locais de aquisição dos produtos, que são identificados através de um campo numérico “CÓDIGO”, composto por 3 dígitos gerados pela POF e 2 dígitos referente ao sequencial do código, considerou-se neste estudo que as Unidade de Orçamento-Despesa que consumiram o produto “CINEMA (INGRESSO)” somente serão classificadas na classe “Filme”, ou seja, consumiu cinema, se consumiram o produto em questão locais detalhados na tabela 6.

Tabela 6 – Locais de aquisição da POF (2008-2009) considerados para o consumo de cinema no estudo experimental

Código	Descrição do local de aquisição
00101	SUPERMERCADO
09504	CINEMA
04801	SHOPPING CENTER
09501	ESTABELECIMENTO CULTURAL
09706	CASA DE SHOWS E ESPETACULOS
01401	LOJA DE DEPARTAMENTO
10501	LOCADORA DE CD, VIDEO GAME, DVD E VIDEOKE
11001	LOJA DE APARELHOS TELEFONICOS

Fonte: IBGE - Documento “Cadastro de produtos da POF 2008-2009”, 2017

Dessa forma, as demais Unidades de Orçamento-Despesa que consumiram o produto “CINEMA (INGRESSO)” em outros locais, sendo estes considerados neste estudo como não oficiais ou indicadores de pirataria, bem como as que não consumiram o produto em questão, serão classificadas como “Não Filme”, ou seja, não consumiram cinema.

4.2 *Software utilizado no estudo experimental*

Os programas utilizados no estudo experimental deste trabalho serão apresentados nesta seção. As etapas de pré-processamento e processamento dos algoritmos de classificação foram suportadas por dois programas, Stata e Weka, respectivamente.

4.2.1 Stata

As etapas de limpeza, integração, seleção, e tratamento do banco de dados objeto deste trabalho foram realizadas no *software* Stata, *Data Analysis and Statistical Software*. O Stata se trata de um pacote de *software* estatístico sob licença da StataCorp LP, disponível para download através do website (<http://www.stata.com>), cujo acesso é pago, provendo diversas funcionalidades de análise e gerenciamento de dados, gráficos e variadas formas de visualização dos dados, além de possuir uma interface de fácil utilização para tais atividades. A versão do *software* utilizada neste trabalho foi a Stata SE/12.0.

4.2.2 Weka

A execução dos algoritmos de Classificação de Dados escopo deste trabalho, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), efetuada durante a etapa de Mineração de Dados, foi realizada utilizando o *software* Weka. O Weka, cujo nome vem de *Waikato Environment for Knowledge Analysis*, trata-se de um *software* de código aberto, *open source*, licenciado pela *General Public License (GNU)* e desenvolvido em linguagem Java na Universidade de Waikato na Nova Zelândia. Esta ferramenta se trata de uma coleção de algoritmos de aprendizado de máquina para aplicação em tarefas de Mineração de Dados. Estes algoritmos podem ser executados através de interface gráfica ou através de linha de comandos executados em código Java.

Além disso, o Weka é composto de funcionalidades que possibilitam tarefas de aprendizado de máquina como pré-processamento, classificação, regressão, *clustering*, associação de regras, além de funcionalidades para visualização dos resultados através de diferentes tipos de gráficos, tabelas e relatórios. A figura 7 apresenta a interface gráfica do Weka Explorer após a importação do banco de dados utilizado neste estudo, apresentando tais funcionalidades do *software* segregadas por aba.

Nota-se, por exemplo, que a importação do arquivo é efetuada na seção de pré-processamento, aba *Preprocess*, que também possibilita o tratamento da base, execução de filtros, bem como a visualização gráfica e quantitativa dos registros por atributo, e a execução dos algoritmos de classificação ocorre na aba *Classify*.

O *software* está disponível para *download* no Website desenvolvido para suporte ao sistema (www.cs.waikato.ac.nz/ml/weka), além disso, seus desenvolvedores disponibilizam uma série de cursos online a respeito do tema de aprendizagem de máquina, suporte a utilização do *software*, bem como exemplos de implementações dos métodos e algoritmos que este possui.

O Weka possui um formato próprio de arquivo, o ARFF, porém, o *software* também é capaz de ler arquivos no formato CSV (*Comma Separated Value*). Dessa forma, não será necessário converter o banco de dados deste trabalho para utilização do Weka.

Figura 7 – Tela da aba de pré-processamento do Weka após importação do banco de dados

Fonte: *Software* Stata, 2017

O Weka foi escolhido para este estudo por apresentar diversas funcionalidades que possibilitam a mineração de dados e transformação de conhecimento e, também, por ser

comumente utilizado em estudos desta natureza, bem como por possibilitar a execução de ambos algoritmos escopo deste trabalho, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM). A versão do Weka utilizada neste trabalho foi a 3.8.1.

Witten, Eike e Hall (2011) foi utilizado como principal referência para a etapa de Mineração de Dados deste estudo, visto que apresentam de forma detalhada os métodos de aprendizagem computacional, funcionamento do *software* Weka, bem como traz detalhamento prático sobre a implementação e execução dos algoritmos, além do detalhamento teórico dos métodos disponíveis no *software* e vastas discussões relacionadas sobre temas e situações relacionados ao como, por quê e quando utilizá-los.

4.3 *Aplicação das técnicas de KDD*

O detalhamento da aplicação das técnicas de Descoberta do Conhecimento (KDD - *Knowledge Discovery in Database*) conduzida neste experimento será detalhado nesta seção. Após o entendimento dos dados contidos da base de dados em questão, a partir de seu *layout*, descritivo de tabelas e relacionamento entre elas, conforme detalhado na seção 4.1, foram aplicadas as etapas previstas no processo de KDD, detalhadas a seguir.

4.4 *Limpeza e integração dos dados*

O apêndice A contém o detalhamento das etapas de limpeza e integração dos dados conduzidas neste estudo experimental de forma conjunta. A limpeza e integração dos dados foi realizada com o intuito de consolidar as tabelas que compõem os registros da POF 2008-2009 para composição do banco de dados contendo os registros e atributos relevantes para a condução deste estudo experimental.

4.5 *Seleção dos dados*

Esta seção contém o detalhamento da seleção dos dados conduzida neste estudo. Nesta etapa, foram selecionados os atributos relevantes para este trabalho após a análise das variáveis presentes no banco de dados resultante da etapa anterior de integração dos dados, a partir de aspectos gerais da teoria econômica.

Jehle e Reny (2000) discutem, a partir da Teoria do Consumidor, que o consumo é gerado por meio de uma tomada de decisão do consumidor, que, de forma geral, possui como

dependência fatores como suas preferências individuais, orçamento, demanda e a oferta do produto. Na teoria do consumidor, as escolhas de consumo são vistas a partir da capacidade de pagamento do consumidor que se encontra diretamente relacionada a sua restrição orçamentária, gerando, assim, a necessidade de se fazer escolhas. A partir disso, as características pessoais dos indivíduos bem como influências psicológicas, sociais e culturais afetam tal tomada de decisão.

Diniz (2010) em seu trabalho de análise do consumo de bens e serviços culturais no Brasil metropolitano, com base nos dados da POF 2002-2003, considerando a abordagem microeconômica do capital humano, em que o consumo de cultura é fortemente determinado pela exposição prévia dos indivíduos a tais bens e serviços, identificou algumas características individuais que influenciam esta exposição. A partir disso, concluiu-se que o consumo cultural, o que inclui bens de informação, como filmes, é determinado por variáveis socioeconômicas, educacionais e sócio demográficas do chefe do domicílio, além das variáveis que indicam a localidade e região dos domicílios.

Os resultados do trabalho em questão sugerem que o consumo cultural no Brasil é fortemente determinado pela renda e grau de educação do consumidor, concentrando-se de forma mais expressiva no rendimento domiciliar. Além disso, apresentou variações de acordo com as regiões do país, o que levou-se a associar tal consumo com as variações de oferta e/ou a distinções histórico-culturais de cada região, bem como que algumas características particulares do chefe do domicílio e, ao mesmo tempo, dos grupos que influenciam o consumo, estas podendo ser, a idade, gênero, raça/etnia e religião, afinal, podem representar influências dos grupos e indicação de pertencimento a estes, bem como as experiências pessoais de cada indivíduo.

Gazley, Clark e Sinha (2011), em seu trabalho de análise das preferências dos indivíduos consumidores de filmes a partir de uma abordagem de análise de fatores para mapear os principais atributos que influenciam a escolha dos consumidores, apresentam como resultado atributos relacionados a preço, gênero do consumidor, país de origem, círculo social, como por exemplo amigos, crítica e estratégia promocional das produtoras de filme.

Adicionalmente, Davis (2006), ao desenvolver um modelo econométrico da demanda de venda de produtos que possuem um local específico, bem como seus consumidores tenham preferências em relação a proximidade geográfica e outras características do produto e loja em que este é fornecido, aplicou tal modelo em uma base de dados de consumo de cinema americano e os resultados sugerem que a proximidade de localização do produto, bem como

seu preço, sua capacidade de substituição, ou seja, alternativas de consumo, e seus distribuidores afetam a decisão consumo.

Com base nas informações apresentadas, este estudo considerou para a seleção dos dados as informações dos indivíduos relacionadas aos domicílios a que pertencem, bem como atributos individuais, referentes às variáveis socioeconômicas, educacionais, sócio demográficas e suas características individuais.

Dessa forma, considerando que a escolha do consumidor depende de suas preferências, foram selecionadas variáveis referentes aos atributos pessoais dos indivíduos como sexo, idade, religião, cor (raça), se possui plano de saúde e seu nível de educação, através de variáveis como grau de escolaridade (nível de instrução), anos de estudo, se é estudante, se sabe ler e escrever e quantidade de horas habitualmente trabalhadas por semana.

Já considerando os determinantes socioeconômicos do indivíduo, considerando o domicílio a que pertence, foram selecionadas neste estudo variáveis que sugerem tais fatores, como tipo de domicílio (casa própria ou alugada), quantidade de banheiros, quantidade de cômodos, quantidade de famílias, quantidade de moradores, quantidade de unidades de consumo, bem como atributo de avaliação da família sobre como seu rendimento total permite levar a vida até o fim do mês.

Adicionalmente, para fatores de oferta do produto, foram considerados os atributos relacionados aos aspectos sócio demográficos dos indivíduos, sendo estes o fator de expansão, que indica os pesos dos indivíduos na amostra da pesquisa, estados em que estão localizados e tipos de regiões que residem, como por exemplo, zona rural e urbana.

Por fim, como não se têm informações relativas aos preços dos produtos consumidos, mas estes são determinados por fatores como renda, foram considerados atributos dos indivíduos, como renda per capita, renda bruta monetária e se possui vínculo empregatício e é considerado uma Unidade de Orçamento Trabalho e/ou Rendimento, bem como sua despesa total com cultura, como atributos indiretos de preço. Adicionalmente, as variáveis relacionadas a idade, bem como se está estudando ou não, já apresentadas anteriormente, podem também ser consideradas indicadores de preço, visto que para o produto cinema, objeto deste estudo, estes fatores impactam o preço dos ingressos.

A partir disso, julgamos que os demais atributos são desnecessários para este estudo, sendo assim, todos estes excluídos do banco de dados através do comando *drop* do Stata.

4.6 Transformação dos dados

O apêndice B contém o detalhamento da etapa de transformação, ou seja, modificação dos dados. A transformação dos dados foi realizada com o intuito de tornar os resultados obtidos pelos classificadores dotados de significados do fenômeno investigado por esta pesquisa.

4.7 Mineração de Dados

A etapa de Mineração de Dados, que consiste na execução dos dois algoritmos de aprendizagem de máquina para Classificação de Dados, NB e SVM, foi conduzida através da ferramenta Weka. A partir disso, importou-se o banco de dados obtido na etapa anterior para o *software* em questão. O Weka disponibiliza funcionalidades de visualização gráfica dos dados e categorias de cada atributo do banco de dados na seção de pré-processamento do *software*, aba *Preprocess*, conforme apresentado na figura 7.

Inicialmente, apesar de o Weka disponibilizar funcionalidades de pré-processamento dos dados para conversão e tratamento destes, bem como aplicação de filtros, estas não seriam utilizadas, visto que esta etapa foi efetuada no *software* Stata. A partir disso, para a execução dos algoritmos e condução da etapa de Mineração de Dados, foi utilizada a funcionalidade (aba) *Classify* do Weka, que possibilita a escolha dos algoritmos de classificação, bem como a definição dos seus parâmetros e a visualização dos resultados de execução, denominados pelo Weka como *output*.

4.7.1 Classes desbalanceadas

Esta seção apresentará uma visão geral sobre classes desbalanceadas, bem como a opção considerada neste estudo para condução da etapa de Mineração de Dados devido a existência de tal desbalanceamento entre as classes.

Ao visualizar graficamente a disposição das classes do atributo binário de classificação dos registros, criado na etapa anterior, conforme apresentado na figura 8, notou-se que se tratam de classes desbalanceadas, ou seja, uma das classes possui uma quantidade de registros muito superior a quantidade da outra, afinal, observou-se que, do total de 128300 registros, 124613 pertencem a classe “não consumiu cinema” (Não filme), o que corresponde

a aproximadamente 97% dos registros, e somente 3687 registros pertencem a classe “consumiu cinema” (Filme), correspondendo aproximadamente aos 3% restantes.

Figura 8 – Funcionalidade de pré-processamento do banco de dados do Weka Explorer, aba “Preprocess”

Fonte: Software Weka, 2017

Gadi (2008) afirma que o desbalanceamento entre as classes dificulta o processo de modelagem, visto que testes estatísticos clássicos podem indicar boa aderência e performance do modelo de classificação dos registros, o que não diretamente significa que os modelos foram capazes de prever os eventos de todas as classes, podendo haver situações que somente a classe predominante apresentou boa acurácia na classificação, o que não será visível na maioria das métricas finais de acuracidade. A partir disso, uma de suas sugestões para este tipo de cenário é a utilização do classificador *Cost Sensitive* para execução dos algoritmos, também disponível no Weka, como uma das alternativas para tais situações em que se têm classes desbalanceadas.

O classificador *Cost Sensitive* é sugerido, pois segundo Gadi (2008), utiliza-se da Meta Heurística para construção e teste dos modelos, trata-se de um algoritmo que possui a

capacidade de execução de transformação dos dados de entrada, executando os métodos de classificação por uma ou mais vezes, o que gerará diversas saídas, mas estas são apresentadas de forma compilada através de uma única saída final. Uma Meta Heurística no Weka, segundo Witten, Eike e Hall (2011), refere-se ao algoritmo que possa ser executado por todos os métodos de classificação sem restrições, mantendo suas respectivas entradas e saídas padronizadas do Weka.

Mais especificamente, o *Cost Sensitive*, segundo Witten, Eike e Hall (2011), consiste em um algoritmo de classificação capaz de balancear a quantidade de eventos de cada classe, através da replicação de registros das classes menos frequentes durante a sua etapa de construção do modelo, eliminando, assim, o desbalanceamento entre as classes, bem como possui a capacidade de gerar uma matriz de custo para ponderação do modelo em construção, minimizando, ao invés da classe mais provável, o seu custo global. Dessa forma, para ser executado no Weka, requer a sua associação a um algoritmo de classificação.

4.7.2 Seleção adicional de atributos por meio de algoritmos de seleção de variáveis

Durante a condução deste estudo, também, surgiu o interesse em verificar se um ou mais atributos, dentre os já pré-selecionados tendo como base a teoria econômica, possuem relevância predominante para a tomada de decisão de consumo de filmes por parte dos indivíduos. Adicionalmente, Witten, Eike e Hall (2011) afirmam que a existência de um número elevado de atributos irrelevantes pode prejudicar o desempenho dos algoritmos de classificação tanto nos aspectos relacionados a acuracidade destes, ou seja, capacidade de prever corretamente os eventos, bem como seu tempo de execução, tornando o processo de execução lento.

Dessa forma, com o objetivo de validar a seleção prévia das variáveis conduzidas na etapa anterior, verificando se um ou mais dos atributos previamente selecionados não possuem relevância significativa sobre as classes, bem como verificar a variação do desempenho dos algoritmos de classificação, NB e SVM, optamos por utilizar a métrica de *Information Gain* para uma segunda seleção dos atributos, a ser conduzida de forma automatizada.

Segundo Rogati e Yang (2002), a métrica *Information Gain* ou Ganho de Informação efetua a seleção dos atributos que melhor separam uma classe da outra, eliminando atributos desnecessários e reduzindo o risco de ocorrência de *overfitting* na Classificação de Dados, ou seja, cenários em que se têm modelos altamente ajustáveis aos dados. Adicionalmente,

sugeriram esta como uma das melhores métricas para seleção de atributos nas situações em que considerar a relevância destes sobre a classe é relevante.

A métrica *Information Gain* se encontra disponível no Weka através do algoritmo de seleção de atributos *InfoGainAttributeEval*, na funcionalidade *SelectAttributes* do programa. Segundo Witten, Eike e Hall (2011), o *InfoGainAttributeEval* trata-se de um algoritmo de seleção de atributos que mensura a informação ganha por cada atributo em relação a classe, a partir do método *MDL-based discretization method*. Adicionalmente, o algoritmo gera como resultado uma lista ordenada dos atributos, visto que está associado ao método de busca denominado *Ranker* no Weka, a partir da relevância de cada um destes em relação ao classificador.

4.7.3 Execução dos algoritmos de classificação

A partir dos aspectos e cenários supracitados nas seções anteriores, optou-se por fragmentar a etapa de Mineração de Dados, considerando a execução de ambos os algoritmos em cada uma delas, NB e SVM, em quatro execuções. Sendo estas: (i) execução dos algoritmos de classificação para o banco de dados completo contendo os 51 atributos previamente selecionados, denominada execução 1; (ii) execução considerando o mesmo banco de dados, porém executando tais algoritmos através do algoritmo de classificação *Cost Sensitive*, denominada execução 2; (iii) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain*, mas não considerando o classificador *Cost Sensitive*, denominada execução 3; (iv) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain* e considerando o classificador *Cost Sensitive*, denominada execução 4.

No ambiente visual do Weka, notou-se que a funcionalidade de execução dos algoritmos de Classificação de Dados do Weka, aba *Classify*, apresenta uma série variada de opções de algoritmos, presentes na seção denominada *Classifier*, bem como apresenta as opções dos tipos de teste para execução do algoritmo de classificação, na seção denominada *Test Options*. A figura 9 apresenta a aba *Classify* do Weka, bem como tais seções mencionadas para configuração do classificador.

Figura 9 – Funcionalidade de execução dos algoritmos de classificação no Weka Explorer, aba “Classify”

Fonte: Software Weka, 2017

A partir disso, para condução da etapa de Mineração de Dados foram considerados os aspectos relacionados a definição dos algoritmos de classificação, bem como a definição dos tipos de teste, conforme detalhados a seguir. Para a escolha do tipo de teste a ser considerado nas execuções dos algoritmos de classificação, notou-se que o Weka disponibiliza as seguintes opções de teste, conforme apresentadas anteriormente na figura 8:

- *Use training set*, funcionalidade para utilização do mesmo conjunto de dados para treinamento e teste do modelo.
- *Supplied test set*, funcionalidade que possibilita para associar um conjunto de dados externo para teste do modelo, dados não pertencentes ao banco de dados utilizado para a sua construção.
- *Cross-validation*, funcionalidade que se refere a técnica de validação cruzada, possibilitando definir a quantidade de partes (*folds*) a ser utilizada. Witten, Eike e Hall (2011) afirmam que uma forma de mitigar os problemas causados por uma amostra específica utilizada para construção e treinamento do modelo de

classificação é repetir todo processo de treinamento e também teste do modelo várias vezes com diferentes amostras de dados aleatórias. A partir disso, é calculada a taxa de erro global do modelo, considerando a média entre as taxas de erro a cada execução. Para isso, utiliza-se a técnica de *Cross-Validation*, ou validação cruzada. Trata-se de uma técnica estatística comumente utilizada para comparação de modelos preditivos, sendo aplicada a partir da fragmentação da amostra de dados em k subconjuntos de dados igualmente distribuídos, também denominados *k-folds*. O processo é repetido até que todos os subconjuntos tenham sido utilizados como conjunto de teste e treinamento, k vezes. A partir disso, considerando, por exemplo, o valor de k igual a 10, o banco de dados é dividido de forma aleatória em 10 partes para a execução da validação cruzada, em que as classes são representadas considerando de forma aproximada as mesmas proporções do conjunto de dados completo. Com isso, as etapas de treinamento e teste do modelo são executadas no total de 10 vezes para cada um dos diferentes conjuntos gerados, conforme o diagrama de *10-fold Cross-validation* apresentado na figura 10, e, por fim, calcula-se a média das 10 estimativas de erro para cálculo da estimativa de erro global.

Figura 10 – Diagrama de 10-fold Cross-validation.

Fonte: Yudong Zhang, Shuihua Wang (2015)

- *Percentage split*, funcionalidade que possibilita a definição do percentual do banco de dados a ser utilizado para treinamento e teste do modelo.

Ambos classificadores escolhidos para este experimento, NB e SVM, tratam-se de algoritmos de aprendizado supervisionado. A partir disso, a etapa de treinamento, ou seja, construção do modelo, requer a utilização de dados “passados”, ou seja, já previamente classificados, para, assim, aprender e ser capaz de identificar padrões para classificação de dados “futuros”, ou seja, não conhecidos pelo modelo.

Por este motivo, Witten, Eike e Hall (2011) reforçam a importância de utilizar dados distintos nas etapas de treinamento (construção) e teste do modelo, visto que, caso isto não seja considerado, os indicadores de precisão e erro calculados na etapa de treinamento não podem ser considerados bons indicadores para execuções futuras dos algoritmos. Dessa forma, foi desconsiderado neste estudo a possibilidade de utilização do tipo de teste *use training set*. Adicionalmente, como foi utilizado neste experimento somente a base de dados da POF e não se tem dados adicionais, a possibilidade de utilização da funcionalidade *supplied test set* também é descartada.

Adicionalmente, pelo fato de não se ter uma conclusão teórica e experimental sobre qual tipo de teste proporciona melhor desempenho dos algoritmos de classificação, visto que esta definição pode variar conforme o banco de dados utilizado, bem como a técnica de classificação a ser executada, surgiu o interesse em analisar se tais opções de teste afetam de forma significativa a performance dos algoritmos deste estudo. Com isso, optamos por executar ambos os algoritmos, NB e SVM, considerando os seguintes tipos de testes:

1. ***Cross-Validation (10 Folds)***: Witten, Eike e Hall (2011) afirmam que o valor de 10 *folds* é o mais comumente utilizado, visto que testes extensivos em vários conjuntos de dados diferentes, com diferentes técnicas de aprendizagem, demonstraram que com 10 partes se obtém a melhor estimativa de erro, além de estudos teóricos que reforçam tal ideia. A partir disso, consideramos o valor de 10 partes (*folds*) para a utilização da técnica de *cross-validation* neste estudo.
2. ***Percentage split***: Consideramos a execução dos algoritmos a partir dos seguintes percentuais de fragmentação dos registros do banco de dados para treinamento do modelo e teste: (i) 75% do total de registro para treinamento e 25% para teste do modelo; e (ii) 50% do total de registros para treinamento e 50% para teste do modelo.

Adicionalmente, para execução dos algoritmos de classificação, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), foram considerados, basicamente, os parâmetros *default* do Weka em todas às quatro execuções objeto deste estudo, conforme detalhamento a seguir de cada uma das execuções, sendo estas: (i) execução dos algoritmos de classificação para o banco de dados completo contendo os 51 atributos previamente selecionados, denominada execução 1; (ii) execução considerando o mesmo banco de dados, porém executando tais algoritmos através do algoritmo de classificação *Cost Sensitive*, denominada execução 2; (iii) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain*, mas não considerando o classificador *Cost Sensitive*, denominada execução 3; (iv) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain* e considerando o classificador *Cost Sensitive*, denominada execução 4.

Execução 1: Banco de dados completo (51 atributos)

A primeira execução foi conduzida no *software* Weka considerando cada um dos dois algoritmos escopo deste estudo experimental, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM).

- *Naïve Bayes* (NB): para a execução do algoritmo de classificação NB no Weka, foi escolhida a classe *NaiveBayes* e foi considerado todos os parâmetros padrão (default) do algoritmo em questão do *software*. O classificador e suas propriedades estão apresentados na figura 11.

Figura 11 – Execução 1: Propriedades do classificador NB no Weka

Fonte: *Software Weka*, 2017

- *Support Vector Machine (SVM)*: Para a execução do algoritmo de classificação SVM no Weka, foi escolhida a classe SMO, implementada para treinamento do classificador SVM, a técnica de *Sequential Minimal Optimization (SMO)* de John Platt, (1998), detalhada anteriormente no Capítulo 2, considerando todos os parâmetros padrão (*default*) do algoritmo em questão no Weka. O classificador e suas propriedades estão apresentados na figura 12.

Figura 12 – Execução 1: Propriedades do classificador SVM no Weka

Fonte: Software Weka, 2017

Execução 2: Banco de dados após a seleção de variáveis por meio do algoritmo *Information Gain*

A segunda execução foi conduzida no *software* Weka realizando, primeiramente, a seleção adicional dos atributos através do algoritmo *Information Gain*. A partir disso, foi

efetuada a execução do algoritmo de seleção de variáveis por meio da funcionalidade *Select Attributes* considerando todos os parâmetros *default* do algoritmo *InfoGainAttributeEval* do Weka. O seletor de variáveis e suas propriedades estão apresentados na figura 13.

Figura 13 – Execução 2: Propriedades do seletor de variáveis *Information Gain* no Weka

Fonte: *Software Weka*, 2017

Adicionalmente, foi definido como método para treinamento e validação do algoritmo de seleção dos atributos o *Cross-validation 10 folds*, cujo detalhamento se encontra descrito na seção 4.7.3 do capítulo 4. O seletor de variáveis, seu método de busca *Ranker*, bem como o método *Cross-validation* para treinamento e execução do algoritmo seleção dos atributos estão apresentados na figura 14.

Figura 14 – Execução 2: Funcionalidade de seleção de variáveis do Weka, aba *Select attributes*

Fonte: Software Weka, 2017

A execução do algoritmo de seleção de variáveis *Information Gain* no software Weka, retornou como resultado o *ranking* dos 51 atributos do banco de dados. A partir disso, notou-se que os seguintes atributos possuem os maiores valores de correlação médio (*average merit*), visto que tal métrica é gerada a partir da média dos valores de correlação obtidos durante os 10 ciclos da validação cruzada (*Cross-validation 10 folds*), ou seja, foram elencados pelo seletor de variáveis como os mais relevantes para o classificador, sendo também os seis primeiros atributos elencados no *ranking* médio (*average rank*), também calculado a partir da média dos rankings obtidos durante os 10 ciclos da validação cruzada. A figura 15 apresenta o resultado parcial da execução do seletor de variáveis, contendo os primeiros atributos do ranking gerado.

Foi considerado como critério para a definição da quantidade de atributos relevantes todos os que possuem valores de *average merit* em duas casas decimais maiores que 0, totalizando, assim, nos seguintes 06 atributos, cujo detalhamento de cada um deles se encontra descrito no apêndice B:

1. Despesa total em Recreação e cultura.
2. Anos de estudos.
3. Renda per capita.

4. Renda bruta monetária.
5. Quantidade de banheiros do domicílio.
6. Se tem plano de saúde.

Figura 15 – Execução 2: Resultado parcial do ranking de atributos gerados pelo algoritmo Infomation Gain no Weka

Attribute selection output		
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===		
average merit	average rank	attribute
0.054 +- 0	1 +- 0	47 v_DI_8_tot
0.027 +- 0	2 +- 0	11 anos_de_estudo
0.023 +- 0	3.2 +- 0.4	46 dummy_renda_percapita
0.022 +- 0	3.8 +- 0.4	45 dummy_renda_bruta_monteria
0.013 +- 0	5.2 +- 0.4	7 qtd_banheiros
0.013 +- 0	5.8 +- 0.4	12 plano_saude
0.009 +- 0	7 +- 0	15 dummy_nivel_instr_bin
0.007 +- 0	8.5 +- 0.5	6 qtd_comodos_domc
0.007 +- 0	8.5 +- 0.5	8 idade_anos
0.005 +- 0	10 +- 0	16 dummy_estudante_bin
0.005 +- 0	11 +- 0	10 cod_sabe_ler
0.003 +- 0	12.4 +- 0.49	1 fator_expansao1
0.003 +- 0	12.6 +- 0.49	44 dummy_cond_precaria
0.003 +- 0	14.3 +- 0.46	51 dummy_estrato_geografico
0.003 +- 0	14.7 +- 0.46	14 dummy_cor_raca_bin
0.001 +- 0	16 +- 0	2 qtd_morador_domc
0.001 +- 0	17 +- 0	35 dummy_SP
0.001 +- 0	18 +- 0	13 num_horas_trabalhadas
0.001 +- 0	19.3 +- 0.46	34 dummy_RJ
0.001 +- 0	19.7 +- 0.46	23 dummy_TO
0 +- 0	21.4 +- 0.49	50 dummy_cod_sit_receita
0 +- 0	21.8 +- 0.75	4 qtd_familia
0 +- 0	23 +- 0.63	25 dummy_PI
0 +- 0	24.6 +- 0.92	26 dummy_CE
0 +- 0	25.1 +- 1.22	48 dummy_religiao
0 +- 0	25.9 +- 1.04	42 dummy_DF

Fonte: Software Weka, 2017

Posteriormente, após a execução da segunda seleção dos atributos, utilizou-se a funcionalidade de filtro do Weka, disponível na seção de pré-processamento, para considerar somente tais atributos do banco de dados, mantendo a quantidade de eventos (linhas). A partir disso, foram executados cada um dos algoritmos de classificação escopo deste estudo experimental, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), utilizando os mesmos parâmetros da execução 1 detalhados na seção anterior.

Execução 3: Banco de dados completo (51 atributos), considerando o algoritmo de classificação *Cost Sensitive*

A terceira execução foi conduzida no *software* Weka associando os dois algoritmos escopo deste estudo experimental, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), ao algoritmo *Cost Sensitive Classifier*, considerando o banco de dados completo de 51 atributos, conforme detalhamento a seguir:

- *Cost Sensitive* e *Naïve Bayes* (NB): Para a execução do algoritmo de classificação *Cost Sensitive* juntamente com o *Naïve Bayes* (NB) no Weka, foi escolhida a classe *CostSensitiveClassifier*, associada a ela o classificador *NaiveBayes* e foram considerados todos os parâmetros padrão (*default*) do algoritmo em questão definidos pelo Weka, com exceção do parâmetro referente a matriz de custo, cujo valor *default* que representa o total de classes é igual 1, gerando uma matriz no formato de 1 x 1, porém, como estamos trabalhando com 2 classes, tal valor foi alterado para 2. O classificador e suas propriedades estão apresentados na figura 16, sendo possível visualizar a associação do NB ao algoritmo *CostSensitive* através do campo *classifier*, bem como o novo valor da matriz de custo definido como 2 x 2, campo *costMatrix*. Adicionalmente, com o objetivo de minimizar o problema de classes desbalanceadas, a matriz em questão foi configurada para aumentar o custo dos eventos em uma relação 10:1, o que significa, segundo Witten, Eike e Hall (2011), que a partir de tal parâmetro, o classificador irá evitar eventos falsos negativos (FN), visto que cada um deles tem um custo equivalente a 10 falsos positivos (FP). Esta configuração foi definida, visto que os eventos FN estão associados a classe minoritária “Filme”, conforme matriz de confusão apresentada a seguir na tabela 8 da seção 4.8 do capítulo 4. Já o valor da relação de custo foi definido devido ao fato de os eventos pertencentes a classe “Não filme” representarem aproximadamente 97% dos registros, conforme detalhado na seção 4.8 do capítulo 4. A figura 17 apresenta a configuração da matriz custo no Weka.

Figura 16 – Execução 3: Propriedades do classificador *CostSensitive* associado ao NB no Weka

Fonte: Software Weka, 2017

Figura 17 – Execução 3: Parametrização da matriz de custo do *Cost Sensitive classifier* no Weka

Fonte: *Software Weka*, 2017

- *Cost Sensitive* e *Support Vector Machine* (SVM): Para a execução do algoritmo de classificação *Cost Sensitive* juntamente com o *Support Vector Machine* (SVM) no Weka, foi escolhida a classe *CostSensitiveClassifier*, associada a ela o classificador *SMO*, conforme apresentado na figura 18, e foram considerados todos os parâmetros padrão (*default*) do algoritmo em questão definidos pelo Weka, com exceção dos parâmetros referentes a matriz de custo, cujos parâmetros definidos são os mesmos descritos na subseção anterior para o algoritmo NB.

Figura 18 – Execução 3: Propriedades do classificador *Cost Sensitive* associado ao SVM no Weka

Fonte: *Software Weka*, 2017

Execução 4: Banco de dados após a seleção de variáveis por meio do algoritmo *Information Gain* considerando o classificador *Cost Sensitive*

A quarta execução foi conduzida no *software* Weka associando os dois algoritmos escopo deste estudo experimental, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), ao algoritmo *Cost Sensitive Classifier*, considerando o banco de dados reduzido a seis atributos após a execução do algoritmo de seleção de variáveis *Information Gain*, conforme descrição do processo e parâmetros já apresentados nesta seção, demoninados execução 2. Adicionalmente, a execução de tais algoritmos combinada ao *Cost Sensitive Classifier* seguiu o mesmo processo e parâmetros já apresentados anteriormente na subseção acima.

4.8 Avaliação ou Pós-processamento

Esta seção apresenta o detalhamento das métricas utilizadas para comparação e análise dos algoritmos a partir do problema de classificação binária apresentado. Para análise-comparativa dos algoritmos, foram consideradas as seguintes métricas calculadas automaticamente pelo *software* Weka após a execução dos algoritmos de classificação, conforme detalhadas por Witten, Eike e Hall (2011):

- Eventos classificados corretamente (*Correctly Classified Instances*), ou seja, valor percentual de eventos classificados corretamente pelo classificador (percentual de acerto).
- Eventos classificados incorretamente (*Incorrectly Classified Instances*), ou seja, valor percentual de eventos classificados incorretamente pelo classificador (percentual de erro).
- Tempo de construção do modelo (*Time taken to build model*), ou seja, total de tempo em segundos para construção do modelo de classificação.
- Tempo de teste do modelo (*Time taken to test model on test split*), ou seja, total de tempo em segundos para teste e validação do modelo de classificação.
- Matriz de confusão (*Confusion Matrix*), ou, também denominada, tabela de contingência, trata-se de uma tabulação cruzada entre a classe prevista pelo modelo e a classe verdadeira dos eventos. Como o problema de classificação objeto deste trabalho é binário, ou seja, possui apenas duas classes, “não consumiu cinema” (Não filme) e “consumiu cinema” (Filme), o resultado da matriz de

confusão é apresentado conforme a tabela 7, considerando “a” como a classe positiva e “b” como a classe negativa. Os verdadeiros positivos ou *true positives* (TP) e verdadeiros negativos ou *true negatives* (TN) se referem a quantidade de eventos classificados corretamente, e falsos positivos ou *false positive* (FP) e falsos negativos ou *false negative* (FN) se referem a quantidade de predições incorretas.

Tabela 7 – Exemplo de matriz de confusão gerada pelo programa Weka

Classe predita		
a	b	Classe real
TP	FN	a= Não filme
FP	TN	b= Filme

Fonte: Nathalia Nieuwenhoff, 2017

A partir da Matriz de confusão, são obtidas as seguintes métricas, que serão também escopo para análise dos algoritmos:

- *TP Rate (Recall) ou Sensitividade*, trata-se do percentual dos eventos classificados como verdadeiros positivos (TP) em relação ao total de eventos da classe real positiva, ou seja, TP+FN, conforme fórmula 1. Dessa forma, nota-se que se trata da métrica da capacidade do algoritmo em prever corretamente os eventos da classe positiva.
- *FP Rate*, trata-se do percentual dos eventos classificados como falsos positivos (FP) em relação ao total de eventos da classe da classe real negativa, ou seja, FP+TN, conforme fórmula 2. Dessa forma, nota-se que se trata da métrica de eventos erroneamente classificadas como positivos dentre toda a classe negativa real.
- *Precision ou Precisão*, trata-se da razão do número total de eventos classificados corretamente pelo número total de eventos pertencentes a classe predita positiva, conforme fórmula 3.
- *F-measure*, trata-se de uma média ponderada entre *Recall* e *Precisão*, conforme apresentado na fórmula 4.

$$TP\ Rate = TP/(TP+FN) \times 100\% \quad (1)$$

$$FP\ Rate = FP/(FP+TN) \times 100\% \quad (2)$$

$$Precision = TP/(TP+FP) \times 100\% \quad (3)$$

$$F\text{-measure} = (2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision}) = (2 \times \text{TP}) / (2 \times \text{TP} + \text{FP} + \text{FN}) \quad (4)$$

Adicionalmente, o Weka também efetua o cálculo de tais métricas por classe e efetua a inversão destas, ou seja, a classe positiva (“a”) passa a ser a negativa (“b”) e vice-versa. A tabela 8 apresenta a matriz de confusão gerada quando a classe “a” passa ser a negativa.

Tabela 8 – Exemplo de matriz de confusão “invertida” gerada pelo programa Weka

Classe predita		
a	b	Classe real
TN	FP	a= Não filme
FN	TP	b= Filme

Fonte: Nathalia Nieuwenhoff, 2017

Segundo Witten, Eike e Hall (2011), existem situações de classificação em que somente considerar os valores de precisão e taxas de erro do classificador não são a forma mais apropriada para a avaliação de modelos de classificação. Uma delas ocorre quando o número de eventos pertencentes a uma classe está muito desbalanceado em relação a outra, ou seja, conceito de classes desbalanceadas apresentado na seção anterior. Por exemplo, suponha a existência de dois classificadores em que um conjunto total de dados esteja classificado 99% em uma classe. O que pode acontecer é o classificador apresentar taxas de erro aproximadas a 1%, visto que foi capaz de prever os eventos da classe predominante, o que significaria em uma análise inicial que se trata de um classificador com elevado grau de precisão, no entanto, o modelo não foi capaz de classificar corretamente os exemplos da classe minoritária. Para este trabalho, este tipo de situação precisa ser considerado, visto que identificamos que tais classes se encontram desbalanceadas e a classe principal objeto deste estudo é a minoritária.

Dessa forma, Siegel e Castellan (1988) sugerem considerar as seguintes métricas, também calculadas automaticamente pelo Weka após a execução dos algoritmos de classificação, para análise comparativa dos classificadores quando se tem este tipo de situação de classes desbalanceadas:

- *Kappa statistic*, segundo Landis e Koch (1977), trata-se de uma medida do grau de concordância da predição com a verdadeira classe, além do que seria esperado somente pelo acaso, ou seja, é uma medida estatística baseada no número de eventos cujo o resultado do classificador é o mesmo da classe real. Sua medida máxima tem o valor 1, que significa total concordância, e valor 0 nenhuma concordância ou que se

trata de uma concordância aleatória, esperada pelo acaso. Adicionalmente, Landis e Koch (1977) reforçam que ter um valor Kappa maior que zero, não significa necessariamente uma elevada concordância, mas que existe alguma. A partir disso, sugeriram a seguinte interpretação apresentada na tabela 9.

Tabela 9 – Tabela de interpretação dos valores de concordância da estatística Kappa

<i>Values of Kappa</i>	<i>Interpretation</i>
<0	<i>No agreement</i>
0-0.19	<i>Poor agreement</i>
0.20-0.39	<i>Fair agreement</i>
0.40-0.59	<i>Moderate agreement</i>
0.60-0.79	<i>Substantial agreement</i>
0.80-1.00	<i>Almost perfect agreement</i>

Fonte: Landis e Koch, 1977

- *ROC Area* ou AUC (*Área under the curve*), segundo Witten, Eike e Hall (2011), numericamente considerada equivalente a estatística de Wilconxon, trata-se de uma medida de probabilidade de escolha aleatória de um evento classificado corretamente nos dados de teste, denominado pelo Weka como evento positivo, esteja acima de um evento negativo, ou seja, que foi classificado incorretamente também selecionado aleatoriamente, ambos baseados em um *ranking* gerado pelo classificador. A medida AUC se refere a uma porção do quadrado (área abaixo) da Curva ROC e, a partir disso, seus valores variam entre 0.0 e 1.0. Dessa forma, tem-se como melhor resultado a situação em que todos os eventos sejam apresentados acima dos negativos, obtendo-se valor AUC igual a 1.0. No pior caso, este valor é 0. No caso que este *ranking* seja essencialmente randômico, ou seja, trata-se de um classificador aleatório, o valor apresentado é 0.5, e se for muito menor do que isso, o classificador é considerado *anti-learning*, ou seja, incapaz de aprender. Adicionalmente, Prati, Batista e Monard (2008) afirmam que esta medida representa o valor escalar da Curva ROC, que se trata de uma representação bidimensional da performance de um classificador, a partir das taxas de verdadeiros positivos (*TP Rate*) e da taxa de falsos positivos (*FP Rate*), ambas calculadas a partir da Matriz de confusão gerada após a execução do algoritmo, cujo gráfico é representado por *FP Rate* no eixo das ordenadas (x) e *TP Rate* no eixo das abscissas (y). Dessa forma, cada algoritmo de classificação é representado por um ponto no espaço ROC. Porém, a performance do algoritmo pode ser representada não somente por um ponto, mas também por vários pontos, que, segundo Prati, Batista e

Monard (2008), estes são variados entre o valor mais restritivo até o mais extremo, sendo representado, assim, através de uma curva no espaço ROC, gerada a partir da ordenação dos registros utilizados para teste de acordo com a classificação efetuada pelo modelo, conhecida como Curva ROC. Prati, Batista e Monard (2008) acrescentam que a comparação dos algoritmos de classificação através da Curva ROC é dada considerando como ideal uma curva convexa e crescente. E caso as curvas dos algoritmos comparados possuam intersecção, considera-se a superioridade da faixa operacional que um possui em relação ao outro. Por fim, caso tais curvas não possuam intersecção, considera-se de melhor performance o algoritmo cuja curva esteja mais próxima do ponto (0.0,1.0), canto superior esquerdo do espaço. Witten, Eike e Hall (2011) acrescentam que a curva ROC é usualmente utilizada para comparação de algoritmos de classificação em situações que a distribuição da classe seja desbalanceada, e funcionam melhor nas situações de duas classes, motivos pelo quais reforçam a escolha desta métrica para análise dos resultados.

4.9 Visualização dos Resultados

O detalhamento da etapa de visualização dos resultados obtidos, bem como a discussão de tais resultados serão apresentados no capítulo 5.

5 Apresentação e análise dos resultados obtidos

Este trabalho alcançou resultados relacionados às quatro execuções efetuadas dos dois algoritmos de aprendizagem de máquina para classificação de dados estudados, *Naïve Bayes* (NB) e *Support Vector Machine* (SVM). Sendo estas: (i) execução dos algoritmos de classificação para o banco de dados completo contendo os 51 atributos previamente selecionados, denominada execução 1; (ii) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain*, mas não considerando o classificador *Cost Sensitive*, denominada execução 2; (iii) execução considerando o mesmo banco de dados, porém executando tais algoritmos através do algoritmo de classificação *Cost Sensitive*, denominada execução 3; (iv) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain* e considerando o classificador *Cost Sensitive*, denominada execução 4. As seções seguintes apresentam os resultados para cada uma das execuções.

5.1 Execução 1: Banco de dados completo (51 atributos)

Esta seção apresenta os resultados obtidos a partir da execução dos algoritmos de classificação, NB e SVM, para o banco de dados completo contendo os 51 atributos previamente selecionados, bem como utilizando os parâmetros *default* do Weka para execução.

A tabela 10 apresenta o resultado da execução dos algoritmos executados no que se refere ao percentual de eventos corretamente classificados (*Correctly Classified Instances*), casos cujo resultado predito pelo classificador é o mesmo do real, e percentual de eventos incorretamente classificados (*Incorrectly Classified Instances*), eventos cujo resultado predito pelo classificador não corresponde ao real. Tais percentuais são apresentados por algoritmo de classificação, bem como por diferentes formas de divisão da base para construção do modelo e validação (teste), sendo estes: 50% dos dados utilizados para construção do modelo e 50% utilizado para teste, 75% dos dados utilizados para construção do modelo e 25% utilizado para teste, e o método *Cross-validation*, considerando 10 *folds* (partições) da base de dados, conforme detalhadas na seção 4.7.3 do capítulo 4.

A partir disso, nota-se que o SVM apresentou este indicador de acuracidade superior ao NB em todas as formas de fragmentação da base para treinamento e teste supracitadas,

visto que seus percentuais de eventos corretamente classificados, sendo estes 97.1723%, 97.1567% e 97.1263%, respectivamente, são superiores ao NB, que apresentou 90.4100%, 90.4380%, 90.7023%, respectivamente, o que diretamente resulta em percentuais de eventos incorretamente classificados do SVM inferiores ao NB.

Observa-se também que os métodos distintos de repartição da base para construção dos modelos de classificação e dos testes apresentaram resultados muito semelhantes nestes percentuais, apresentando variações inferiores a 0,3%, afinal, a variação dos percentual de eventos corretamente classificados do NB entre os métodos 50% treinamento e 50% teste e 75% treinamento e 25% teste corresponde a 0,028% (90,4380% - 90,4100%), já entre os métodos 75% treinamento e 25% teste e *cross-validation* corresponde a 0,2643% (90,7023% - 90,4380%) e, por fim, entre os métodos *cross-validation* e 50% treinamento e 50% teste corresponde a 0,2923% (90,7023% - 90,4100%). Já para o classificador SVM, as variações são inferiores a 0,05%, visto que a variação entre os métodos 50% treinamento e 50% teste e 75% treinamento e 25% teste corresponde a 0,0156% (97,1723% - 97,1567%), já entre os métodos 75% treinamento e 25% teste e *cross-validation* corresponde a 0,0304% (97,1567% - 97,1263%) e, por fim, entre os métodos *cross-validation* e 50% treinamento e 50% teste corresponde a 0,046% (97,1723% - 97,1263%), o que também resulta diretamente em valores reduzidos das variações dos percentuais dos eventos incorretamente classificados entre os métodos de fragmentação dos dados para construção do modelo e teste.

Adicionalmente, nota-se nesta execução que o método 50% treinamento e 50% teste apresentou o maior percentual de eventos corretamente classificados do classificador SVM, seguido por 75% treinamento e 25% teste e *cross-validation*. Já para o classificador NB, os resultados se inverteram, o melhor percentual é apresentado para o *cross-validation*, seguido por 75% treinamento e 25% teste e 50% treinamento e 50% teste.

Tais resultados possibilitam, assim, sugerir que para esta base de dados e o problema de classificação proposto, a escolha de tais métodos em questão de repartição dos dados para construção do modelo e testes não é um fator que influencia de forma significativa o desempenho do classificador.

Tabela 10 – Execução 1: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste

	NB		SVM	
	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>
<i>Split da base para treinamento e teste</i>				

<i>50% training/50% test</i>	90.4100%	9.5900%	97.1723%	2.8277%
<i>75% training/25% test</i>	90.4380%	9.5620%	97.1567%	2.8433%
<i>Cross-validation (10 Folds)</i>	90.7023%	9.2977%	97.1263%	2.8737%

Fonte: Nathalia Nieuwenhoff, 2017

A tabela 11 apresenta os valores de tempo de execução dos algoritmos, sendo estes subdivididos em tempo de construção do modelo (*Time taken to build model*) e tempo de teste do modelo para repartição de dados de teste (*Time taken to test model on test split*). A partir disso, nota-se que o algoritmo NB apresentou tempo de construção do modelo significativamente inferior ao SVM em todos os métodos de repartição da base de dados. Sendo os tempos do NB, 1.12 segundos (50% treinamento e 50% teste), 2.39 segundos (75% treinamento e 25% teste) e 2.1 segundos (*cross-validation*), e do SVM, 1497.73 segundos (50% treinamento e 50% teste), 1499.46 segundos (75% treinamento e 25% teste) e 1480.87 segundos (*cross-validation*). Já para o tempo de teste dos modelos, o SVM apresentou tempos inferiores, mas tais diferenças são bem mais reduzidas, não atingindo 1 segundo completo. Adicionalmente, nota-se que o *cross-validation* não possui valores de teste do modelo apresentados, visto que conforme descrito na seção 4.7.3 do capítulo 4, para este método, a ideia de fragmentação da base entre construção e teste não é aplicável, visto que ele efetua a validação cruzada de todos os dados em ambas através de conjuntos de dados aleatórios, sendo assim utilizado o banco de dados por completo para treinamento e teste do modelo.

Tabela 11 – Execução 1: Tabela de tempo de execução dos algoritmos por tipo de teste

	NB		SVM	
<i>Split da base para treinamento e teste</i>	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>
<i>50% training/50% test</i>	1.12 segundos	2.08 segundos	1497.73 segundos	0.14 segundos
<i>75% training/25% test</i>	2.39 segundos	1.52 segundos	1499.46 segundos	0.49 segundos
<i>Cross-validation (10 folds)</i>	2.1 segundos	-	1480.87 segundos	-

Fonte: Nathalia Nieuwenhoff, 2017

Conforme detalhado na seção 4.8 do capítulo 4, quando se têm classes desbalanceadas, é importante considerar outras variáveis para comparação do desempenho em relação aos fatores de acuracidade dos algoritmos de classificação, conforme apresentadas na tabela 12. Optamos por apresentar tais variáveis de desempenho dos algoritmos utilizando apenas o método *cross-validation (10 folds)*, já que, conforme apresentado, a variação dos resultados gerais entre os métodos de divisão das bases para treinamento e teste apresentados anteriormente é irrelevante neste estudo.

A partir disso, nota-se que o SVM apresentou valores médios finais de *TP Rate*, *FP Rate* e *F-measure*, detalhados na linha *Weighted Avg*, sendo estes 0.971, 0.971, 0.957, respectivamente, superiores ao NB, em que são 0.917, 0.484, 0.930, respectivamente. Porém, nota-se que o valor de precisão, métrica *Precision*, do NB é 0.960, superior ao SVM, 0.943. A métrica de precisão do SVM inferior ao NB se explica pelo fato do SVM ter classificado mais eventos como falsos positivos que o NB.

Analisando estas mesmas medidas por classe, sendo estas, “Filme”, ou seja, se o indivíduo consumiu cinema, e “Nao filme”, se não consumiu cinema, observamos que os resultados de um algoritmo em relação ao outro apresentam variações. Considerando apenas a classe “Filme”, em que esta passa a ser a classe positiva (“a”) da matriz de confusão, conforme inversão da matriz anteriormente apresentada na tabela 8 da seção 4.8 do capítulo 4, nota-se que todas as métricas de acuracidade do SVM apresentaram valor igual a zero, o que indica que o algoritmo classificou todos os eventos desta classe incorretamente, além disso, os valores de tais métricas passam ser inferiores ao NB, que apresenta valores maiores que zero para todas.

Já para a classe “Nao filme”, a superioridade do SVM acaba sendo mantida, visto que todos os valores destas métricas são superiores ao NB, atingindo o valor máximo (1.0), exceto em relação a métrica de precisão (*Precision*). Estes resultados demonstram que o SVM foi capaz de prever corretamente todos os eventos da classe em questão, bem como ao se analisar a métrica de precisão, também tais resultados já apresentados são confirmados, afinal, devido ao fato de o SVM ter classificado todos os eventos da classe “Filme” incorretamente (FP), seu valor de precisão é impactado, tornando-se, assim, inferior ao NB nesta métrica em todas as análises, tanto no valor final médio quanto por classe.

As matrizes de confusão de ambos os classificadores, conforme apresentadas na tabela 13, reforçam tais resultados apresentados, afinal, nota-se na matriz do SVM que apesar de seu valor de falso negativo (FN) ser zero para a classe “Não filme”, o que significa todos os eventos foram classificados corretamente, verdadeiros positivos (TP), seu valor de

verdadeiros negativos (TN) para a classe “Filme” é também zero, o que significa que todos os eventos da classe foram classificados incorretamente (FP). Já o NB, apresenta 10099 eventos classificados incorretamente da classe “Não filme” (FN), e 1830 classificados incorretamente da classe “Filme” (FP), que corresponde a aproximadamente 50% dos eventos da classe em questão.

Tabela 12 – Execução 1: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka

Classe	NB				SVM			
	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>
Filme	0.504	0.081	0.155	0.237	0.000	0.000	0.000	0.000
Não Filme	0.919	0.496	0.984	0.950	1.000	1.000	0.971	0.985
<i>Weighted Avg.</i>	0.917	0.484	0.960	0.930	0.971	0.971	0.943	0.957

Fonte: Nathalia Nieuwenhoff, 2017

Tabela 13 – Execução 1: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação no Weka

NB			SVM		
a	b	<-- classified as	a	b	<-- classified as
114514	10099	a = naofilme	124613	0	a = naofilme
1830	1857	b = filme	3687	0	b = filme

Fonte: Nathalia Nieuwenhoff, 2017

Adicionalmente, os valores de estatística Kappa (*Kappa statistic*) e Área ROC, conforme apresentados na tabela 14, também, reforçam a superioridade de performance em relação às medidas de acuracidade do NB em relação ao SVM quando não se considera apenas o valor médio entre as classes, mas a capacidade de predição geral dos algoritmos para cada uma delas.

O NB apresenta valor de estatística Kappa igual a 0.2024, que conforme a Tabela sugerida por Landis e Koch (1977) apresentada na seção 4.8 do capítulo 4, corresponde a *Fair agreement*, e apesar de ser um resultado de concordância entre baixa e intermediária, ou seja, apesar do valor final de concordância da predição realizada pelo algoritmo com a verdadeira classe (classe real) ser fraca, ainda é superior ao do SVM, cujo valor é igual a zero, o que corresponde a *Poor agreement*, demonstrando nenhuma concordância ou que a concordância

do classificador foi exatamente a esperada pelo acaso. Este valor é zero devido ao fato de o SVM não ter sido capaz de prever algum evento da classe “Filme” corretamente.

Por fim, nota-se que a Área ROC, cujos valores variam entre 0.0 e 1.0, sendo 1.0 o melhor caso, conforme detalhado na seção 4.8 do capítulo 4, também apresenta valor superior no classificador NB em relação ao SVM, sendo estes, 0.865 e 0.500, respectivamente, o que indica que a probabilidade de escolha aleatória de um evento classificado corretamente nos dados de teste esteja acima de um evento classificado incorretamente também selecionado aleatoriamente é superior no algoritmo NB em relação ao SVM. Além disso, segundo Witten, Eike e Hall (2011), este valor da Área ROC obtido após a execução do classificador SVM igual a 0.5 indica que este comportou-se como um classificador aleatório.

As figuras 19 e 20 apresentam as Curvas ROC de ambos classificadores, NB e SVM, respectivamente, para a classe “Filme” geradas pela ferramenta Weka, notando-se, assim, que a curva do algoritmo NB, está mais próxima da ponto ideal (0.0,1.0), conforme apresentado na seção 4.8 do capítulo 4, sendo mais uma métrica de acuracidade que o algoritmo apresentou valor superior em relação ao SVM.

Tabela 14 – Execução 1: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka

NB		SVM	
<i>Kappa statistic</i>	<i>ROC Area</i>	<i>Kappa statistic</i>	<i>ROC Area</i>
0.2024	0.865	0	0.500

Fonte: Nathalia Nieuwenhoff, 2017

Figura 19 – Execução 1: Ilustração da curva ROC do NB gerada pelo Weka

Fonte: *Software Weka*, 2017

Figura 20 – Execução 1: Ilustração da curva ROC do SVM gerada pelo Weka

Fonte: *Software Weka*, 2017

Dessa forma, para esta execução, sugere-se que o NB apresentou desempenho superior ao SVM, pois apesar de as métricas referentes à performance média serem inferiores ao SVM, quando se observa tais métricas de acuracidade dos algoritmos por classe, os valores de tempos de execução, estatística Kappa e área e curva ROC, bem como o fato de não haver uma classe mais relevante que a outra para o problema de classificação estudado, nota-se que o *Naïve Bayes* apresentou desempenho mais equilibrado, sendo capaz de prever eventos de ambas as classes, afinal, o SVM não foi capaz de prever corretamente algum evento da classe “Filme”. Witten, Eike e Hall (2011) sinalizam que resultados similares a estes são obtidos quando se executam alguns algoritmos de Classificação de Dados, o que inclui o SVM, para banco de dados que possuem classes desbalanceadas e, com isso, este é apenas capaz de treinar a classe que contém a maioria dos registros, o que é confirmado pelos resultados obtidos neste experimento.

5.2 Execução 2: Banco de dados após a seleção de variáveis por meio do algoritmo *Information Gain*

Esta seção apresenta os resultados obtidos a partir da execução dos algoritmos de classificação, NB e SVM, para o banco de dados reduzido após a execução do algoritmo de seleção de atributos *Information Gain*, considerando apenas os seis primeiros atributos elencados por ele, bem como os parâmetros *default* do Weka para execução de cada um dos algoritmos de classificação.

A tabela 15 apresenta o resultado da execução de ambos algoritmos, NB e SVM, no que se refere ao percentual de eventos corretamente classificados (*Correctly Classified Instances*) e percentual de eventos incorretamente classificados (*Incorrectly Classified Instances*). Tais percentuais também são apresentados por algoritmo de classificação, bem como por diferentes formas de divisão da base para construção do modelo e validação (teste), sendo estes: 50% dos dados utilizados para construção do modelo e 50% utilizado para teste, 75% dos dados utilizados para construção do modelo e 25% utilizado para teste, e o método *Cross-validation*, já apresentado anteriormente, considerando *10 folds* (partições) da base de dados.

A partir disso, nota-se que os resultados para o algoritmo SVM não sofreram alterações em relação a execução 1, considerando o banco de dados completo, conforme detalhado na seção 6.1, visto que tais percentuais se mantiveram exatamente os mesmos. Já algoritmo NB apresentou os seguintes percentuais de eventos corretamente classificados:

92,9525%, 92,8698%, 93,0881%, respectivamente. Dessa forma, nota-se que tais percentuais do SVM são superiores ao NB, e o que diretamente resulta em percentuais de eventos incorretamente classificados do SVM inferiores ao NB.

Adicionalmente, nota-se também que as diferenças de acuracidade entre os métodos de repartição da base para construção dos modelos de classificação e dos testes permanecem apresentando resultados muito semelhantes conforme apresentados para a base completa, visto que para o algoritmo NB tais variações também são inferiores a 0,3%, afinal, a variação dos percentual de eventos corretamente classificados entre os métodos 50% treinamento e 50% teste e 75% treinamento e 25% teste, corresponde a 0,0827%, entre os métodos 75% treinamento e 25% teste e *cross-validation* corresponde a 0,1356% e, por fim, entre os métodos *cross-validation* e 50% treinamento e 50% teste corresponde a 0,2923%.

Já para o classificador SVM, as variações permanecem inferiores a 0,05%, visto que, conforme descrito anteriormente, o algoritmo não apresentou variações em tais percentuais em relação a execução do algoritmo para a base completa e, com isso, o método de repartição da base 50% treinamento e 50% teste continuou apresentando o maior percentual de eventos corretamente classificados, seguido por 75% treinamento e 25% teste e *cross-validation*. Porém, para o classificador NB, o melhor percentual apresentado continuou sendo o *cross-validation*, porém, neste caso, foi seguido pelo método 50% treinamento e 50% teste e 75% treinamento e 25% teste. Dessa forma, os resultados em questão para a base reduzida em relação à quantidade de atributos também sugerem que a escolha de tais métodos em questão de repartição dos dados para construção do modelo e testes não é um fator significativo para o desempenho dos classificadores, considerando esta base de dados e o problema de classificação proposto.

Tabela 15 – Execução 2: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste

<i>Split</i> da base para treinamento e teste	NB		SVM	
	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>
<i>50% training/50% test</i>	92.9525%	7.0475%	97.1723%	2.8277%
<i>75% training/25% test</i>	92.8698%	7.1302%	97.1567%	2.8433%
<i>Cross-validation (folds 10)</i>	93.0881%	6.9119%	97.1263%	2.8737%

Fonte: Nathalia Nieuwenhoff, 2017

A tabela 16 apresenta os valores de tempo de execução dos algoritmos considerando a seleção das variáveis, sendo estes subdivididos em tempo de construção do modelo (*Time taken to build model*) e tempo de teste do modelo para repartição de dados de teste (*Time taken to test model on test split*). A partir disso, nota-se que o algoritmo SVM apresentou significativa redução do tempo de execução para construção do modelo em relação a base completa, cujo valor médio entre os três métodos de fragmentação da base é 1492.69 segundos, e considerando a seleção das variáveis tais valores de tempo são 21.25 segundos (50% treinamento e 50% teste), 22.25 segundos (75% treinamento e 25% teste) e 18.45 segundos (*Cross-validation*).

O algoritmo NB também apresentou redução no tempo de construção do modelo, permanecendo mais rápido que o SVM em todos os métodos de repartição da base de dados para treinamento e teste abordados neste estudo. Em relação ao tempo de teste dos modelos, o SVM apresentou tempo inferior em relação ao NB, considerando a fragmentação da base 75% treinamento e 25% teste, sendo 0.06 segundos e 0.19 segundos, respectivamente, e o NB apresentou melhor tempo para a fragmentação da base em 50% treinamento e 50% teste, sendo 0.16 segundos e 0.36 segundos respectivamente, mas tais diferenças permanecem reduzidas, não atingindo 1 segundo completo.

Tabela 16 – Execução 2: Tabela de tempo de execução dos algoritmos por tipo de teste

<i>Split da base para treinamento e teste</i>	NB		SVM	
	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>
<i>50% training/50% test</i>	0.74	0.16	21.25	0.36
	segundos	segundos	segundos	segundos
<i>75% training/25% test</i>	0.28	0.19	22.25	0.06
	segundos	segundos	segundos	segundos
<i>Cross-validation (10 folds)</i>	0.43	-	18.54	-
	segundos		segundos	

Fonte: Nathalia Nieuwenhoff, 2017

Adicionalmente, considerando as métricas que contribuem para validação da acuracidade dos algoritmos nas situações de classes desbalanceadas, apresentadas na tabela 17, mostram-se os resultados de tais métricas considerando também apenas o método de fragmentação dos dados para treinamento e teste *cross-validation (10 folds)*, já que a variação

dos resultados gerais entre os métodos apresentados anteriormente permanecem irrelevantes para este estudo.

Nota-se que o SVM apresentou os mesmos valores de *TP Rate*, *FP Rate* e *F-measure* da execução 1, detalhada na seção 5.1 deste capítulo, conforme a linha *Weighted Avg*, sendo estes 0.971, 0.971, 0.957, respectivamente. E apesar do NB apresentar alterações positivas em tais valores, 0.931, 0.595, 0.943, respectivamente, visto que todos aumentaram, os do SVM permanecem superiores, exceto pela métrica de previsão que o NB apresentou valor superior. A métrica de precisão do SVM inferior ao NB se explica pelo fato de o SVM ter classificado mais eventos como falsos positivos que o NB.

Adicionalmente, ao analisar estas mesmas medidas por classe, sendo estas, “Filme” e “Nao filme”, nota-se que o SVM também apresentou os mesmos resultados obtidos na execução 1, conforme detalhada na seção 5.1 deste capítulo, o que significa que não foi capaz de prever corretamente nenhum evento da classe “Filme”. Dessa forma, os valores de tais métricas passam ser inferiores ao NB, que apresenta valores maiores que zero para todas.

Já para a classe “Nao filme”, a superioridade do SVM acaba sendo mantida em relação ao NB, visto que todos os valores destas métricas permanecem com o valor máximo (1.0), exceto em relação a métrica de precisão (*Precision*). Estes resultados demonstram que o SVM foi capaz de prever corretamente todos os eventos da classe em questão, bem como ao se analisar a métrica de precisão, também tais resultados já apresentados são confirmados, afinal, devido ao fato de o SVM ter classificado todos os eventos da classe “Filme” incorretamente (FP), seu valor de precisão é impactado, tornando-se, assim, inferior ao NB nesta métrica em todas as análises, tanto no valor final médio quanto por classe.

Ao analisar as matrizes de confusão de ambos os classificadores, conforme apresentadas na tabela 18, nota-se que o SVM novamente não apresentou alterações de desempenho em relação aos resultados da sua execução para a base completa na execução 1 descrita na seção 5.1 deste capítulo. Já o NB, apresenta 6614 eventos classificados incorretamente da classe “Não filme” (FN), e 1433 classificadas incorretamente da classe “Filme” (FP).

Por fim, para os valores de estatística Kappa (*Kappa statistic*) e Área ROC, apresentados na tabela 19, o SVM não apresenta variações, permanecendo, assim, com o valor de estatística Kappa igual a 0 e Área ROC igual 0.500. Já o NB apresentou valor de estatística Kappa é 0.2132, que permanece com o resultado de concordância *Fair agreement*, conforme tabela sugerida por Landis e Koch (1977) presente na seção 4.8 do capítulo 4. Já o seu valor da Área ROC é 0.851, permanecendo superior ao SVM. As figuras 21 e 22 abaixo apresentam

as Curvas ROC de ambos classificadores, NB e SVM, respectivamente, para a classe “Filme” geradas pela ferramenta WEKA, notando-se, assim, que a curva do algoritmo NB, ainda continua mais próxima da ponto ideal (0.0,1.0), indicando acuracidade superior do algoritmo para a classe em questão, em relação ao SVM, que não sofreu alterações.

Tabela 17 – Execução 2: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka

Classe	NB				SVM			
	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>
Filme	0.389	0.053	0.178	0.244	0.000	0.000	0.000	0.000
Não Filme	0.947	0.611	0.981	0.964	1.000	1.000	0.971	0.985
<i>Weighted Avg.</i>	0.931	0.595	0.958	0.943	0.971	0.971	0.943	0.957

Fonte: Nathalia Nieuwenhoff, 2017

Tabela 18 – Execução 2: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação

NB			SVM		
a	b	<-- classified as	a	b	<-- classified as
117999	6614	a = naofilme	124613	0	a = naofilme
2254	1433	b = filme	3687	0	b = filme

Fonte: Nathalia Nieuwenhoff, 2017

Tabela 19 – Execução 2: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka

NB		SVM	
<i>Kappa statistic</i>	<i>ROC Area</i>	<i>Kappa statistic</i>	<i>ROC Area</i>
0.2132	0.851	0	0.500

Fonte: Nathalia Nieuwenhoff, 2017

Figura 21 – Execução 2: Ilustração da curva ROC do NB gerada pelo Weka

Fonte: Nathalia Nieuwenhoff, 2017

Figura 22 – Execução 2: Ilustração da curva ROC do SVM gerada pelo Weka

Fonte: Nathalia Nieuwenhoff, 2017

Dessa forma, para esta execução, nota-se que o NB permaneceu apresentando desempenho superior ao SVM, pois apesar de as métricas referentes a performance final continuarem serem inferiores ao SVM, conforme também identificadas na Execução 1, quando tais métricas de acuracidade dos algoritmos são observadas por classe, os valores de tempos de execução, estatística Kappa e área e curva ROC, bem como o fato de não haver uma classe mais relevante que a outra para o problema de classificação estudado, nota-se que o *Naïve Bayes* apresentou desempenho mais equilibrado, sendo capaz de prever eventos de ambas as classes, afinal, o SVM não foi capaz de prever corretamente algum evento da classe “Filme”. Conforme apresentação na seção 5.1 deste capítulo, sugere-se que a incapacidade do SVM de predição de eventos da classe “Filme” é dada devido ao desbalanceamento das classes.

5.3 Execução 3: Banco de dados completo (51 atributos), considerando o algoritmo de classificação *Cost Sensitive*

Esta seção apresenta os resultados obtidos a partir da execução dos algoritmos de classificação, NB e SVM, para o banco de dados completo contendo os 51 atributos previamente selecionados utilizando os parâmetros *default* do Weka, porém executando tais algoritmos combinados com o algoritmo de classificação *Cost Sensitive*, conforme apresentado na seção 4.7.1 do capítulo 4. Optamos por efetuar esta execução utilizando apenas o método *cross-validation (10 folds)*, já que, conforme resultados obtidos nas execuções 1 e 2, apresentadas nas seções anteriores 5.1 e 5.2, respectivamente, a variação dos resultados gerais entre os métodos de divisão das bases utilizados possuem valores muito pequenos, considerados irrelevantes neste estudo.

A tabela 20 apresenta o resultado da execução de ambos algoritmos executados a partir do classificador *Cost Sensitive* no que se refere ao percentual de eventos corretamente classificados (*Correctly Classified Instances*) e percentual de eventos incorretamente classificados (*Incorrectly Classified Instances*). A partir disso, nota-se que o SVM apresentou este indicador de acuracidade superior ao NB, visto que seu percentual de eventos corretamente classificados corresponde a 93.2214 %, enquanto o NB apresentou 86.3352 %, o que diretamente resulta em percentuais de eventos incorretamente classificados do SVM inferiores ao NB.

Tabela 20 – Execução 3: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste

	NB		SVM	
	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>
<i>Split da base para treinamento e teste</i>				
<i>Cross-validation (folds 10)</i>	86.3352 %	13.6648 %	93.2214 %	6.7786%

Fonte: Nathalia Nieuwenhoff, 2017

A tabela 21 apresenta os valores de tempo de execução dos algoritmos, sendo estes subdivididos em tempo de construção do modelo (*Time taken to build model*) e tempo de teste do modelo para repartição de dados de teste (*Time taken to test model on test split*). A partir disso, nota-se que o algoritmo NB apresentou tempo de construção do modelo significativamente inferior ao SVM. Adicionalmente, nota-se que o *cross-validation* não possui valores de teste do modelo apresentados, visto que conforme descrito na seção 4.7.3 do capítulo 4, para este método, a ideia de fragmentação da base entre construção e teste não é aplicável, visto que ele efetua a validação cruzada de todos os dados em ambas através de conjuntos de dados aleatórios, sendo assim utilizado o banco de dados por completo para treinamento e teste do modelo.

Tabela 21 – Execução 3: Tabela de tempo de execução dos algoritmos por tipo de teste

	NB		SVM	
	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>
<i>Split da base para treinamento e teste</i>				
<i>Cross-validation (10 folds)</i>	1.21 segundos	-	5365.56 segundos	-

Fonte: Nathalia Nieuwenhoff, 2017

Conforme detalhado na seção 4.8 do capítulo 4, quando se têm classes desbalanceadas, é importante considerar outras variáveis para comparação do desempenho em relação aos fatores de acuracidade dos algoritmos de classificação, conforme apresentadas na tabela 21. A partir disso, nota-se que o SVM apresentou valores médios finais de todas as métricas, sendo estas *TP Rate*, *FP Rate*, *Precision* e *F-measure*, detalhados na linha *Weighted Avg*, sendo estes 0.932, 0.425, 0.965, 0.946, respectivamente, superiores ao NB, em que são 0.863, 0.359, 0.963 e 0.905 respectivamente.

Analisando estas mesmas medidas por classe, sendo estas, “Filme” e “Nao filme”, observamos que os resultados de um algoritmo em relação ao outro apresentam variações. Considerando apenas a classe “Filme”, em que esta passa a ser a classe positiva (“a”) da matriz de confusão, conforme inversão da matriz anteriormente apresentada na tabela 22, nota-se que as métricas *TP Rate* e *FP Rate* do SVM são inferiores ao NB, mas o SVM apresentou valores superiores para as métricas *Precision* (precisão), dado que classificou menos eventos como falsos negativos que o NB, e *F-measure*, que se trata da média ponderada entre as métricas de precisão e *TP Rate*, apresentando assim valor superior dado que foi impactada pelo valor da métrica de precisão.

Já para a classe “Nao filme”, a superioridade do SVM acaba sendo mantida, exceto para a métrica *Precision*, que pode ser explicado pelo fato de o NB ter classificado menos eventos como falsos positivos que o SVM.

As matrizes de confusão de ambos os classificadores, conforme apresentadas na tabela 23, reforçam tais resultados apresentados, afinal, nota-se que o SVM classificou mais eventos como verdadeiros positivos (TP) e menos eventos como falsos negativos (FN) que o NB. Porém, nota-se que o NB classificou mais eventos como verdadeiros negativos (TN) e menos como falsos positivos (FP) que o SVM. Tais resultados expressam que o SVM foi capaz de prever corretamente mais eventos da classe “Não filme” que o NB, bem como o NB foi capaz de prever corretamente mais eventos da classe “Filme” que o SVM. Adicionalmente, apesar de apresentar valores inferiores que o NB para a classe “Filme”, nota-se que, nesta execução, o SVM foi capaz de prever eventos desta classe, diferentemente do que foi apresentado nas execuções 1 e 2, conforme detalhado nas seções 5.1 e 5.2, respectivamente.

Tabela 22 – Execução 3: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka

Classe	NB				SVM			
	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>
Filme	0.634	0.130	0.126	0.210	0.564	0.057	0.227	0.323
Não Filme	0.870	0.366	0.988	0.925	0.943	0.436	0.986	0.964
<i>Weighted Avg.</i>	0.863	0.359	0.963	0.905	0.932	0.425	0.965	0.946

Fonte: Nathalia Nieuwenhoff, 2017

Tabela 23 – Execução 3: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação

NB			SVM		
a	b	<-- classified as	a	b	<-- classified as
108431	16182	a = naofilme	117525	7088	a = naofilme
1350	2337	b = filme	1609	2078	b = filme

Fonte: Nathalia Nieuwenhoff, 2017

A tabela 24 apresenta os valores de estatística Kappa (*Kappa statistic*) e Área ROC para ambos os algoritmos. A partir disso, nota-se que o SVM apresentou valor de estatística Kappa superior ao NB, porém, ambos correspondem a *Fair agreement*, conforme a tabela sugerida por Landis e Koch (1977) apresentada na seção 4.8 do capítulo 4, e apesar de ser um resultado de concordância entre baixa e intermediária, ou seja, apesar de o valor final de concordância da predição realizada pelo algoritmo com a verdadeira classe (classe real) ser fraca, o SVM não apresentou nesta execução o indicador correspondente a nenhuma concordância ou que a concordância do classificador foi exatamente a esperada pelo acaso, o que é explicado devido ao fato de ter sido capaz de prever eventos para ambas as classes.

Por fim, nota-se que a Área ROC, cujos valores variam entre 0.0 e 1.0, sendo 1.0 o melhor caso, conforme detalhado na seção 4.8 do capítulo 4, o algoritmo NB apresentou valor superior ao SVM, sendo estes, 0.865 e 0.753, respectivamente, o que indica que a probabilidade de escolha aleatória de um evento classificado corretamente nos dados de teste esteja acima de um evento classificado incorretamente também selecionado aleatoriamente é superior no algoritmo NB em relação ao SVM. Adicionalmente, nota-se que nesta execução o valor da Área ROC do SVM não corresponde a 0.5 ou valor inferior, o que, segundo Witten, Eike e Hall (2011), indica que este não se comportou como um classificador aleatório, indicador, também, explicado devido ao fato de ter sido capaz de prever eventos para ambas as classes.

As figuras 23 e 24 apresentam as Curvas ROC de ambos classificadores, NB e SVM, respectivamente, para a classe “Filme” geradas pela ferramenta Weka, notando-se, assim, que a curva do algoritmo NB, está mais próxima da ponto ideal (0.0,1.0), conforme apresentado na seção 4.8 do capítulo 4, sendo mais uma métrica de acuracidade que o algoritmo apresentou valor superior em relação ao SVM.

Tabela 24 – Execução 3: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka

NB		SVM	
<i>Kappa statistic</i>	<i>ROC Area</i>	<i>Kappa statistic</i>	<i>ROC Area</i>
0.1707	0.865	0.2944	0.753

Fonte: Nathalia Nieuwenhoff, 2017

Figura 23 – Execução 3: Ilustração da curva ROC do NB gerada pelo Weka

Fonte: Nathalia Nieuwenhoff, 2017

Figura 24– Execução 3: Ilustração da curva ROC do SVM gerada pelo Weka

Fonte: Nathalia Nieuwenhoff, 2017

Dessa forma, para esta execução, nota-se que o SVM apresentou melhor desempenho para a maioria das métricas de acuracidade em relação ao NB, sendo estas: percentual de eventos corretamente classificados, valores médios das métricas *TP Rate*, *FP Rate*, *Precision* e *F-measure*, bem como valores de *TP Rate*, *FP Rate* e *F-measure* para a classe “Não Filme” e valores de *Precision* e *F-measure* para a classe “Filme”. E apesar de apresentar a maioria dos valores das métricas *TP Rate* e *FP Rate* inferiores ao NB para a classe “Filme”, bem como o valor de precisão para a classe “Não filme”, nota-se que estes possuem variações muito reduzidas, correspondendo a 0.07, 0.073 e 0.002, respectivamente.

Adicionalmente, apesar de seu tempo de execução ter sido bem superior ao NB, consideramos que este não é um fator crítico para este tipo de estudo. E mesmo também apresentando valor da área ROC inferior ao NB, nota-se que nesta execução o SVM também foi capaz de prever eventos de ambas as classes, diferentemente do que foi apresentado nas execuções 1 e 2, conforme detalhado nas seções 5.1 e 5.2 deste capítulo, respectivamente. Este fato reforça uma das indicações de Gadi (2008) para situações em que se têm classes desbalanceadas, como o banco de dados objeto deste estudo, em que Witten, Eike e Hall (2011) sinalizam a limitação de alguns algoritmos em treinar a classe que contém a minoria

dos registros, uma das formas de minimizar o problema de desbalanceamento é considerar a execução dos algoritmos a partir do classificador *Cost Sensitive*.

5.4 Execução 4: Banco de dados após a seleção de variáveis por meio do algoritmo *Information Gain* considerando o classificador *Cost Sensitive*

Esta seção apresenta os resultados obtidos a partir da execução dos algoritmos de classificação, NB e SVM, para o banco de dados reduzido após a execução do algoritmo de seleção de atributos *Information Gain*, utilizando os parâmetros *default* do Weka, porém executando tais algoritmos combinados com o algoritmo de classificação *Cost Sensitive*, conforme apresentado na seção 4.7.1 do capítulo 4. Optamos por efetuar esta execução utilizando apenas o método *cross-validation (10 folds)*, já que, conforme resultados obtidos nas execuções 1 e 2, apresentadas nas seções 5.1 e 5.2 deste capítulo, respectivamente a variação dos resultados gerais entre os métodos de divisão das bases utilizados possuem valores muito pequenos, considerados irrelevantes neste estudo.

A tabela 25 apresenta o resultado da execução de ambos algoritmos executados a partir do classificador *Cost Sensitive* no que se refere ao percentual de eventos corretamente classificados (*Correctly Classified Instances*) e percentual de eventos incorretamente classificados (*Incorrectly Classified Instances*). A partir disso, nota-se que ambos apresentaram tais percentuais de eventos corretamente classificados superiores a execução 3, conforme apresentado na seção 5.3 deste capítulo. Adicionalmente, nota-se que o SVM apresentou este indicador de acuracidade superior ao NB, visto que seu percentual de eventos corretamente classificados corresponde a 94.6726%, enquanto o NB apresentou 88.1372%, o que diretamente resulta em percentuais de eventos incorretamente classificados do SVM inferiores ao NB.

Tabela 25 – Execução 4: Tabela do percentual de eventos corretamente e incorretamente classificados por algoritmo e por tipo de teste

<i>Split</i> da base para treinamento e teste	NB		SVM	
	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>	<i>Correctly Classified Instances</i>	<i>Incorrectly Classified Instances</i>
<i>Cross-validation (folds 10)</i>	88.1372%	11.8628%	94.6726%	5.3274%

A tabela 26 apresenta os valores de tempo de execução dos algoritmos, sendo estes subdivididos em tempo de construção do modelo (*Time taken to build model*) e tempo de teste do modelo para repartição de dados de teste (*Time taken to test model on test split*). A partir disso, nota-se que o algoritmo NB apresentou tempo de construção do modelo significativamente inferior ao SVM, bem como ambos algoritmos apresentaram tempos inferiores quando comparados com a execução 3, conforme apresentada na seção 5.3 deste capítulo. Adicionalmente, nota-se que o *cross-validation* não possui valores de teste do modelo apresentados, visto que conforme descrito no Capítulo 5, para este método, a ideia de fragmentação da base entre construção e teste não é aplicável, visto que ele efetua a validação cruzada de todos os dados em ambas através de conjuntos de dados aleatórios, sendo assim utilizado o banco de dados por completo para treinamento e teste do modelo.

Tabela 26 – Execução 4: Tabela de tempo de execução dos algoritmos por tipo de teste

	NB		SVM	
<i>Split</i> da base para treinamento e teste	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>	<i>Time taken to build model</i>	<i>Time taken to test model on test split</i>
<i>Cross-validation (10 folds)</i>	0.34 segundos	-	593.25 segundos	-

Fonte: Nathalia Nieuwenhoff, 2017

Conforme detalhado na seção 4.8 do capítulo 4, quando se têm classes desbalanceadas, é importante considerar outras variáveis para comparação do desempenho em relação aos fatores de acuracidade dos algoritmos de classificação, conforme apresentadas na tabela 27. A partir disso, nota-se que o SVM apresentou valores médios finais de todas as métricas, sendo estas *TP Rate*, *FP Rate* e *F-measure*, detalhados na linha *Weighted Avg*, sendo estes 0.947, 0.624 e 0.952, respectivamente, superiores ao NB, em que são 0.881, 0.424 e 0.915, respectivamente. Já para a métrica *Precision*, apesar de bem próximos, o algoritmo NB apresentou valor superior ao SVM, sendo estes 0.961 e 0.959. Adicionalmente, com exceção da métrica *Precision*, os valores de tais métricas de ambos algoritmos são superiores em relação a execução 3, conforme apresentada na seção 5.3 deste capítulo,

Analisando estas mesmas medidas por classe, sendo estas, “Filme” e “Nao filme”, observamos que os resultados de um algoritmo em relação ao outro apresentam variações. Considerando apenas a classe “Filme”, em que esta passa a ser a classe positiva (“a”) da matriz de confusão, conforme inversão da matriz anteriormente apresentada na tabela 8 da

seção 4.8 do capítulo 4, nota-se que as métricas *TP Rate* e *FP Rate* do SVM são inferiores ao NB, mas o SVM apresentou valores superiores para as métricas *Precision* (precisão), dado que classificou menos eventos como falsos negativos que o NB, e *F-measure*, que se trata da média ponderada entre as métricas de precisão e *TP Rate*, apresentando assim valor superior dado que foi impactada pelo valor da métrica de precisão.

Já para a classe “Nao filme”, a superioridade do SVM acaba sendo mantida, exceto para a métrica *Precision*, que pode ser explicado pelo fato de o NB ter classificado menos eventos como falsos positivos que o SVM.

As matrizes de confusão de ambos os classificadores, conforme apresentadas na tabela 28, reforçam tais resultados apresentados, afinal, nota-se que o SVM classificou mais eventos como verdadeiros positivos (TP) e menos eventos como falsos negativos (FN) que o NB. Porém, nota-se que o NB classificou mais eventos como verdadeiros negativos (TN) e menos como falsos positivos (FP) que o SVM. Tais resultados expressam que o SVM foi capaz de prever corretamente mais eventos da classe “Não filme” que o NB, bem como o NB foi capaz de prever corretamente mais eventos da classe “Filme” que o SVM. Adicionalmente, apesar de apresentar valores inferiores que o NB para a classe “Filme”, nota-se que nesta execução o SVM foi capaz de prever eventos desta classe, como na execução 3 detalhada na seção 5.3 deste capítulo, e diferentemente do que foi apresentado nas execuções 1 e 2, conforme detalhado nas seções 5.1 e 5.2 deste capítulo, respectivamente.

Tabela 27 – Execução 4: Tabela das métricas de acuracidade por classe após a execução dos algoritmos na ferramenta Weka

Classe	NB				SVM			
	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>
Filme	0.567	0.109	0.133	0.215	0.359	0.036	0.228	0.279
Não Filme	0.891	0.433	0.986	0.936	0.964	0.641	0.981	0.972
<i>Weighted Avg.</i>	0.881	0.424	0.961	0.915	0.947	0.624	0.959	0.952

Fonte: Nathalia Nieuwenhoff, 2017

Tabela 28 – Execução 4: Matrizes de confusão geradas a partir da execução dos algoritmos de classificação

NB	SVM
----	-----

a	b	<-- classified as	a	b	<-- classified as
110990	13623	a = naofilme	120141	4472	a = naofilme
1597	2090	b = filme	2363	1324	b = filme

Fonte: Nathalia Nieuwenhoff, 2017

A tabela 29 apresenta os valores de estatística Kappa (*Kappa statistic*) e Área ROC para ambos os algoritmos. A partir disso, nota-se que o SVM apresentou valor de estatística Kappa superior ao NB, porém, ambos correspondem a *Fair agreement*, conforme a tabela sugerida por Landis e Koch (1977) apresentada na seção 4.8 do capítulo 4, e apesar de ser um resultado de concordância entre baixa e intermediária, ou seja, apesar de o valor final de concordância da predição realizada pelo algoritmo com a verdadeira classe (classe real) ser fraca, como na execução 3 detalhada na seção 5.3 deste capítulo, o SVM não apresentou nesta execução o indicador correspondente a nenhuma concordância ou que a concordância do classificador foi exatamente a esperada pelo acaso, o que é explicado devido ao fato de ter sido capaz de prever eventos para ambas as classes. Adicionalmente, comparando tais valores de estatística Kappa com os obtidos na execução 3, conforme detalhado na seção 5.3 deste capítulo, nota-se que o valor para o NB apresentou pequena variação superior e para o SVM apresentou redução.

Por fim, nota-se que a Área ROC, cujos valores variam entre 0.0 e 1.0, sendo 1.0 o melhor caso, conforme detalhado na seção 4.8 do capítulo 4, o algoritmo NB apresentou valor superior ao SVM, sendo estes, 0.851 e 0.662, respectivamente, o que indica que a probabilidade de escolha aleatória de um evento classificado corretamente nos dados de teste esteja acima de um evento classificado incorretamente também selecionado aleatoriamente é superior no algoritmo NB em relação ao SVM. Adicionalmente, nota-se que nesta execução o valor da Área ROC do SVM não corresponde a 0.5 ou valor inferior, o que, segundo Witten, Eike e Hall (2011), indica que este não se comportou como um classificador aleatório, indicador, também, explicado devido ao fato de ter sido capaz de prever eventos para ambas as classes. Adicionalmente, comparando tais valores de área ROC com os obtidos na execução 3, conforme detalhado na seção 4.8 do capítulo 4, nota-se que o NB manteve o mesmo valor e o SVM apresentou redução.

As figuras 25 e 26 apresentam as Curvas ROC de ambos classificadores, NB e SVM, respectivamente, para a classe “Filme” geradas pela ferramenta Weka, notando-se, assim, que a curva do algoritmo NB, está mais próxima da ponto ideal (0.0,1.0), conforme apresentado

na seção 4.8 do capítulo 4, sendo mais uma métrica de acuracidade que o algoritmo apresentou valor superior em relação ao SVM.

Tabela 29 – Execução 4: Tabela de valores do Índice Kappa e Área ROC gerados após a execução dos algoritmos de classificação no Weka

NB		SVM	
<i>Kappa statistic</i>	<i>ROC Area</i>	<i>Kappa statistic</i>	<i>ROC Area</i>
0.1772	0.851	0.253	0.662

Fonte: Nathalia Nieuwenhoff, 2017

Figura 25 – Execução 4: Ilustração da curva ROC do NB gerada pelo Weka

Fonte: Nathalia Nieuwenhoff, 2017

Figura 26 – Execução 4: Ilustração da curva ROC do SVM gerada pelo Weka

Fonte: Nathalia Nieuwenhoff, 2017

Dessa forma, para esta execução, os resultados gerais se assemelham aos da execução 3, conforme apresentados na seção 5.3 deste capítulo. Nota-se que o SVM apresentou melhor desempenho para a maioria das métricas de acuracidade em relação ao NB, sendo estas: percentual de eventos corretamente classificados, valores médios das métricas *TP Rate*, *FP Rate* e *F-measure*, bem como valores de *TP Rate*, *FP Rate* e *F-measure* para a classe “Não Filme” e valores de *Precision* e *F-measure* para a classe “Filme”. E apesar de apresentar a maioria dos valores das métricas *TP Rate* e *FP Rate* inferiores ao NB para a classe “Filme”, bem com o valor de precisão médio e para a classe “Não filme”, nota-se que estes possuem variações muito reduzidas, correspondendo a 0.208, 0.073, 0.005 e 0.002, respectivamente.

Adicionalmente, apesar de seu tempo de execução ter sido bem superior ao NB, consideramos que este não é um fator crítico para este tipo de estudo. E mesmo também apresentando valor da área ROC inferior ao NB, nota-se que nesta execução o SVM também foi capaz de prever eventos de ambas as classes, conforme também apresentado na execução 3, detalhado na seção 5.3 deste capítulo, e diferentemente do que foi apresentado nas execuções 1 e 2, conforme detalhado nas seções 5.1 e 5.2 deste capítulo, respectivamente.

5.5 Comparação geral entre as execuções

Esta seção apresenta os resultados obtidos a partir da execução dos algoritmos de classificação, NB e SVM, para o banco de dados completo contendo os 51 atributos previamente selecionados, bem como utilizando os parâmetros *default* do Weka para execução.

A tabela 30 apresenta o resultado das quatro execuções dos algoritmos no que se refere ao percentual de eventos corretamente classificados (*Correctly Classified Instances*) e eventos incorretamente classificados (*Incorrectly Classified Instances*). A partir disso, nota-se que:

- O algoritmo SVM apresentou este indicador de acuracidade superior ao NB em todas as execuções.
- O melhor percentual de ambos algoritmos se refere a execução 2, referente a execução dos algoritmos para o banco de dados reduzido, após a execução do algoritmo de seleção de variáveis *Information Gain*, e sem a utilização do classificador *Cost Sensitive*, sendo que para o SVM, o resultado apresentado é mesmo da execução 1.
- Para ambos algoritmos, as execuções referentes às bases reduzidas, após a execução do algoritmo de seleção de variáveis *Information Gain*, sendo estas 2 e 4, apresentaram percentuais superiores aos das execuções efetuadas com o banco de dados completo, 1 e 3.
- Para ambos algoritmos, as execuções referentes a combinação destes com o classificador *Cost Sensitive*, sendo estas 3 e 4, apresentaram tais percentuais inferiores às execuções sem considerar o algoritmo em questão, sendo estas 1 e 2.

Tabela 30 –Tabela de percentuais de eventos corretamente e incorretamente classificados para as quatro execuções deste estudo experimental

	NB				SVM			
	Exe 1	Exe 2	Exe 3	Exe 4	Exe 1	Exe 2	Exe 3	Exe 4
<i>Correctly Classified Instances (%)</i>	90.70	93.09	86.34	88.14	97.13	97.13	93.22	94.67
<i>Incorrectly Classified Instances (%)</i>	9.30	6.91	13.66	11.86	2.87	2.87	6.78	5.33

Fonte: Nathalia Nieuwenhoff, 2017

Ao comparar ambos algoritmos em relação às quatro execuções no que se refere às métricas *TP Rate*, *FP Rate*, *Precision* e *F-measure*, para a classe “Filme”, ou seja, indivíduos que consumiram o produto cinema, conforme apresentado na tabela 31, nota-se que:

- O algoritmo SVM não foi capaz de prever eventos da classe em questão nas execuções 1 e 2, visto que todas as métricas possuem valor igual a zero. Resultado que se difere nas execuções 3 e 4, quando se considera o algoritmo classificador *Cost Sensitive*.
- O algoritmo SVM apresentou o seu melhor resultado referente a tais métricas na execução 3. O algoritmo NB apresentou melhor resultado para as métricas *TP Rate* e *FP Rate* também na execução 3, mas para as métricas de *Precision* e *F-measure* na execução 2.
- O melhor resultado da métrica *TP Rate* foi obtido pelo algoritmo NB na execução 3; o melhor resultado da métrica *FP Rate* foi obtido pelo algoritmo NB na execução 3; o melhor resultado da métrica *Precision* foi obtido pelo algoritmo SVM na execução 4; o melhor resultado da métrica *F-measure* foi obtido pelo algoritmo SVM na execução 3. Tais resultados sugerem que os melhores resultados de tais métricas foram obtidos nas execuções considerando o algoritmo classificador *Cost Sensitive*, sendo estas 3 e 4.
- As métricas *TP Rate* e *FP Rate*, para ambos algoritmos apresentaram nas execuções que envolvem o banco de dados reduzido, após a execução do algoritmo de seleção de variáveis *Information Gain*, sendo estas 2 e 4, pequenas reduções quando comparadas às execuções para o banco de dados completo, 1 e 3. Sendo que para o SVM considera-se somente a comparação entre a execução 3 e 4, visto que para as execuções 1 e 2 apresentou valores iguais a zero.
- Para todas as execuções, o algoritmo NB apresentou resultados referentes às métricas *TP Rate* e *FP Rate* superiores ao SVM. Já o SVM apresentou nas execuções 3 e 4, considerando o algoritmo classificador *Cost Sensitive*, resultados referentes às métricas *Precision* e *F-measure* superiores ao NB.

Tabela 31 – Tabela das métricas de acuracidade para a classe “Filme” referente às quatro execuções deste estudo experimental

NB				SVM			
<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>

Execução 1	0.504	0.081	0.155	0.237	0	0	0	0
Execução 2	0.389	0.053	0.178	0.244	0	0	0	0
Execução 3	0.634	0.13	0.126	0.21	0.564	0.057	0.227	0.323
Execução 4	0.567	0.109	0.133	0.215	0.359	0.036	0.228	0.279

Fonte: Nathalia Nieuwenhoff, 2017

Ao comparar ambos algoritmos em relação às quatro execuções referente às métricas *TP Rate*, *FP Rate*, *Precision* e *F-measure*, para a classe “Não filme”, ou seja, indivíduos que não consumiram o produto cinema, conforme apresentado na tabela 32, nota-se que:

- O algoritmo SVM foi capaz de prever corretamente todos os eventos da classe em questão nas execuções 1 e 2, visto que as métricas *TP Rate* e *FP Rate* possuem valores iguais a um.
- O algoritmo SVM apresentou o seu melhor resultado referente a tais métricas nas execuções 1 e 2. O algoritmo NB apresentou melhor resultado para as métricas *TP Rate* e *FP Rate* também na execução 2, mas para as métricas de *Precision* e *F-measure* na execução 1.
- O melhor resultado da métrica *TP Rate* foi obtido pelo algoritmo SVM nas execuções 1 e 2; o melhor resultado da métrica *FP Rate* foi obtido pelo algoritmo SVM nas execuções 1 e 2; o melhor resultado da métrica *Precision* foi obtido pelo algoritmo NB na execução 3; o melhor resultado da métrica *F-measure* foi obtido pelo algoritmo SVM nas execuções 1 e 2; Tais resultados sugerem que os melhores resultados de tais métricas foram obtidos nas execuções considerando o algoritmo *Cost Classifier*.
- Ao comparar as execuções 1 e 2 com as execuções 3 e 4, nota-se que ambos algoritmos apresentaram melhores resultados gerais para as métricas em questão nas execuções 1 e 2. Sendo que na execução 2, considerando o banco de dados reduzido, ambos apresentaram seus melhores resultados.
- As métricas *TP Rate*, *FP Rate* e *F-measure*, para ambos algoritmos, apresentaram aumento dos valores nas execuções que envolvem o banco de dados reduzido, após a execução do algoritmo de seleção de variáveis *Information Gain*, sendo estas 2 e 4, quando comparadas às execuções para o banco de dados completo, 1 e 3. Sendo que para o SVM considera-se somente a comparação entre a execução 3 e 4, visto que para as execuções 1 e 2 apresentou valores iguais a zero.
- Com execução da métrica precisão, ambos algoritmos apresentaram melhores resultados quando executados para o banco de dados reduzido, execuções 2 e 4.

- Para todas as execuções, o algoritmo SVM apresentou resultados referentes às métricas *TP Rate*, *FP Rate* e *F-measure* superiores ao NB. Já o NB apresentou resultados referentes à métrica *Precision* superiores ao NB.

Tabela 32 – Tabela das métricas de acuracidade para a classe “Não Filme” referente às quatro execuções deste estudo experimental

	NB				SVM			
	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>	<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision</i>	<i>F-measure</i>
Execução 1	0.919	0.496	0.984	0.95	1	1	0.971	0.985
Execução 2	0.947	0.611	0.981	0.964	1	1	0.971	0.985
Execução 3	0.87	0.366	0.988	0.925	0.943	0.436	0.986	0.964
Execução 4	0.891	0.433	0.986	0.936	0.964	0.641	0.981	0.972

Fonte: Nathalia Nieuwenhoff, 2017

Ao comparar ambos algoritmos em relação às quatro execuções referente às métricas Área ROC e estatística *Kappa*, conforme apresentado na tabela 33, nota-se que:

- O algoritmo NB apresentou redução dos valores de estatística *Kappa* para ambas as execuções envolvendo o classificador *Cost Sensitive*, 3 e 4, quando comparadas com as execuções sem considerar tal algoritmo, sendo estas 1 e 2. Já o SVM apresentou comportamento inverso, ou seja, apresentou valores superiores nas execuções 3 e 4.
- Para as execuções 1 e 2, o SVM apresentou valores de ambas as métricas que indicam: (i) nenhuma concordância ou concordância “pobre”, segundo Landis e Koch (1977), referente a estatística *Kappa*; e comportamento de um classificador aleatório, segundo Witten, Eike e Hall (2011), referente ao valor apresentado da área ROC.
- O algoritmo NB apresentou valores da Área ROC superiores ao SVM em todas as execuções, o que indica superior probabilidade de escolha aleatória de um evento classificado corretamente nos dados de teste esteja acima de um evento que foi classificado incorretamente também selecionado aleatoriamente, segundo Witten, Eike e Hall (2011).
- Os valores de área ROC do NB apresentam redução nos resultados obtidos das execuções referentes ao banco de dados reduzido, sendo estas 2 e 4, indicando que ao reduzir os atributos, a probabilidade mencionada no item acima também reduz. Porém, notam-se variações positivas referentes aos valores de estatística *Kappa*.

- Considerando as execuções 3 e 4, o SVM apresentou pequenas reduções dos valores de tais métricas quando executado para o banco de dados reduzido, execução 4.

Tabela 33 – Tabela das métricas de estatística Kappa e área ROC referente às quatro execuções deste estudo experimental

	NB				SVM			
	Exe 1	Exe 2	Exe 3	Exe 4	Exe 1	Exe 2	Exe 3	Exe 4
Kappa statistic	0.2024	0.2132	0.1707	0.1772	0	0	0.2944	0.253
ROC Area	0.865	0.851	0.865	0.851	0.5	0.5	0.753	0.662

Fonte: Nathalia Nieuwenhoff, 2017

6 Conclusão

Os padrões de decisão do consumo, relacionados diretamente com as preferências dos indivíduos, dependem de uma composição de características individuais, aspectos culturais, econômicos e sociais segregados e agrupados. Especificamente relacionado ao consumo cultural, considerando o bem de informação filmes em salas de cinema, cuja abordagem também engloba a noção de cidadania impactada pelo acesso, classificar os indivíduos e identificar padrões de consumo no Brasil, cujas informações relativas a este tema se encontram dispostas em microdados, torna-se uma tarefa desafiadora, não somente pelo fato de se ter uma quantidade elevada de registros e atributos, o que impossibilita uma análise manual, bem como pela complexidade de tais fatores que influenciam a decisão de consumo e a heterogeneidade envolvida no objeto de estudo em questão.

Em paralelo, a utilização da técnica de extração de conhecimento em banco de dados volumosos, não-triviais e dinâmicos, também conhecida como Descoberta do Conhecimento ou *Knowledge Discovery in Databases* (KDD), para geração de conhecimento a partir de padrões de dados válidos, complexos e potencialmente úteis, conforme sugerida por Fayyad, Shapiro e Smyth, 1996, vem sendo comumente aplicada nos mais variados cenários, provocando, assim, o amplo estudo e desenvolvimento de técnicas computacionais eficientes e eficazes para suportar este processo. Neste contexto, como parte da etapa de Mineração de Dados, que se trata de uma etapa do processo de KDD, a tarefa de Classificação de Dados se aplica como uma alternativa para Classificação de Dados que envolvem a decisão de consumo dos bens de informação, visto que é capaz de examinar o conjunto de registros marcados com seus atributos correspondentes e identificar as características (variáveis) que definem tais classes, sendo capaz, assim de prever as classes que os registros ainda não classificados pertencem.

Dessa forma, este trabalho realizou o estudo experimental a partir da aplicação do processo de Descoberta do Conhecimento (KDD), o que inclui as etapas de seleção e Mineração de Dados, para um problema de classificação binária, indivíduos brasileiros que consomem e não consomem um bem de informação, filmes em salas de cinema. A etapa de seleção dos dados foi conduzida a partir de aspectos gerais da teoria econômica, após entendimento e análise do banco de dados objeto deste estudo. O banco de dados se refere aos microdados da Pesquisa de Orçamento Familiar (POF), realizada em um período de 12 meses entre os anos 2008 e 2009 e disponibilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), contendo informações de consumo de cinema para um período de referência de 30

dias. Adicionalmente, todas as etapas que envolvem o pré-processamento do banco de dados, o que também inclui a seleção inicial dos dados, foram executadas na ferramenta Stata. A etapa de Mineração de Dados foi conduzida através da execução de dois algoritmos de aprendizagem de máquina, sendo estes *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), para treinamento e teste dos modelos para classificação de tais dados na ferramenta Weka.

A análise comparativa dos algoritmos foi efetuada para as quatro execuções efetuadas neste experimento considerando medidas relacionadas à acuracidade das técnicas aplicadas e seus desempenhos computacionais. Sendo estas: (i) execução dos algoritmos de classificação para o banco de dados completo contendo os 51 atributos previamente selecionados, denominada execução 1; (ii) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain*, mas não considerando o classificador *Cost Sensitive*, denominada execução 2; (iii) execução considerando o mesmo banco de dados, porém executando tais algoritmos através do algoritmo de classificação *Cost Sensitive*, denominada execução 3; (iv) execução de tais algoritmos considerando somente os atributos do banco de dados elencados como mais relevantes após a execução do algoritmo de seleção de variáveis *Information Gain* e considerando o classificador *Cost Sensitive*, denominada execução 4.

Os resultados obtidos nas execuções dos algoritmos indicam que analisar apenas os valores médios das métricas em situações em que se possuem classes desbalanceadas pode sugerir resultados que são correspondem ao desempenho do algoritmo para ambas as classes, conforme sinalizado por Witten, Eike e Hall (2011), visto que a partir das execuções 1 e 2 do algoritmo SVM, referentes às execuções sem a combinação do algoritmo classificador *Cost Sensitive*, tais resultados apresentam valores médios, bem como percentual de eventos corretamente classificados que indicam elevada acuracidade de desempenho do algoritmo. Afinal, por exemplo, apresentou valores médios para as métricas *TP Rate*, *FP Rate*, *Precision* e *F-measure* correspondentes a 0.971, 0.971, 0.943 e 0.957, respectivamente, cujo melhor caso é 1, bem como percentuais de eventos corretamente classificados em valores aproximados a 97%, porém, nestas mesmas execuções não foi capaz de prever corretamente pelo menos evento da classe “Filme”, apresentando valores iguais a zero para estas mesmas métricas quando analisada somente a classe em questão.

A partir disso, uma análise inicial somente considerando tais métricas médias poderia sugerir um resultado que não atende ao problema de classificação proposto, visto que neste estudo não há relevância entre classes, o que sugere que quando não se utilizar procedimentos adicionais para a tratativa do desbalanceamento entre as classes, o SVM é um algoritmo que

apresenta uma limitação para este tipo de banco de dados, sugerindo, assim, que o NB tenha apresentado melhor desempenho, pois apesar de não apresentar os valores médios destas métricas de acuracidade superiores ao SVM, comportou-se de forma balanceada, sendo capaz de prever eventos de ambas as classes, o que é reforçado pelos resultados das métricas de precisão (*Precisão*), visto que como tal medida é composta pela quantidade de eventos de ambas as classes, caso os algoritmos apresentem desempenhos desbalanceados entre elas, seu valor será impactado. E dependendo desta variação, pode também impactar a métrica *F-measure*, visto que é calculada a partir da métrica de precisão, por se tratar de uma média aritmética entre precisão e *recall*. Adicionalmente, as métricas de estatística Kappa, área e curva ROC também reforçam este aspecto de superioridade final do NB em relação ao SVM considerando o problema de classificação estudado, afinal, nestas mesmas execuções, o NB apresentou resultados superiores ao SVM para tais métricas. Estes resultados são explicados pelo fato de tais métricas não serem fortemente impactadas pela quantidade de eventos que cada classe possui, considerando apenas o desempenho médio dos algoritmos, mas por indicarem a acuracidade geral dos algoritmos para cada uma das classes, independente da proporção de eventos que cada uma possui.

Witten, Eike e Hall (2011) sinalizam que resultados similares a estes são obtidos quando se executam alguns algoritmos de Classificação de Dados, o que inclui o SVM, para banco de dados que possuem classes desbalanceadas e, com isso, este é apenas capaz de treinar a classe que contém a maioria dos registros, o que é confirmado pelos resultados obtidos neste experimento, visto que a classe “Filme”, que representa os indivíduos que consumiram o bem de informação, filmes em salas de cinema, corresponde a aproximadamente apenas 3% dos registros do banco de dados objeto deste estudo, sendo assim, a classe minoritária.

Os resultados das execuções sugerem que a utilização do algoritmo *Cost Sensitive*, conforme sugerido por Gadi (2008), apresentou resultados positivos, afinal, ao combinar o algoritmo em questão com o SVM, apesar de ter classificado incorretamente alguns dos eventos da classe majoritária “Não filme”, que representa os indivíduos que não consumiram cinema, notou-se que este foi capaz de prever os eventos de ambas as classes, apresentando, assim, um desempenho mais equilibrado atendendo ao problema de classificação composto, além de apresentar superioridade em relação ao NB para a maioria das métricas de acuracidade.

Adicionalmente, os resultados obtidos em todas as execuções indicam que caso sejam estudados problemas de classificação em que se tenha uma classe mais relevante que a outra,

caso esta seja a minoritária, o NB apresentou resultados superiores em relação às métricas de acuracidade em relação ao SVM. Porém, caso a classe prioritária seja a majoritária, o SVM apresentou resultados superiores, bem como foi capaz de prever corretamente todos os eventos da classe em questão nas execuções 1 e 2, independente da redução da quantidade de atributos considerada entre uma execução e a outra.

A partir dos resultados obtidos, notou-se, também, para ambos algoritmos, que quando as métricas de acuracidade de uma classe apresentam melhorias nos resultados, as da outra apresentam reduções, o que indica que a quantidade de registros por classe é um fator que influencia diretamente o desempenho de tais algoritmos. Também foi observado que para todas as execuções, o algoritmo NB apresentou os melhores resultados em relação às métricas de área e curva ROC, o que possibilita afirmar que a probabilidade de escolha aleatória de um evento classificado corretamente nos dados de teste esteja acima de um evento classificado incorretamente também selecionado aleatoriamente é para o algoritmo NB superior em relação ao SVM.

Adicionalmente, os resultados obtidos em relação aos valores de estatística Kappa mostram que ambos algoritmos apresentaram em todas as execuções valores que indicam concordância “Frac”, segundo a classificação sugerida por Landis e Koch (1977), ou seja, o grau de concordância das predições realizadas pelos algoritmos com as classes verdadeiras (classes reais) é fraco.

Em relação aos métodos de fragmentação da base para treinamento e teste considerados na primeira e segunda execuções, sendo eles: (i) 50% treinamento e 50% teste; (ii) 75% treinamento; (iii) 25% teste e *cross-validation (10 folds)*, os resultados apresentaram variações irrelevantes, indicando, assim, que os métodos em questão de repartição dos dados para construção do modelo e testes não é um fator significativo para o desempenho dos classificadores, considerando este banco de dados e o problema de classificação analisado neste estudo. Já em relação ao tempo de execução dos algoritmos, o NB apresentou tempo significativamente inferior ao SVM em todas as execuções.

Ao comparar as execuções dos algoritmos entre o banco de dados completos composto pelos 51 atributos após a seleção das variáveis conduzida a partir de aspectos gerais da teoria econômica, após entendimento e análise do banco de dados objeto deste estudo, com o banco de dados reduzido, a partir da pré-seleção adicional dos atributos através do algoritmo *Information Gain*, não apresentam variações significativas. Adicionalmente, algumas métricas apresentaram melhores resultados nas execuções considerando o banco de dados reduzido quando comparadas às execuções utilizando o banco de dados completo, como, por exemplo:

(i) percentuais gerais de eventos corretamente classificados; (ii) métricas *TP Rate*, *FP Rate* e *Precision* referentes a classe majoritária; (iii) métrica *Precision* referente a classe minoritária e (iv) estatística Kappa.

A execução da seleção adicional dos atributos através do algoritmo *Information Gain* resultou em seis atributos com maiores valores médios de correlação (*average merit*), ou seja, sendo considerados o que possuem maior relevância para os classificadores. Sendo estes: Despesa total em Recreação e cultura; Anos de estudos; Renda per capita; Renda bruta monetária; Quantidade de banheiros do domicílio; e Se o indivíduo possui plano de saúde.

Os atributos selecionados pelo algoritmo de seleção de variáveis sugerem que a decisão de consumo de cultura, mais especificamente do bem de informação, filmes em cinema, está fortemente relacionada a fatores dos indivíduos relacionados a: (i) renda, visto que além das variáveis de renda bruta e per capita retornadas pelo algoritmo como relevantes, as variáveis relacionadas a quantidade de banheiros e a que indica se o indivíduo possui plano de saúde, também podem ser consideradas indicadoras de renda; (ii) nível de educação; e (iii) suas preferências por bens culturais, obtido a partir do valor gasto pelos indivíduos com bens culturais.

Adicionalmente, os resultados obtidos neste estudo a partir das execuções dos algoritmos sugerem que a escolha dos algoritmos de aprendizagem de máquina para Classificação de Dados está diretamente relacionada a fatores como: (i) importância das classes para o problema a ser estudado; (ii) balanceamento entre as classes; (iii) universo de atributos a serem considerados em relação a quantidade e grau de importância destes para o classificador.

A partir dos resultados obtidos neste trabalho, surgiram as seguintes propostas de trabalhos futuros: (i) Avançar os estudos para análise da contribuição específica dos atributos que são determinantes para a classificação de indivíduos consumidores e não consumidores de bens culturais, mais especificadamente, filmes em salas de cinema; (ii) Aplicar outros tipos de algoritmos, como redes neurais e árvores de decisão, para o problema de classificação estudado, para análise de seus desempenhos; (iii) Avançar os estudos relacionados a técnicas computacionais que visam minimizar problemas de desbalanceamento entre classes; (iv) Analisar o padrão dos usuários de banda larga ou telefones celulares que devem ter classes mais balanceadas entre consumidores e não consumidores para verificar se os resultados obtidos com ambos algoritmos, NB e SVM, apresentam variações significativas; (v) Aplicar os algoritmos para o banco de dados da POF 2015- 2016, quando for disponibilizado pelo

IBGE, para analisar se o padrão de consumo referente ao problema de classificação proposto se mantém ao longo do tempo.

Referências¹

- AGÊNCIA NACIONAL DE CINEMA. Informe Anual 2009, 2009. Disponível em: <<http://oca.ancine.gov.br/media/SAM/Informes/2009/InformeAnual2009.pdf>>. Acesso em: 10 mai. 2016.
- AGÊNCIA NACIONAL DE CINEMA. Valor Adicionado pelo setor audiovisual, 2015. Disponível em: <<http://oca.ancine.gov.br/media/SAM/Estudos/VALOR-ADICIONADO-PELO-SETOR-AUDIOVISUAL.pdf>>. Acesso em: 10 mai. 2016.
- ALI M., JUNG L.T. Confidentiality based file attributes and data classification using TsF-KNN. In: 2015 5th International Conference on IT Convergence and Security, [S.I.], 2015, ICITCS 2015 – Proceedings.
- AMO, S.; ROCHA, A.R. Mining Sequential Patterns using Genetic Programming. In: International Conference on Artificial Intelligence, Las Vegas, 2003, p. 451-456.
- ATTAL, F.; MOHAMMED, S.; DEDABRISHVILI, M.; CHAMROUKHI, F.; OUKHELLOU, L.; AMIRAT, Y. In: Physical human activity recognition using wearable sensors. Sensors (Switzerland), Suíça, v. 15, n. 12, p. 31314-31338, 2015.
- BIOLCHINI, J.; MIAN, P. G.; NATALI, A. C. C.; TRAVASSOS, G. H. **Systematic review in software engineering**. Rio de Janeiro: System Engineering and Computer Science Dept., COOPE-UFRJ, 2005. Technical report, RT-ES 679/05.
- BOEACHAT, A.A. **Algoritmos para a detecção de drifting em sensores de fundo de poço**. Santa Catarina: Universidade Federal de Santa Catarina, 2012.
- BUENO, Michel Ferreira; VIANA, Maury Reis; Mineração de Dados: aplicação, eficiência e resultados. In: Congresso de iniciação científica do INATEL – INCITEL, 2012, Santa Rita do Sapucaí. **Anais...** Minas Gerais: INATEL-INCITEL, 2012. p. 86-88.
- CAMILO, C. O.; SILVA, J. C. da. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Relatório Técnico – Universidade Federal de Goiás. Goiás: Universidade Federal de Goiás, 2009. 29 p.
- CANCLINI, N. G. **Consumidores e cidadãos: conflitos multiculturais da globalização**. 4 ed. Rio de Janeiro: Editora UFRJ, 1999.
- CHAO, Y. G.; CHEN, C. Semi-supervised method for gene expression data classification with Gaussian fields and harmonic functions. In: 2008 19th International Conference on Pattern Recognition, [S.I.], 2008, p. 1-4.
- COHEN, I.; COZMAN, F. G.; SEBE, N.; CIRELO, M. C.; HUANG, T. S. Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

interaction. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.I.], v. 26, n. 12, 1553-1566, 2004.

CZAIKOWKA, J.; RUDZKI, M.; CZAJKOWSKI, Z. A new fuzzy support vectors machine for biomedical data classification. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, [S.I.], 2008, p. 4676-4679.

DAVIS, P.; Spatial Competition in Retail Markets: Movie Theaters. **The RAND Journal of Economics**. Winter-Blackwell Publishing, [S.I.], v. 37, n. 4, p. 964-982, 2006.

DEHZANGI, O. ; MA, B.; CHNG, E. S.; LI, H. Fuzzy rule selection using Iterative Rule Learning for speech data classification. In: 2008 19th International Conference on Pattern Recognition, [S.I.], 2008, p. 1-4.

DINIZ, S.C. **Análise do consumo de bens e serviços artístico-culturais no Brasil metropolitano**. Minas Gerais: Universidade Federal de Minas Gerais, 2010.

ELBASI, E. Fuzzy Logic-Based Scenario Recognition from Video Sequences. **Journal of applied research and technology**, [S.I.], v. 11, n. 5, p. 702-707, 2013.

ESSIET, I. O.; AUDU, G. A. Determination of food age using neural network. In: IEEE 6th International Conference on Adaptive Science & Technology (ICAST), [S.I.], 2014, p. 1-4.

FAYYAD, U.; SHAPIRO, G.P.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, United States, v. 17, n. 3, 1996.

GADI, M. F. A. Uma comparação de métodos de classificação aplicados à detecção de fraude em cartões de crédito. 2008. 201 p. Tese (Mestrado em Ciências) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2008.

GALLAGHER, C; MADDEN, M.G.; D'ARCY, B. A Bayesian Classification Approach to Improving Performance for a Real-World Sales Forecasting Application. In: 2015 IEEE 14th International Conference on Machine Learning and Applications, Florida, 2015.

HAN, J.; KRAMER, M. **Data mining Concepts and Techniques**. 2nd ed. California: Morgan Kaufmann San Francisco, 2006.

HAYKIN, S. **Neural networks and learning machines**. 3. ed. [S.I.]: Prentice Hall, 2009.

HOFSTEDE, G. J.; MINKOV, M.; **Cultures and Organizations: Software of the Mind**. 3rd ed. [S.I.]: McGraw-Hill, 2010.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA: banco de dados da Pesquisa de Orçamento Familiar (POF) 2008-2009. Disponível em: <
http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/pof/2008_2009/microdados.shtm>. Acesso em: 05 Jun. 2015.

JEHLE, G. A.; RENY, P.J. **Advanced Microeconomic Theory**. 2nd ed., [S.I.]: Addison Wesley, 2000.

KINTO, E. A. Otimização e análise das máquinas de vetores de suporte aplicadas à classificação de documentos. 2011. 145p. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2011.

KRAWCZYK, B.; SCHAEFER, G. Ensemble fusion methods for medical data classification. In: 11th Symposium on Neural Network Applications in Electrical Engineering, [S.I.], 2012, p. 143-146.

LANDIS, J. R.; KOCH, G. G. **The Measurement of Observer Agreement for Categorical Data**. [S.I.]: International Biometric Society, 1977, v. 33, v. 1, p. 159-174.

LEE, J.; LEE, J. H. K-means clustering based SVM ensemble methods for imbalanced data problem. In: 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS), [S.I.], 2014, p. 614-617.

LEWICKE, A.; SAZONOV, E.; CORWIN, M. J.; NEUMAN, M.; SCHUCKERS, S. Sleep Versus Wake Classification From Heart Rate Variability Using Computational Intelligence: Consideration of Rejection in Classification Models. **IEEE Transactions on Biomedical Engineering**, [S.I.], v.55, n.1, p. 108- 118, 2008.

LIU H.; COQUIN, D.; VALET, L.; CERUTTI, G. Leaf species classification based on a botanical shape sub-classifier strategy. In: Proceedings - International Conference on Pattern Recognition, [S.I.], 2014, p. 1496- 1501.

LORENA, A.C.; CARVALHO, A. C. P. L. F. **Introdução às Máquinas de Vetores Suporte**. São Paulo: ICMC-USP, 2003. 201 p. Relatório técnico, ISSN - 0103-2569.

MINISTÉRIO DA CULTURA. Economia e Política Cultural: acesso, emprego e financiamento. Brasília, 2007.

MITCHELL, T.M. **Machine Learning**.1.ed. [S.1]: McGraw-Hill Science/Engineering/Math, 1997.

MITRA, V.; WANG C.; BANERJEE, S. Lidar detection of underwater objects using a neuro-SVM-based architecture. **IEEE Transactions on Neural Networks**, [S.I.], v. 16, n.3, p. 717-731, 2006.

NEF, T.; URWYLER, P.; BUCHLER, M.; TARNANAS, I.; STUCKI, R.; CAZZOLI, D.; MURI, R.; MOSIMANN, U. Evaluation of three state-of-the-art classifiers for recognition of activities of daily living from smart home ambient data. **Sensors (Switzerland)**, Suíça, v. 15, n. 5, p. 11725-11740, 2015.

NI, J.; LIANG, L. A Gait Recognition Method Based on KFDA and SVM. In: 2009 International Workshop on Intelligent Systems and Applications, [S.I.], 2009, p. 1-4.

NI, J.; LIANG, L. A New Method Based on KFDA and SVM for Gait Identification. In: 2009 International Workshop on Intelligent Systems and Applications, [S.I.], 2009, p. 1-3.

OSKOEI, M. A.; GAN, J. Q.; HU, H. Adaptive schemes applied to online SVM for BCI data classification. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, [S.I.], 2009, p. 2600-2003.

OYANG, Y.; HWANG S.; OU Y.; CHEN, C.; CHENG Z. Data classification with radial basis function networks based on a novel kernel density estimation algorithm. **IEEE Transactions on Neural Networks**, [S.I.], v. 16, n. 1, p. 225-236, 2005.

PAPAKOSTAS, G.A.; DIAMANTARA, K.I.; PAPADIMITRIOU, T. Parallel pattern classification utilizing GPU-based kernelized Slackmin algorithm. **Journal of Parallel and Distributed Computing**, [S.I.], v.99, p. 90-99, 2017.

PIMENTEL, E. P.; OMAR, N. Métricas para o Mapeamento do Conhecimento do Aprendiz em Ambientes Computacionais de Aprendizagem. In: XVII Simpósio Brasileiro de Informática na Educação, Brasília. **Anais do XVII Simpósio Brasileiro de Informática na Educação**. Brasília: [s.n.], 2006. p. 247-256.

PLATT, J. C. **Fast Training of Support Vector Machines Using Sequential Minimal Optimization, Advances in Kernel Methods - Support Vector Learning** 1st. ed. Cambridge, MA, USA: MIT Press, 1999.

POLAT, K. Data weighting method on the basis of binary encoded output to solve multi-class pattern classification problems. **Expert Systems with Applications**, [S.I.], v. 40, n. 11, p. 4637-4647, 2014.

PRATI, R. C., BATISTA, G. E. A. P. A., MONARD, M. C. Curvas ROC para a Avaliação de Classificadores. **Revista IEEE América Latina**, [S.I.], 2008.

ROGATI, M.; YANG, Y. High-performing feature selection for text classification. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2002. (CIKM '02), p. 659–661.

SCHALKOFF, R. J.; **Pattern Recognition: Statistical, Structural and Neural Approaches**. New York: J. Wiley 1991, 364 p.

SEGARAM, T. Advance classification: kernel methods and SVMs. In: Programming Collective Intelligence: Build Smart Web 2.0 Applications, O'Reilly, 2007.

SHAPIRO, C.; VARIAN, H., **Information Rules: a strategic guide to the network economy**. Boston: Harvard Business School Review Press, 1999.

SIGEL S., CASTELLAN N. **Nonparametric Statistics for the Behavioral Sciences**. 2nd ed. New York: McGraw-Hill, USA, 1988. p. 284-285.

SILVA, A. L. L. **Consumo de produtos culturais em São Paulo: Análise dos Fatores Antecedentes e Propostas de Modelo**. São Paulo: Universidade de São Paulo, 2008.

SILVA, S. da; DIAS, M. J.; LOPES, V. J. Damage detection in a benchmark structure using AR-ARX models and statistical pattern recognition. **Journal of the Brazilian Society of Mechanical Sciences and Engineering**, [S.I.], v.29, n. 2, p. 174-184, 2007.

SOLÍS, M.; BENITEZ, H.P.; RUBIO, E.; MEDINA-GÓMEZ, L.; MORENO, E.; GONZALEZ, G.; LEIJA, L. Pattern Classification of Decomposed Wavelet Information using ART2 Networks for echoes Analysis. **Journal of applied research and technology**, [S.I.], v. 6, n. 1, p. 33-44, 2008.

TABAKHI, S.; NAJAFI, A.; RANJBAR, R.; MORADI, P. Gene selection for microarray data classification using a novel ant colony optimization. **Neurocomputing**, [S.I.], v. 168, n. 6, p. 1024-1036, 2015.

THROSBY, T. Cultural Capital. **Journal of Cultural Economics**, [S.I.], v. 23, p. 3-12, 1999.

ULUTAGAY, G.; KANTARCI, S. An Extension of Fuzzy L-R Data Classification with Fuzzy OWA Distance. **International Journal of Intelligent Systems**, [S.I.], v. 30, n. 9, p. 1006-1020, 2015.

VAPNIK, V. **The nature of Statistical Learning Theory**. [S.I.]: Springer Science & Business Media, 2013, 314 p.

VEIJA, J. L. B. C.; CARVALHO, A. A. de; SILVA, I. C. da; REBELLO, J. M. A. The use of artificial neural network in the classification of pulse-echo and TOFD ultra-sonic signals. **Journal of the Brazilian Society of Mechanical Sciences and Engineering**, [S.I.], v. 27, n. 4, p. 394-398, 2005.

WANG, J. A novel K-NN classification algorithm for privacy preserving in cloud computing. **Journal of Applied Sciences - Engineering and Technology**, [S.I.], v. 4, n. 22, p. 4865-4870, 2012.

WANG, K.-J.; ADRIAN, A.M.; CHEN, K.-H.; WANG, K.-M. An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. **Journal of Biomedical Informatics**, [S.I.], v. 54, p. 220-229, 2015.

WITTEN, I.H.; FRANK, E.; HALL, A. M. **Data Mining Practical Machine Learning Tools and Techniques**, 3rd ed., [S.I.]: Elsevier, 2011.

WU, Q.; JIA, C.; CHEN, W. A Novel Classification-Rejection Sphere SVMs for Multi-class Classification Problems. In: Third International Conference on Natural Computation (ICNC 2007), [S.I.], 2007, p. 34- 38.

XIE, J.; WANG C.; ZHANG, Y.; JIANG, S. Clustering *Support Vector Machines* for unlabeled data classification. In: 2009 International Conference on Test and Measurement, [S.I.], 2009, p. 34-38.

XU, Z.; WATADA, J.; WU, M.; IBRAHIM, Z.; KHALID, M. Solving the imbalanced data classification problem with the particle swarm optimization based *Support Vector Machine*. **IEEJ Transactions on Electronics - Information and Systems**, [S.I.], v. 134, n. 6, p. 788-795, 2014.

XUE, Z.; LUO, J.; LIU, W. QP Representable Mahalanobis Ellipsoidal Learning Machine for Imbalanced Data Classification. In: 2012 Fifth International Joint Conference on Computational Sciences and Optimization, [S.I.], 2012, p. 316-319.

YANG, X.; CAO, A.; SONG, Q.; SHAEFER, G.; SU, Y. Vicinal support vector classifier using supervised kernel-based clustering. **Artificial Intelligence in Medicine**, [S.I.], v. 60, n. 3, p. 189-196, 2014.

ZHANG, B.; LU, J.; GUO, H.; XU J.; ZHAO, C. Object-Oriented Change Detection for Multi-Source Images Using Multi-Feature Fusion. In: 2016 3rd International Conference on Artificial Intelligence and Pattern Recognition, [S.I.], 2016, p. 75-80.

ZHANG, B.; LU, J.; GUO, H.; XU, J.; ZHAO, C. Object-oriented change detection for multi-source images using multi-feature fusion. In: 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), [S.I.], 2016, p. 1-6.

ZHANG, X.; DING, S.; SUN, T. Multi-class LSTMSVM based on optimal directed acyclic graph and shuffled frog leaping algorithm. **International Journal of Machine Learning and Cybernetics**, [S.I.], v. 7, n. 2, p. 241-251, 2016.

Apêndice A – Detalhamento dos procedimentos de limpeza e integração dos dados

O detalhamento das duas primeiras etapas previstas no processo de KDD foram conduzidas neste estudo de forma conjunta e serão apresentadas neste apêndice. A seção 4.1 apresentou como a base de dados da POF está estruturada e, a partir disso, nota-se que os dados estão segmentados em diferentes tabelas. Dessa forma, a partir do objetivo deste estudo e da análise dos fatores que podem afetar a decisão de consumo de cinema pelos indivíduos, a tabela 12 – despesa individual pertencente a POF 4 foi utilizada como base de dados origem para composição do banco de dados, porém, esta tabela não possui todos os atributos necessários para treinamento e validação dos algoritmos, visto que possui somente os atributos referentes às despesas dos indivíduos. Os atributos relacionados aos aspectos sociais, socioeconômicos e características dos indivíduos, bem como de seus domicílios, estão contidos em outras tabelas da POF.

A partir disso, foi efetuada a integração de algumas das tabelas da POF para composição do banco de dados, através do comando *merge* do *software* Stata. O comando em questão se trata da junção da tabela, ou apenas atributos desta, atualmente utilizada no *software*, denominada pelo Stata como tabela *in memory* ou master banco de dados, com uma outra tabela a ser referenciada, denominada *using* banco de dados.

Dessa forma, após estudo e entendimento da estrutura dos dados da POF, foram considerados como campos-chave para identificação dos domicílios e indivíduos, de forma respectiva, os seguintes atributos, cujas descrições foram obtidas através do arquivo “Descrição dos Registros da POF 2008-2009” da POF, pertencente a documentação dos microdados e disponível no site do IBGE:

1 Atributos de identificação dos domicílios

Os seguintes atributos foram considerados para identificação dos domicílios:

- Código da UF (campo: *cod_uf*), atributo numérico de 02 dígitos que identifica a Unidade da Federação.
- Número sequencial (campo: *num_seq*), atributo numérico de 03 dígitos que identifica o número sequencial de cada setor da amostra de domicílios da POF.

- DV do sequencial (campo: *num_dv*), atributo numérico de 01 dígito que identifica o verificador do Número sequencial de cada setor da amostra de domicílios da POF.
- Número do domicílio (campo: *cod_domc*), atributo numérico de 02 dígitos que identifica o domicílio, foi atribuído sequencialmente a cada domicílio selecionado em cada setor da amostra da POF.
- Estrato geográfico (campo: *num_est_renda*), atributo numérico de 02 dígitos que identifica os estratos do Plano Amostral da pesquisa: estratificações Geográfica e Estatística.
- Fator de expansão (campo: *fator_expansao1*), atributo numérico de 14 dígitos com até 8 casas decimais que identifica o fator de expansão atribuído ao domicílio pertencente a um setor. Dessa forma, cada domicílio corresponde a um fator de expansão, cujo valor é também atribuído para as unidades de consumo que o pertencem.

2 Atributos de identificação dos indivíduos

Somados aos atributos descritos acima para identificação dos domicílios, foram também considerados os seguintes atributos para identificação dos indivíduos:

- Número da UC (campo: *num_uc*), atributo numérico de 01 dígito que identifica a unidade de consumo (UC) pertencente ao domicílio pesquisado, variando, assim, de 1 até o número total de unidades de consumo do domicílio.
- Número do informante (campo: *num_inf*), atributo numérico de 02 dígitos para identificação do número de ordem no qual o morador foi registrado no quadro 3 do POF1. Conforme o arquivo de descrição dos registros da POF, este número é utilizado para identificar o informante em todos questionários da pesquisa.

A partir disso, considerando os atributos previamente descritos para composição dos campos-chave de identificação dos domicílios e dos indivíduos, foi executado o comando *merge* da tabela 12- Despesas individuais com as seguintes tabelas:

- 01 – Domicílio, para obtenção das características socioeconômicas, sociais, geográficas, bem como sobre a infraestrutura do domicílio a que o indivíduo pertence.
- 02 – Pessoas, para obtenção das características pessoais dos indivíduos.

- 04 – Condições de vida, para obtenção de atributos indicadores das condições de vida dos indivíduos.
- 14 – rendimentos e deduções, para obtenção das informações referentes aos rendimentos e vínculo empregatício dos indivíduos.

Para a execução do comando *merge* na ferramenta Stata, a tabela 12 da POF foi definida como *master* banco de dados (banco de dados *in memory*), a partir disso, através do menu *Data > Combine banco de dados > Merge two banco de dados*, foram definidos os seguintes parâmetros:

- Aba *Main*:
 - Tipo de *merge*, conforme detalhado no *Help* do Stata, o *software* disponibiliza quatro (04) tipos de junção de tabelas, sendo estes:
 - *One-to-one on key variables*, utilizado para combinação de um registro do *master* banco de dados (origem) com um registro do *using* banco de dados. Aplica-se em situações que os registros de ambas tabelas se referem a um mesmo evento, porém, possuem informações distintas.
 - *Many-to-one key variables*, utilizado para combinação de vários (muitos) registros do *master* banco de dados (origem) com um registro do *using* banco de dados. Aplicável em situações de dados hierárquicos, ou seja, um ou mais registros do banco de dados origem estão relacionados a um registro do *using* banco de dados.
 - *One-to-many on key variables*, o conceito é o mesmo apresentado para o *Many-to-one*, porém, neste caso, o *master* banco de dados contém um registro associado a muitos do *using* banco de dados.
 - *Many-to-many on key variables*, utilizado para combinação de vários (muitos) registros do *master* banco de dados (origem) com vários registros do *using* banco de dados. Aplicável em situações de dados hierárquicos, ou seja, um ou mais registros do banco de dados origem estão relacionados a um registro do *using* banco de dados. A partir disso, analisando os dados das tabelas da POF, notou-se que o tipo mais adequado de *merge* a ser executado é o *Many-to-one*, visto que o *master* banco de dados (Tabela 12 da POF, referente às despesas individuais), conforme descrito na seção 5.1, dispõe os registros por Unidade de Orçamento-Despesa, porém, não se tem um registro único por cada indivíduo entrevistado, visto que caso este tenha consumido mais de um produto durante o período de referência ou tenha

consumido um mesmo produto em locais de aquisição e formas de pagamentos distintas, tais registros são apresentados em linhas diferentes. Com isso, pode-se ter mais de um registro para uma mesma Unidade de Orçamento-Despesa. Já as demais tabelas a serem associadas ao *master* banco de dados apresentam um registro, ou seja, linha da tabela, por domicílio (Tabela 01 – Domicílio) ou por indivíduo (Tabelas 02 – Pessoa, 14 – Condições de vida e 14 – Rendimentos e deduções). Dessa forma, tem-se o conceito de *Many-to-One*, em que os registros de um ou muitos itens consumidos pelas Unidades de Orçamento-Despesa presentes no *master* banco de dados (*Many*) estão relacionadas a indivíduos que pertencem a um único domicílio (*One*) e que possuem características pessoais únicas (*One*), presentes em cada *using* banco de dados.

- Atributos-chave, que se tratam das variáveis selecionadas para identificação e associação dos registros entre as tabelas. Dessa forma, neste parâmetro foram inseridas as variáveis detalhadas no início desta seção.
- *Filename of* banco de dados *on disk*, indicação da localização do *using* banco de dados.

A figura 27 apresenta a aba *Main* do Stata contendo os parâmetros supracitados.

Figura 27 – Aba *Main* do *software* Stata para configuração dos parâmetros do comando *merge*

Fonte: *Software* Stata, 2017

- Aba *Options*: Nesta aba, são definidos os campos da tabela a ser efetuada a junção que se deseja associar a tabela origem (*Variables to keep from* banco de dados *on disk*), dessa forma, para cada junção foram considerados todos os atributos de cada tabela para posterior análise e limpeza, bem como possibilita a criação de uma nova variável contendo o resultado do comando por registro da tabela. Dessa forma, segundo o manual do *software* Stata, é possível visualizar o detalhamento dos resultados da junção entre as tabelas, visto que a variável apresentará observações de que o *merge* entre as tabelas ocorreu com sucesso ou se não foi possível efetuar a junção destas, pois o registro foi identificado em apenas uma delas. Os tipos de resultados a serem obtidos são definidos pelo usuário. A figura 28 apresenta a aba *Options* do Stata contendo os parâmetros descritos.

Figura 28 – Aba *Options* do *software* Stata para configuração dos parâmetros do comando *merge*

Fonte: *Software* Stata, 2017

- Aba *Results*, utilizada para definir o detalhamento dos resultados que se deseja obter com a execução do comando *merge*. Neste trabalho, foram considerados os parâmetros *Record appeared in master only* e *Record appeared in both*, na seção *Results to keep*, que irá manter além dos registros em que a junção ocorreu, todos os registros do *master* banco de dados, mesmo que a junção com as demais tabelas não aconteça, ou seja, caso não tenham registros em comum a serem associados. Esta decisão foi tomada com o objetivo de manter todos os registros de consumo inicialmente presentes na tabela 12 da POF, possibilitando, assim, avaliações posteriores da necessidade de limpeza dos dados, caso haja registros em branco. A figura 29 apresenta a aba *Results* do Stata contendo o parâmetro descrito.

Figura 29 – Aba *Results* do *software* Stata para configuração dos parâmetros do comando *merge*

Fonte: *Software* Stata, 2017

Dessa forma, as etapas apresentadas foram executadas quatro vezes, executando-se, assim, o comando *merge* iniciado com o *master* banco de dados com a tabela 12 da POF, e cada tabela gerada com o resultado do comando, foi sendo considerada como *master* banco de dados das demais execuções, até que se juntaram os campos de cada uma das tabelas detalhadas na seção 4.1.

Adicionalmente, notou-se que a POF apresenta as informações de despesa somente por item, com o objetivo de compor um valor total de despesa do indivíduo com Cultura e Recreação no período de referência da pesquisa, 30 dias, foi utilizado o pacote Data Zoom para composição deste atributo. O Data Zoom, cujo acesso é aberto, foi desenvolvido pelo Departamento de Economia da PUC-Rio com financiamento da FINEP, para disponibilização gratuita de pacotes em Stata para a leitura dos microdados das pesquisas domiciliares do IBGE, o que inclui a POF 2008-2009. Para geração de tais bases, requer que se tenha os dados originais, bem como o programa disponível no próprio *website* do Data Zoom para *download* (<http://www.econ.puc-rio.br/datazoom/index.html>).

A partir disso, para composição do atributo de despesa total, foi instalado o programa seguindo as instruções disponibilizadas no Website previamente citado e foi executado o comando `db datazoom_pof2008` no Stata para execução do programa. O Data Zoom disponibiliza três funcionalidades, sendo estas:

- bases Padronizadas, que possibilita a construção de uma base com gastos em itens agregados, seguindo as tabulações do IBGE;
- gastos selecionados, que possibilita a personalização da cesta de itens a partir de suas categorias, bem como a geração de gastos a partir de itens desagregados; e
- tipos de registros, que possibilita a extração da base dos dados originais sem modificações.

A figura 30 apresenta o menu inicial do Data Zoom para a POF 2008-2009 no Stata, após a execução do comando supracitado e contendo tais opções previamente descritas. Para a composição do valor agregado (total) de despesa com Cultura e Recreação dos indivíduos, utilizou-se a opção de Gastos Selecionados, considerando, assim, os seguintes parâmetros:

- Tipo de identificação: Unidade de consumo.
- Itens à cesta: Outros gastos.
- Itens selecionados: Recreação e cultura, o que também inclui o produto CINEMA (INGRESSO), objeto deste trabalho.

Figura 30 – Tela inicial para acesso aos microdados da POF do pacote Data Zoom no Stata

A figura 31 apresenta o detalhamento de tais parâmetros definidos no Stata.

Após a geração do atributo correspondente ao valor total de despesa com cultura e recreação, executou-se a junção deste ao *master* banco de dados, também através do comando *merge* do Stata, conforme detalhado nas seções anteriores.

Figura 31 – Tela do pacote Data Zoom para configuração da cesta de despesas para geração dos valores agregados no Stata.

Fonte: *Software Stata*, 2017

Inicialmente, a tabela 12 da POF sem modificações possuía um total de 1.365.107 registros e 26 atributos. Após a execução dos procedimentos de integração do banco de dados, que consistiu na sua junção com as quatro tabelas apresentadas anteriormente, bem como com o campo de valor total de despesas com recreação e cultura gerado, a quantidade de registros se manteve, e passou a totalizar 189 atributos.

Apêndice B – Detalhamento dos procedimentos de transformação dos dados

Ao analisar os registros do banco de dados utilizado neste trabalho, notou-se que alguns destes possuem diversas categorias, que não apresentam relação do ponto de vista de significado uma em relação a outra, bem como a existência de variáveis referentes a elevados valores financeiros, podendo gerar problemas dimensionais de valores discrepantes, ambos podendo, assim, afetar o desempenho dos algoritmos de classificação. A partir disso, optou-se por efetuar as seguintes transformações no banco de dados: criação de variáveis binárias e a conversão de valores monetários para os atributos detalhados a seguir.

1 Criação de variáveis binárias

Foram criadas as seguintes variáveis binárias:

- Classificador, o campo em questão foi criado a partir dos seguintes campos “código do item” (campo: *cod_item*), que, conforme o arquivo de descrição dos registros da POF, identifica o produto consumido pela Unidade de Consumo, bem como pelo campo “código do local de aquisição” (campo: *cod_local_compra*), que identifica o tipo de local em que o produto foi adquirido. Para criação do campo binário, considerou-se: se o atributo referente ao código do item possuir valor que corresponde ao consumo do produto “CINEMA (INGRESSO)” e o atributo código do local de aquisição for igual a um dos locais listados na seção 5.1 (Filme), caso contrário (Não filme).
- Condição de orçamento trabalho e/ou rendimento (campo: *cod_sit_receita*), que, conforme o arquivo de descrição dos registros da POF, identifica se o indivíduo a partir dos 10 anos de idade é uma Unidade de Orçamento Trabalho e/ou Rendimento, possuindo como possíveis categorias: (-1) Não Aplicável, (01) Sim, (02) Não, (03) Recusa e (04) Outro. Para criação do campo binário, estes foram convertidos para os seguintes valores: Sim (1), demais categorias (0).
- Tipo de domicílio, a partir do campo “tipo de domicílio” (campo: *cod_tipo_domc*), que, conforme o arquivo de descrição dos registros da POF, identifica o tipo de domicílio, podendo assumir as seguintes categorias: (01) Casa, (02) Apartamento e (03) Cômodo. Para criação do campo binário, estes foram convertidos para os seguintes valores: Se categorias 01 e 02 (1), Categoria 03 (0).

- Condição precária, a partir do campo indicador da condição de vida da família (campo: *cod_cond_vida*), que, conforme o arquivo de descrição dos registros da POF, identifica a avaliação da família sobre como seu rendimento total permite levar a vida até o fim do mês, possuindo como possíveis categorias: (0) Não respondeu, (1) Muita dificuldade; (2) Dificuldade, (3) Alguma dificuldade, (4) Alguma facilidade, (5) Facilidade e (6) Muita facilidade. Para criação do campo binário, estes foram convertidos para os seguintes valores: se categorias 1, 2 ou 3 (1), demais categorias ou não respondeu (0);
- Unidades da Federação (Estados brasileiros), que correspondem a 27 campos binários criados um por estado do Brasil, a partir do campo “código da unidade da Federação” (campo: *cod_uf*), que, conforme o arquivo de descrição dos registros da POF, identifica o código das unidades da federação, tais códigos se encontram detalhados na seção 4.1. Para criação dos campos binários por estado, estes foram convertidos para os seguintes valores: Se o valor do campo “código da unidade da Federação” corresponde ao estado em questão (1), demais estados (0).
- Zona rural ou urbana, a partir dos campos “estrato geográfico” (campo: *num_ext_renda*) e campo “código da unidade da Federação” (campo: *cod_uf*), previamente descrito. O campo referente ao estrato geográfico, conforme o arquivo de descrição dos registros da POF, identifica os estratos do plano amostral da pesquisa por estratificações geográficas e estatísticas. A nível geográfico, a estratificação compreende na área urbana para o município da capital, resto da região metropolitana, resto da UF e área rural. Para criação do campo binário, foram convertidos os códigos de estrato geográfico conforme detalhados na tabela de estrato do plano amostral da POF, considerando cada estado da Federação, visto que esta atribuição é efetuada considerando o código da unidade da Federação, sendo, assim necessário executar a atribuição dos valores para cada um dos Estados. Sendo, por fim, convertidos para os seguintes valores: Urbana (1), Rural (0).
- Sexo, a partir do campo “código do sexo” (campo: *cod_sexo*) que, conforme o arquivo de descrição dos registros da POF, identifica o sexo dos indivíduos, podendo assumir as seguintes categorias: (01) Masculino e (02) Feminino. Para criação do campo binário, estes foram convertidos para os seguintes valores: Masculino (1), Feminino (0).

- Plano de saúde, a partir do campo “tem plano de saúde” (campo: *plano_saude*) que, conforme o arquivo de descrição dos registros da POF, identifica se o indivíduo possui plano ou seguro-saúde, podendo assumir as seguintes categorias: (01) Sim e (02) Não. Para criação do campo binário, estes foram convertidos para os seguintes valores: Sim (1), Não (0).
- Cor ou raça, a partir do campo “cor ou raça” (campo: *cod_cor_raca*) que, conforme o arquivo de descrição dos registros da POF, identifica a Cor ou Raça declarada pelo indivíduo, podendo assumir as seguintes categorias: (01) Branca, (02) Preta, (03) Amarela, (04) Parda, (05) Indígena e (09) Não Sabe. Para criação do campo binário, estes foram convertidos para os seguintes valores: Branco (1), Demais categorias (0).
- Religião, a partir do campo “código da religião” (campo: *religiao*), que, conforme o arquivo de descrição dos registros da POF, identifica a seita, culto ou ramo da religião declarada pelo indivíduo. Para criação do campo, utilizou-se a tabela Cadastro de Religião POF 2008-2009, contendo a descrição de todos os códigos de religião, sendo estes convertidos para os seguintes valores: Católico (1), demais religiões (0);
- Sabe ler e escrever, a partir do campo “sabe ler e escrever” (campo: *cod_sabe_ler*) que, conforme o arquivo de descrição dos registros da POF, identifica se o indivíduo sabe ler e escrever, podendo assumir as seguintes categorias: (01) Sim e (02) Não. Para criação do campo binário, estes foram convertidos para os seguintes valores: Sim (1), Não (0).
- Estudante, a partir do campo “frequência à escola ou creche” (campo: *cod_freq_escola*), que, conforme o arquivo de descrição dos registros da POF, identifica se o indivíduo frequentava escola ou creche no período da pesquisa, podendo assumir as seguintes categorias: (01) Sim, Rede Particular (02) Sim, Rede Pública, (03) Não, mas já Frequentou, (04) Nunca Frequentou. Para criação do campo binário, estes foram convertidos para os seguintes valores: Se categorias 01 ou 02 (1), Demais categorias (0).
- Nível de instrução, para identificar se o indivíduo possui ensino superior, a partir do campo “curso mais elevado que frequentou” (campo: *cod_niv_instr*), que, conforme o arquivo de descrição dos registros da POF, identifica o curso mais elevado que o indivíduo frequentou, podendo assumir as seguintes categorias: (-1)

Não Aplicável, (01) Creche, (02) Pré-escolar, (03) Classe de Alfabetização de Crianças, (04) Alfabetização de Adultos, (05) Antigo Primário, (06) Antigo Ginásio, (07) Antigo Clássico, Científico, etc. (08) Regular do Ensino Fundamental, (09) Educação de Jovens e Adultos ou Supletivo do Ensino Fundamental ou 1º grau, (10) Regular do Ensino Médio, (11) Educação de Jovens e Adultos ou Supletivo do Ensino Médio ou 2º grau, (12) Tecnológico Superior, (13) Pré-Vestibular, (14) Superior-Graduação, (15) Especialização Superior e (16) Mestrado ou Doutorado. Para criação do campo binário, estes foram convertidos para os seguintes valores: Se categorias 12m 14, 15 ou 16 (1), demais categorias (0).

2 *Conversão de valores monetários*

Visando evitar problemas dimensionais de valores discrepantes, os atributos abaixo foram alterados considerando a seguinte fórmula: $\log(\text{variável} + 1)$.

- Renda bruta monetária, a partir do campo Renda monetária mensal do domicílio (campo: *renda_bruta_monetaria*) que, conforme o arquivo de descrição dos registros da POF, identifica o valor em reais do rendimento bruto monetário mensal do domicílio, obtido através do somatório dos rendimentos brutos monetários mensais de todos os moradores do domicílio através do trabalho, transferências e outras rendas, acrescido da variação patrimonial, que compreende vendas de imóveis, recebimentos de heranças e o saldo positivo da movimentação financeira.
- Renda per capita, a partir do campo Renda per capita da unidade de consumo (campo: *renda_percapita*) que, conforme o arquivo de descrição dos registros da POF, identifica o valor da renda total mensal da unidade de consumo dividida pelo número de integrantes da unidade de consumo.

3 *Procedimentos adicionais de transformação dos dados*

Após a criação e modificação de tais variáveis, foi efetuada a exclusão de registros duplicados, visto que como se não tem o atributo correspondente a quantidade dos itens consumidos por cada indivíduo, bem como as variáveis de forma de pagamento e locais de compra não são relevantes para a etapa seguinte do trabalho, Mineração de Dados, notou-se a

existência de registros de um mesmo item consumido por um mesmo indivíduo em linhas de registros distintas. Dessa forma, todos os registros com tais características foram removidos do banco de dados através da função *duplicates drop* do Stata, considerando as variáveis de identificação dos indivíduos anteriormente descritas, bem como o atributo referente aos produtos consumidos para remoção dos registros duplicados.

Adicionalmente, as variáveis utilizadas para identificação dos indivíduos e domicílios, conforme descritas na seção 5.3.1, com exceção do atributo correspondente ao Fator de Expansão, bem como os atributos originais que suportaram as etapas previamente descritas foram excluídos do banco de dados, através da função *drop* do Stata, obtendo-se, assim, o banco de dados final contendo 128.300 registros e 51 atributos. Este total de atributos se refere a tais atributos criados e modificados, bem como os seguintes atributos, que concluímos não ser necessário transformá-los:

- Fator de expansão (campo: *fator_expansao1*), que conforme descrito na seção 5.3.1, identifica o fator de expansão atribuído ao domicílio pertencente a um setor. Cada domicílio corresponde a um fator de expansão, o qual é também atribuído para as suas unidades de consumo.
- Quantidade de moradores (campo: *qtd_morador_domc*), que, conforme o arquivo de descrição dos registros da POF, identifica o número total de pessoas que residem no domicílio.
- Quantidade de unidades de consumo (campo: *qtd_uc*), que, conforme o arquivo de descrição dos registros da POF, identifica o número total de unidades de consumo do domicílio
- Quantidade de famílias (campo: *qtd_familia*), que, conforme o arquivo de descrição dos registros da POF, identifica o número total de famílias no domicílio.
- Quantidade de cômodos do domicílio (campo: *qts_comodos_domc*), que, conforme o arquivo de descrição dos registros da POF, identifica o total de cômodos existente no domicílio.
- Quantidade de banheiros do domicílio (campo: *qtd_banheiros*), que, conforme o arquivo de descrição dos registros da POF, identifica o número de banheiros ou sanitários localizados dentro da propriedade e de uso exclusivo dos moradores do domicílio.
- Idade calculada em anos (campo: *idade_anos*), que, conforme o arquivo de descrição dos registros da POF, refere-se a idade do indivíduo em anos completos.

- Anos de estudos (campo: *anos_de_estudo*), que, conforme o arquivo de descrição dos registros da POF, identifica a idade do morador em anos completos.
- Número de horas trabalhadas (campo: *num_horas_trabalhadas*), que, conforme o arquivo de descrição dos registros da POF, corresponde ao número total de horas habitualmente trabalhadas por semana referente ao trabalho informado pelo indivíduo.
- Despesa total em Recreação e cultura (campo: *v_DT_8_tot*), valor total agregado de despesa com recreação e cultura gerado através do programa Data Zoom por indivíduo.

O banco de dados final obtido foi exportado, considerando todos os seus atributos, da ferramenta Stata através do comando *export* para um arquivo no formato CSV (*Comma Separated Value*).