UNIVERSIDADE FEDERAL DO CEARÁ

***CAMPUS* SOBRAL**

CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

LUCINARA KECIA SILVA FERNANDES

APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DE PREDISPOSIÇÃO AO MEDO DO CRIME

SOBRAL

2022

LUCINARA KECIA SILVA FERNANDES

APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DE PREDISPOSIÇÃO AO MEDO DO CRIME

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Carlos Alexandre Rolim Fernandes

Coorientador: Prof. Dr. James Ferreira Moura Júnior

SOBRAL

2022

LUCINARA KECIA SILVA FERNANDES

APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DE PREDISPOSIÇÃO AO MEDO DO CRIME

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Carlos Alexandre Rolim Fernandes (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. James Ferreira Moura Júnior (Coorientador)
Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab)

Prof. Dr. Iális Cavalcante de Paula Júnior
Universidade Federal do Ceará (UFC)

Prof. Me. David Nascimento Coelho
Universidade Federal do Ceará (UFC)

Aos meus pais, que estiveram presentes ao longo desta caminhada me auxiliando com muito amor e dedicação.

AGRADECIMENTOS

Ao Prof. Dr. Carlos Alexandre por me orientar em meu trabalho de conclusão e durante todos os passos do desenvolvimento da pesquisa.

Ao Prof. Dr. James, por me orientar nos aspectos relacionados a área de psicologia abordados neste trabalho.

À Banca Examinadora pela disposição em contribuir na homologação do trabalho proposto e sugestões em prol de melhorias.

À todos os professores da UFC por me proporcionar conhecimento e incentivo à minha formação.

Às amigas e companheiras de curso Beatriz Martins, Kamila Amélia e Thaís Félix pelo apoio durante e além da tragetória acadêmica.

À amiga Larissa Teixeira e ao companheiro Almir Rodrigues por todo apoio emocional e técnico, me motivando sempre em prol da conclusão deste ciclo.

Aos meus pais, Claudia Fernandes e Francisco Fernandes, e irmã, Tainara Fernandes, por serem sempre meu alicerce e apoio incondicional em todos os momentos da minha vida, a eles devo todo meu sucesso.

"Science is much more than a body of knowledge. It is a way of thinking. This is central to its success. Science invites us to let the facts in, even when they don't conform to our preconceptions."

(Carl Sagan)

RESUMO

Diante do quadro de violência vivenciado pela população brasileira, fatores característicos como aspectos sociodemográficos e propensão à posições autoritárias impactam na predisposição ao medo do crime por parte dos cidadãos. Perante essa conjuntura, o Aprendizado de Máquina se mostra uma ferramenta útil na análise dessas relações, por já ser um artifício cada vez mais aplicado à dados sociais no contexto de predição. Com base nisso, o objetivo deste trabalho está na verificação dos melhores modelos para análise de dados e técnicas de Aprendizado de Máquina (AM), nos quais os níveis de indicadores de medo do crime são previstos, e na análise de quais atributos são mais relevantes para a previsão da predisposição ao medo do crime, utilizando o banco de dados da pesquisa intitulada "Medo da violência e o apoio ao autoritarismo no Brasil". Coordenada pelo Fórum Brasileiro de Segurança Pública, os dados da pesquisa foram coletados em formato de questionário com assertivas de cunho sociodemográfico, relacionadas a situações de vivência como vítima de crimes e propensão ao apoio a posições autoritárias. Como metodologia de desenvolvimento, é abordado no trabalho a utilização do método *Knowledge Discovery in Databases* (KDD), partindo da análise dos dados coletados na pesquisa à disposição de simulações com 3 cenários de dados propostos e com combinações de classificadores, sendo eles *Support Vector Machine* (SVM), *Random Forest* (RF) e *K-Nearest Neighbors* (KNN). Ainda nas etapas do método, é realizada análise de atributos, com os algoritmos *Sequential Forward Select* (SFS) e *Sequential Backward Selection* (SBS), bem como aplicação da técnica de redução de dimensionalidade *Principal Component Analysis* (PCA). Em síntese, a melhor acurácia foi obtida utilizando os dados normalizados, o algoritmo SFS e o classficador SVM. Além disso, são observados como atributos mais importantes, a partir do cálculo dos coeficientes de Gini e Entropia, os relacionados a idade, escolaridade e índices sintéticos obtidos das escalas propostas pela pesquisa.

Palavras-chave: Aprendizado de Máquina. KDD. Dados Sociais. Crime. Autoritarismo

ABSTRACT

Given the situation of violence experienced by the Brazilian population, characteristic factors such as sociodemographic aspects and propensity to authoritarian positions impact the predisposition to fear of crime on part of citizens. Given this context, Machine Learning seems to be a useful tool in the analysis of these relationships, as it is already an artifice increasingly applied to social data in the context of prediction. Based on this, the objective of this work is to verify the best models for data analysis and Machine Learning (ML) techniques, in which the levels of fear of crime indicators are predicted, and in the analysis of which attributes are most relevant for predicting the predisposition to fear of crime, above the research database entitled "Fear of violence and support for authoritarianism in Brazil". Coordinated by the Brazilian Public Security Forum, the research data were collected in a questionnaire format with sociodemographic assertions, related to situations of living as a victim of crimes and propensity to support authoritarian positions. As a development methodology, the use of the Knowledge Discoveryin Databases (KDD) method is considered in the work, starting from the analysis of the data collected in the research available and simulations with 3 proposed data scenarios and with combinations of classifiers, namely Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbors (KNN). Still in the steps of the method, an analysis of attributes is performed, with the algorithms Sequential Forward Select (SFS) and Sequential Backward Selection (SBS), as well as application of the technique of dimensionality reduction Principal Component Analysis (PCA). In summary, the best results are verified, using accuracy values as a performance metric, in the model formed by applying the normalized data to the SFS algorithm and soon after to the SVM classifier. In addition, the most important attributes, based on the calculation of the Gini and Entropy coefficients, are those related to age, education and synthetic indices obtained from the scales proposed by the research.

Keywords: Machine Learning. KDD. Social Data. Crime. Authoritarism

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
KDD	Descoberta de Conhecimento em Bases de Dados, *Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbors*
PCA	*Principal Component Analysis*
RF	*Random Forest*
SBS	*Sequential Backward Selection*
SFS	*Sequential Forward Select*
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

O Brasil vive o drama da violência, que gera por volta de 60 mil mortes intencionais por ano e faz com que mais de 50 milhões de pessoas que compõem a população adulta do país conheçam pessoas que foram assassinadas (G1, 2017). Motivada por um fator gestão do medo e pelo sentimento de insegurança das populações das cidades, é evidente a promoção de um cenário que favorece o apoio da população a posições mais autoritárias (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018).

O medo do crime, seja ele proveniente de conflitos dentre relações sociais ou de como é pautada a atuação do Estado na acareação de violações, está associado à uma série de déficits civis e democráticos (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018). Por isso, fatores característicos como aspectos sociodemográficos e propensão à posições autoritárias impactam de forma significativa na predisposição do indivíduo a esse sentimento. Diante disso, a relevância do debate acerca de elementos que podem vir a relacionar o contexto sócio-político brasileiro e a violência é considerável.

Nesse contexto, a inteligência artificial apresenta-se como uma ferramenta útil e promissora a ser aplicada. Por meio de uma ampla gama de aplicações, os métodos de Aprendizagem de Máquina (AM) ligados a estruturas de descoberta de conhecimento estão influenciando diversos domínios que afetam a vida das pessoas, seja por meio de seus novos algoritmos e teorias especializadas ou por fatores além desses (RUDIN; WAGSTAFF, 2014). Em casos relacionados a pesquisas que envolvem coleta de dados a partir da aplicação de questionários ou formulários, técnicas de AM estão sendo utilizadas por pesquisadores e cientistas sociais para vários aspectos, incluindo processamento de dados no contexto de predição (BUSKIRK *et al.*, 2018).

A presente pesquisa consiste no estudo e na análise de algoritmos de aprendizado de máquina a serem aplicados no banco de dados da pesquisa intitulada “Medo da violência e o apoio ao autoritarismo no Brasil”, coordenada pelo Fórum Brasileiro de Segurança Pública, em parceria com universidades públicas e realizada pelo Instituto Datafolha, no ano de 2017. Envolvendo 2.087 pessoas de todas as regiões do país, a pesquisa coletou dados a partir da aplicação de um questionário com assertivas relacionadas a escalas utilizadas para mensurar medo do crime, vitimização do crime, chances de ocorrência de crimes e propensão ao apoio a posições autoritárias, além de questões de cunho sociodemográfico e acesso à bens e serviços (FILHO *et al.*, 2018).

A partir do estudo detalhado dessa base de dados e da aplicação de técnicas de

AM, são destacados propostas de melhores modelos para previsão de predisposição ao medo do crime, tendo como referencial valores de um índice sintético que visa conhecer o grau de medo da população brasileira em relação à uma série de eventos (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018), como forma de colaborar com setores públicos contribuindo na construção de soluções e projetos. Além da proposição do modelo de previsão, este projeto também evidencia quais os fatores que mais influenciam na previsão da predisposição ao medo do crime. Fazendo uso do processo de KDD, são dispostas simulações com 3 cenários de dados propostos e com combinações de classificadores, sendo eles *Support Vector Machine* (SVM), *Random Forest* (RF) e *K-Nearest Neighbors* (KNN), e usando estratégias de análise de atributos, sendo tais *Sequential Forward Select* (SFS), *Sequential Backward Selection* (SBS), bem como técnica de redução de dimensionalidade, sendo ela *Principal Component Analysis* (PCA). Com relação ao estudo dos fatores que mais influenciam na previsão, são usados os parâmetros Coeficiente de Gini e Ganho de Informação. Para validação dos modelos, é utilizado o método K-fold, com k = 10.

Este trabalho foi desenvolvido em 7 capítulos. No Capítulo 2 são expostos o objetivo geral e os objetivos específicos, que sintetizam o que se pretende alcançar no trabalho. O Capítulo 3 apresenta os fundamentos teóricos necessários para melhor compreensão dos aspectos do estudo. No Capítulo 4, trabalhos relacionados ao contexto ao qual o trabalho está pautado são citados e comentados. A explanação de como é disposta a metodologia encontra-se no Capítulo 5, sendo dividida em três seções referentes à base de dados, à aplicação do método KDD e à importância dos atributos. Já no Capítulo 6, são apresentados os resultados obtidos, seguidos de discussões acima do que é apurado pela execução da metodologia proposta. Por fim, são mostradas no Capítulo 7 as considerações finais do estudo, assim como propostas para trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo Geral

O objetivo do estudo consiste na identificação de melhores modelos de análise de dados e técnicas de AM, dentre os propostos, para previsão de níveis indicados de predisposição ao medo do crime, bem como na análise de quais atributos são os mais relevantes para a previsão da predisposição ao medo do crime.

2.2 Objetivos Específicos

- Aplicar a metodologia KDD no processo de desenvolvimento dos modelos preditivos propostos.
- Realizar simulações a partir de cenários propostos em prol da verificação da viabilidade da aplicação dos métodos de AM aplicados.
- Identificar, a partir de análises, as variáveis mais relevantes a serem consideradas.
- Gerar modelos de predição do grau relativo dos índices de medo do crime.
- Analisar o desempenho dos modelos estudados a partir de técnicas de validação cruzada.

3 FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão apresentados os aspectos teóricos fundamentais ao desenvolvimento do estudo, sendo eles os Índices de propensão ao apoio a posições autoritárias, de medo do crime, de vitimização do crime e de chances de ocorrência de crimes, o Critério Brasil de Classes, a Descoberta de Conhecimento em Bases de Dados (KDD), Aprendizagem de Máquina (AM) e os classificadores utilizados, como também o método Análise de Componentes Principais (PCA) e o Ganho de informação e Coeficiente de Gini.

3.1 Índice de Propensão ao apoio a Posições Autoritárias

O Índice de Propensão ao apoio a Posições Autoritárias é calculado a partir de uma escala psicométrica de 17 assertivas com listagem estruturada segundo estudo feito por (CROCHIK, 2017) acima da escala F de Adorno (ADORNO *et al.*, 1950), projetada para medir tendência a atitudes antidemocráticas implícitas na personalidade de indivíduos e distribuídas em dimensões como submissão à autoridade, agressividade autoritária e convencionalismo. As assertivas selecionadas por Crochik seguem cargas fatoriais de forma equilibrada de acordo com as sub-dimensões conceituadas como fenômenos associados a posturas autoritárias, onde um ranking de 1 a 6 pontos, sendo quanto mais próximo de 1 menor o apoio a posições autoritárias e quanto mais próximo de 6 maior a adesão e apoio a elas. Tais assertivas estão colocadas na quarta seção no Anexo A.

3.2 Índices de Medo, Vitimização e de Chances de ocorrência de crime

Os Índices de Medo, Vitimização e de Chances de ocorrência de crime tratam-se de métrica obtidas a partir da listagem de itens, pautados dentro da análise realizada pela pesquisa "Medo da violência e o apoio ao autoritarismo no Brasil", do Fórum Brasileiro de Segurança Pública e do Instituto Datafolha, e estudada em (FILHO *et al.*, 2018). Os itens são ilustrados em formato de assertivas dispostos em escala, onde um índice sintético é calculado a partir da média aritmética dos valores obtidos. O formulário utilizado na pesquisa está disponível no Anexo A, onde as assertivas estão organizadas pelas seções respectivas.

O Índice de medo do crime é obtido da listagem de 16 itens relacionados ao grau de medo da população frente a situações criminais, ilustrados em formato de assertivas com possibilidades de respostas "Sim" e "Não" (correspondendo aos valores numéricos 1 e 0, res-

pectivamente). O Índice de Vitimização do Crime é obtido a partir de uma lista de 15 assertivas sobre que tipo de crime o indivíduo teria sido vítima no mês anterior (mesmos itens da escala de medo do crime, exceto menções de itens vitimizados por homicídio). Já o Índice de Chances de ocorrência de crime é proveniente de uma escala de probabilidade de ocorrência de vitimização por crime, composta por 16 itens, representando a probabilidade de um indivíduo vir a sofrer um crime no próximo mês (por exemplo, "Ter sua residência invadida ou arrombada no próximo mês", "De sofrer sequestro relâmpago no próximo mês"), respondida através de uma escala *Likert*, variando de 0 (nenhuma chance de acontecer) a 10 (alta chance de acontecer).

3.3 Critério Brasil de Classificação Econômica

O Critério Brasil de Classificação Econômica (ABEP, 2017) é um padrão de classificação socioeconômica desenvolvida pela Associação Brasileira de Empresas de Pesquisa (ABEP), pautada na mensuração do poder de compra da população, tornando possível a estratificação dos indivíduos nas classes A, B1, B2, C1, C2 e D/E. Tal fatoração considera aspectos como bens de consumo, estrutura física da residência e escolaridade dos membros da família, através de um sistema de pontuação obtido através de assertivas com valores tabelados, de acordo com o documento descritivo publicado pelo órgão.

3.4 Descoberta de Conhecimento em Bases de Dados (KDD)

A partir de conceitos propostos inicialmente por Fayyad, Piatetsky-Shapiro e Smyth em 1996, o KDD se refere a um processo interativo de identificação de novos padrões em dados que sejam válidos, novos, potencialmente úteis e interpretáveis (FAYYAD *et al.*, 1996). O processo KDD envolve uma série de passos de acordo com decisões tomadas pelos usuários, porém conta com um fluxo básico de fases, sendo elas:

- Seleção: agrupamento e seleção dos dados a serem analisados;
- Pré-processamento: operações básicas de limpeza dos dados, remoção de ruído, definições de estratégias para lidar com campos omissos, etc;
- Transformação: seleção de atributos úteis visando a efetivação do objetivo central da análise, podendo conter aplicação de técnicas de redução de dimensionalidade;
- Mineração de dados: levantamento e aplicação de algoritmos de mineração de dados como Classificação, Regressão, Sumarização, etc;

- Interpretação: análise exploratória dos padrões minerados visando a verificação do conhecimento descoberto, sendo válido o retorno a fases anteriores para fins de consolidação.

3.5 Aprendizagem de Máquina (AM)

A AM se trata de um campo de estudo direcionado à análise e implementação de algoritmos capazes de operar na construção, de forma indutiva, de modelos de previsão ou decisão a partir de dados. De acordo com Mitchell, considerar que um programa de computador aprende com certa experiência vinculada a alguma classe de tarefas e métricas de performance é afirmar que o desempenho das tarefas nessa classe, mensurado por tais métricas, evolui com a experiência (MITCHELL, 1997).

Os métodos de AM são categorizados em 3 abordagens: aprendizado supervisionado, onde são observados exemplos de pares entrada-saída, aprendendo uma função de mapeamento; aprendizado não-supervisionado, onde são assimilados padrões na entrada, mesmo sem nenhum retorno explícito fornecido; e aprendizado supervisionado por reforço, que ocorre através de uma série de recompensas ou punições (RUSSELL; NORVIG, 2010).

O AM apresenta modelos de aprendizagem dentro das seguintes tarefas:

- Classificação, que se baseia na previsão de categoria de uma observação dada;
- Regressão, onde é estimado uma valor numérico de uma observação;
- Agrupamento, onde são agrupadas observações em "clusters".

No trabalho serão utilizados modelos dentro da abordagem de Classificação.

3.6 Classificadores

3.6.1 Support Vector Machine (SVM)

O SVM consiste em uma técnica de classificação supervisionada não paramétrica (RUSSELL; NORVIG, 2010) binária. O algoritmo SVM funciona a partir da identificação de um hiperplano em um espaço N-dimensional, onde N é o número de atributos que classifica de forma distinta os pontos dos dados.

3.6.2 K-Nearest Neighbors (KNN)

O método KNN é também uma técnica de classificação supervisionada, não paramétrica. O algoritmo assume que todas as instâncias correspondem a pontos no espaço N-dimensional, onde os ”vizinhos mais próximos” são definidos em termos relativos à distância (MITCHELL, 1997).

3.6.3 Árvore de Decisão

Classificadores baseados em árvores de decisão utilizam-se de algoritmos que subdividem gradualmente os dados em conjuntos menores e mais específicos, de acordo com atributos, até atingirem um tamanho simplificado o suficiente para serem rotulados, treinando-os dentro do modelo em prol da aplicação deste em dados novos. Um dos algoritmos pautados nesse método é o *Random Forest*, que consiste na execução de uma árvore de decisão várias vezes, onde cada uma delas utiliza um subconjunto dos atributos e um dos dados de forma aleatória para o treinamento.

3.7 Análise de Componentes Principais (PCA)

A análise de componentes principais ou PCA (*Principal Component Analysis*) é uma técnica pautada na análise multivariada de interrelações entre um número significativo de variáveis, explicando-as em componentes principais, ou seja, dimensões inerentes. O objetivo da utilização do método está na redução do espaço de dimensão original para um com a maior parcela possível de informação (BISHOP, 2006).

3.8 Ganho de informação e Coeficiente de Gini

Ganho de informação e Coeficiente de Gini são critérios utilizados em prol da análise de importância de atributos, a partir da capacidade de informação dos próprios atributos. O ganho de informação está relacionado a medida de entropia (GÉRON, 2019), que se utiliza da estratégia de redução da impureza medindo o nível de aleatoriedade dos ramos em uma árvore de decisão, já o coeficiente de Gini calcula a impureza relacionada a ocorrência ou não de um evento (CERIANI; VERME, 2012).

4 REVISÃO BIBLIOGRÁFICA

Nesta seção serão abordados trabalhos relacionados à análise e a construção do banco de dados a ser utilizado (FILHO *et al.*, 2018) (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018), assim como alguns trabalhos que abordam metodologias que aplicam técnicas de AM, como *K-Nearest Neighbors* (KNN) (ROCHA *et al.*, 2020), *Support Vector Machine* (SVM) (DI *et al.*, 2019) e Árvore de Decisão (SÁNCHEZ-MAROÑO *et al.*, 2017), em dados coletados a partir de formulários de pesquisas de cunho comportamental e/ou sócio-demográfico.

No trabalho (FILHO *et al.*, 2018), são apresentadas análises interseccionais de como se dá a interferência de marcadores de raça/classe no medo do crime e no autoritarismo a partir de dados da pesquisa "Medo da violência e o apoio ao autoritarismo no Brasil", realizada pelo Fórum Brasileiro de Segurança Pública e pelo Instituto Datafolha em 2017 (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018), apurados a partir de questionários aplicados a 2087 participantes. Além disso, o trabalho expõe uma interpretação teórica da estrutura da base de dados coletada. A exploração das informações ao nível conceitual possui suma importância tanto para a seleção da porção a ser utilizada na implementação dos modelos preditivos quanto para a proposição das abordagens e cenários a serem selecionados.

Além do que é abordado em (FILHO *et al.*, 2018), há um texto de debate descritivo sobre a pesquisa de (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018), elaborado pelo próprio órgão. No texto, é explanado tanto sobre os aspectos dentro da Psicologia social, como também sobre a construção e validação das escalas psicométricas utilizadas na coleta dos dados e nas análises quantitativas realizadas.

No trabalho (ROCHA *et al.*, 2020), são aplicadas duas técnicas de aprendizado de máquina a partir de um algoritmo não-supervisionado, no caso o *K-means*, e de um algoritmo supervisionado, KNN, visando a detecção de possíveis problemas psicológicos em indivíduos, baseada na predição do fator preponderante do teste de personalidade *Big Five*, que se dá por um conjunto de perguntas. Como resultado observado, o algoritmo KNN se sobressaiu com 70% de acurácia ao K-*means* com 60%.

Em (DI *et al.*, 2019), propõe-se a adoção do método de aprendizagem de máquina SVM com o intuito de detectar dependência à internet (*Internet Addiction Disorder*; IAD) de estudantes universitários chineses, utilizando dados de 2397 alunos obtidos por meio de questionários. Os itens que compõem os formulários são dispostos em valores escalados de acordo com abordagens direcionadas ao cálculo de índices como o *Chinese Big Five Personality*

Inventory (CBF-PI), que se trata de uma versão do utilizado em (ROCHA *et al.*, 2020). Para validação dos modelos propostos, é utilizada a técnica 10-*fold*. Dentre os resultados obtidos, o melhor valor de acurácia apresentada é de 96,32%.

Já em (SÁNCHEZ-MAROÑO *et al.*, 2017), são utilizados algoritmos de Árvore de Decisão para classificação de modelos comportamentais baseados em fatores demográficos e psicológicos individuais que influenciam no comportamento pró-ambiental. Para aquisição de dados, é disponibilizado aos participantes de uma pesquisa um questionário composto por blocos de questões, onde algumas seguem o modelo de escala, outras o modelo de valores booleano para assertivas duais e ainda outras com valores exatos.

5 METODOLOGIA

A metodologia proposta no trabalho consiste na condução de etapas do método KDD, a partir da base de dados utilizada, do delineamento de cenários de abordagem dos dados, das propostas de modelos de simulações e da análise de atributos mais relevantes.

5.1 Base de Dados

A base de dados utilizada apresenta informações relacionadas à medo, vitimização e chances de ocorrência de crimes, dados sociodemográficos, pautados dentro das diretrizes do Critério Brasil de Classificação Econômica de 2017 (ABEP, 2017), e grau de concordância em posicionamentos autoritários, de acordo com os relatos coletados em forma de formulário, estruturado em blocos (FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA, 2018). No estudo, considera-se *Bloco* como seção de sentenças referentes à uma abordagem específica do formulário e *Índice* como a taxa calculada a partir da média aritmética entre os resultados de cada sentença de um bloco.

Sobre os dados do formulário (Anexo A), temos:

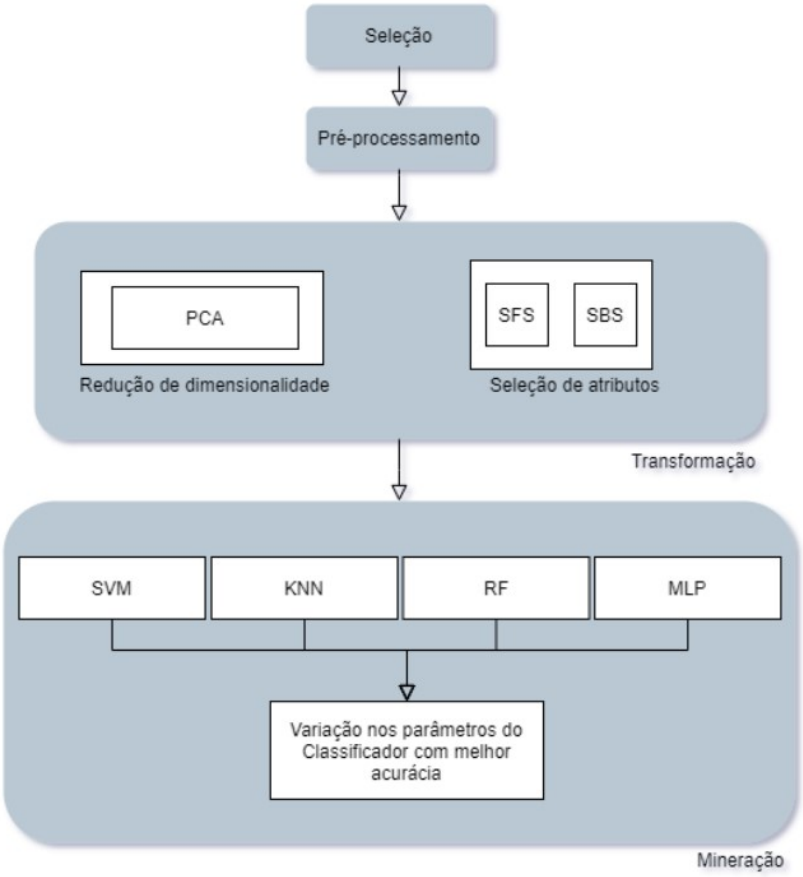
- Bloco 1: Medo do crime;
- Bloco 2: Escala de chances de ocorrência de vitimização por crime;
- Bloco 3: Escala de vitimização do crime;
- Dados Pessoais e sócio-econômicos (Bloco 5);
- Bloco 4: Autoritarismo.

Sobre as informações a cerca das dimensionalidades do base de dados, têm-se 2087 amostras no total, entre válidas e nulas, e 83 atributos, sendo distribuídos entre os blocos e abordados através de métodos de mensuração distintas, como apresentados no Anexo B. O processo de coleta de dados foi baseado em critérios representativos, estratificados e aleatórios, conforme atualização dos dados populacionais brasileiros a partir do censo de 2010 (FILHO *et al.*, 2018).

5.2 Aplicação do método KDD

Nesta seção, são explanadas as etapas predefinidas do KDD para as simulações, assim como também as estratégias e algoritmos utilizados em cada uma delas. Na Figura 1, temos ilustrada em formato de diagrama a estrutura de aplicação do método.

Figura 1 – Diagrama com etapas do KDD das simulações



Fonte: Elaborado pela autora

5.2.1 Seleção e Pré-processamento

Para análise dos dados, é aplicado o método KDD, onde na etapa que corresponde ao pré-processamento dos dados são descartadas as amostras nulas e inválidas aos itens, resultando em 1757 amostras. Nessa etapa, há um incremento no número de variáveis, pois os atributos "Sexo" e "Cor" passam pelo processo de binarização, totalizando em 88 variáveis. O resultado dessa etapa encontra-se ilustrado na Tabela 1.

5.2.2 Transformação

Na etapa de transformação, tais amostras passam pelos processos de normalização e padronização por *z-score*, não sendo aplicadas ao mesmo tempo e sim aplicadas à posteriori a cada um dos cenários propostos.

Sobre a estratégia de abordagem da utilização dos dados, para a aplicação dos modelos preditivos, 3 cenários de entradas e saídas do algoritmo são estipulados, considerando as abordagens de valores das sentenças, como também valores de índices calculados. Para o

Tabela 1 – Base de dados pós pré-processamento

Blocos	nº de itens	Método de mensuração
Dados Pessoais	3	Numérico para *Idade* e binário para *Sexo*
Bloco 1 - Medo do crime	16	Binário, 1 para "Sim"e 2 para "Não"
Bloco 2 - Escala de chances de ocorrência de vitimização por crime	16	Escala de Likert com valores de 0 (nenhuma chance de acontecer) à 10 (muita chance de acontecer)
Bloco 3 - Escala de vitimização do crime	15	Binário, 1 para "Sim"e 2 para "Não"
Bloco 4 - Assertivas relacionadas a propensão ao apoio a posições autoritárias	17	Escala de Likert com valores de 1 (Discordo totalmente) à 6 (Concordo totalmente)
Bloco 5 - Dados sócio-econômicos	21	Numérico, dependendo da sentença

Fonte: Elaborado pela autora.

procedimento, sendo *indB1* o valor do índice do Bloco 1 de cada amostra e *medB1* o valor da mediana dos valores de *indB1* de todas as amostras, são considerados:

- Primeiro Cenário:
  - Entrada: Valores dos itens de Dados Pessoais, Valores dos itens do Bloco 2, Valores dos itens do Bloco 3, Valores dos itens do Bloco 4, Valores dos itens do Bloco 5;
  - Classe: Para [indB1] < [medB1] terá valor 0, para [indB1] >= [medB1] terá valor 1
- Segundo Cenário:
  - Entrada: Valores dos itens de Dados Pessoais, Índice do Bloco 2, Índice do Bloco 3, Índice do Bloco 4, Valores dos itens do Bloco 5;
  - Classe: Para [indB1] < [medB1] terá valor 0, para [indB1] >= [medB1] terá valor 1
- Terceiro Cenário:
  - Entrada: Valores dos itens de Dados Pessoais, Valores dos itens do Bloco 5;
  - Classe: Para [indB1] < [medB1] terá valor 0, para [indB1] >= [medB1] terá valor 1

Nos passos de análises seguintes, relacionados a fase de transformação dentro do método KDD, são utilizadas as estratégias de seleção de atributos *Sequential Forward Select* (SFS) e *Sequential Backward Selection* (SBS) para realização dos testes de identificação de possíveis melhores performances e resultados. Como estratégias de redução de dimensionalidade, é utilizada a técnica de *Principal Component Analysis* (PCA).

5.2.3 Mineração

Sobre os modelos de classificação, no que se refere a etapa de mineração no método KDD, é optado pela realização de testes utilizando os classificadores SVM, KNN e *Random*

Forest (RF), implementados na linguagem *Python* através da biblioteca *Scikit-Learn* (GÉRON, 2019). Para validação dos modelos, é utilizado o método *K-fold*, com k = 10.

Dentre os parâmetros abordados no modelo SVM, os definidos para o *Kernel* são: Linear, Polinomial, *Radial Basis Function* e *Sigmoid*, com valores do parâmetro de regularização C em 0,01, 0,1, 1, 10 e 100, e do coeficiente do *kernel* γ em 0,002, 0,05, 1, 22, seguindo valores em sequência exponencial crescente por demostrarem bons resultados em experimentos (HSU *et al.*, 2003). Para a aplicação do algoritmo KNN, são utilizados para o número de vizinhos *k* os valores 5 (padrão da biblioteca), 1, 3, 7 e 10, valores próximos para evitar ruídos, no caso de *k* muito pequeno, e abrangências desnecessárias à análise, no caso de grandes valores de *k* (RUSSELL; NORVIG, 2010). Como valores de parâmetros do algoritmo RF, são consideradas as métricas de pureza dos dados Entropia e Coeficiente de Gini e as quantidades de estimadores 10, 100 e 1000, seguindo valores em sequência exponencial crescente de base 10, abordagem muito utilizada para presumir o melhor valor em simulações (GÉRON, 2019).

5.3 Importância dos Atributos

Para a análise dos atributos mais relevantes, serão utilizados os parâmetros Ganho de Informação e Coeficiente de Gini, obtidos a partir da execução do algoritmo de árvore de decisão RF.

6 RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados os resultados obtidos, de acordo com a metodologia utilizada, bem como é feita uma discussão dos resultados. Na seção 6.1 é mostrada análise dos atributos, partindo da observação dos resultados obtidos através dos cálculos de correlação dos atributos, com valores e representações gráficas. Nas seções 6.2, 6.3 e 6.4, estratégias dos modelos são comparadas através dos valores de acurácia média, calculadas a partir da média aritmética dos valores de acurácia obtidos de 10 execuções de cada algoritmo completo. Na seção 6.5 é feito o levantamento da importância dos atributos, a partir da perspectiva de comparativo entre os valores dos coeficientes de Gini e entropia.

6.1 Análise de variáveis

Como mencionado anteriormente, a base de dados possui 88 variáveis, sendo distribuídos em 3 cenários de estudo onde são aplicados os métodos de *Machine Learning*. Para avaliação da relação entre pares de variáveis, é utilizado a técnica do Coeficiente de Correlação de Person (r), ilustrado com os respectivos valores através do gráfico na Figura 2. Devido à quantidade significativa de variáveis no banco de dados, uma análise decomposta dentro dos cenários de estudo apresentados mostra-se mais efetiva.

Seguindo a sugestão de interpretação de direção e magnitude dos valores obtidos sugerida por (COHEN, 1992), podemos elencar 3 tipos de classificação:

- $|0{,}10| \leq r \leq |0{,}29|$ -> correlação fraca;
- $|0{,}30| \leq r \leq |0{,}49|$ -> correlação moderada;
- $|0{,}50| \leq r \leq |1{,}00|$ -> correlação forte.

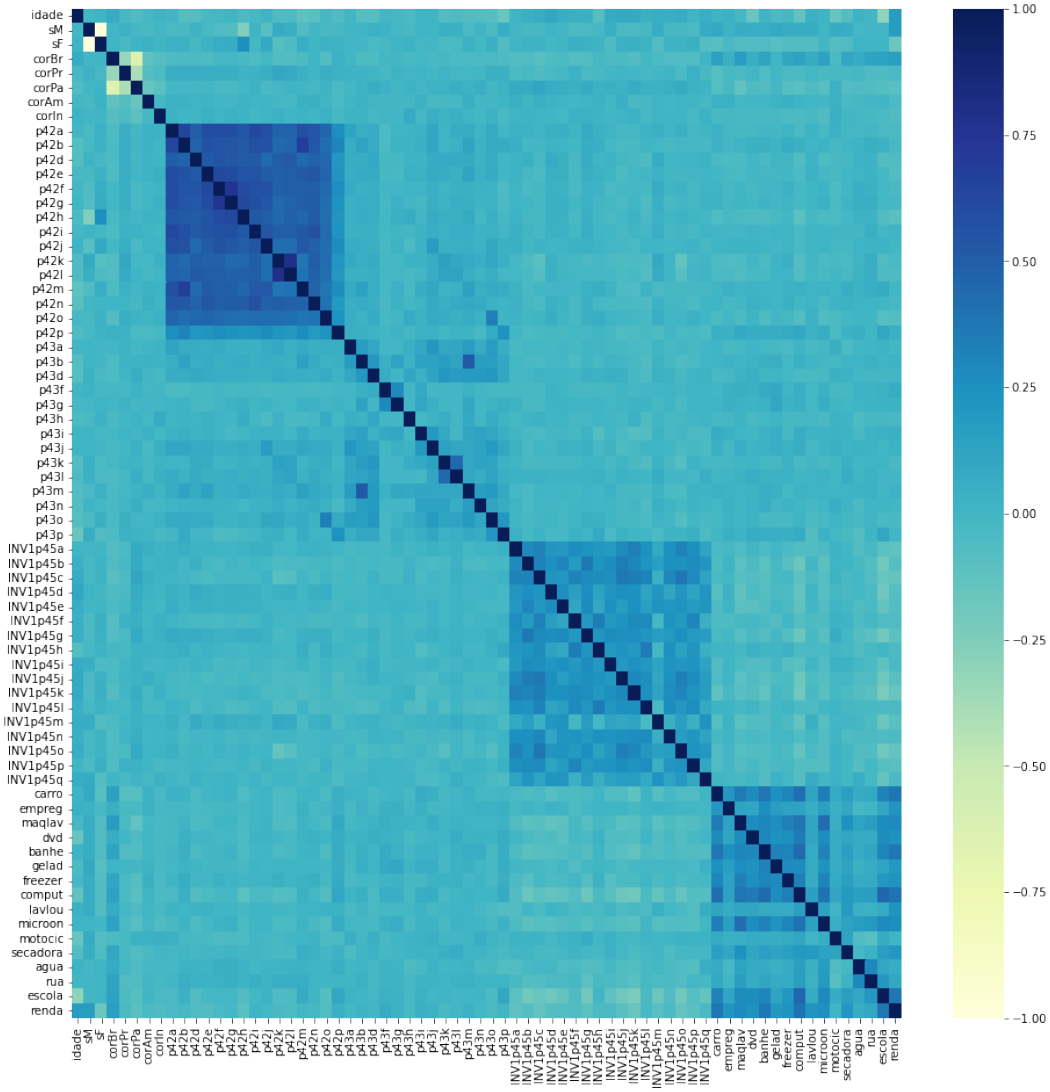
De acordo com os valores obtidos, analisando-os em quantidade percentual de pares correlatos, ilustrados na Tabela 2, e a matriz de correlação de cada cenário, nas figuras 2, 3 e 4, nota-se que a correlação à nível fraca e moderada entre as variáveis de entrada prevalece. Tal ponto mostra-se positivo, pois correlação forte entre os atributos pode indicar redundância.

Tabela 2 – Porcetagem de pares correlatos por níveis de classificação

Cenários	Fraca	Moderada	Forte
Cenário 1	80,94 %	15,51 %	3,55 %
Cenário 2	64,47 %	31,41 %	4,12 %
Cenário 3	58,68 %	36,46 %	4,86%

Fonte: elaborado pelo autor.

Figura 2 – Matriz de Correlação referente a todas as variáveis

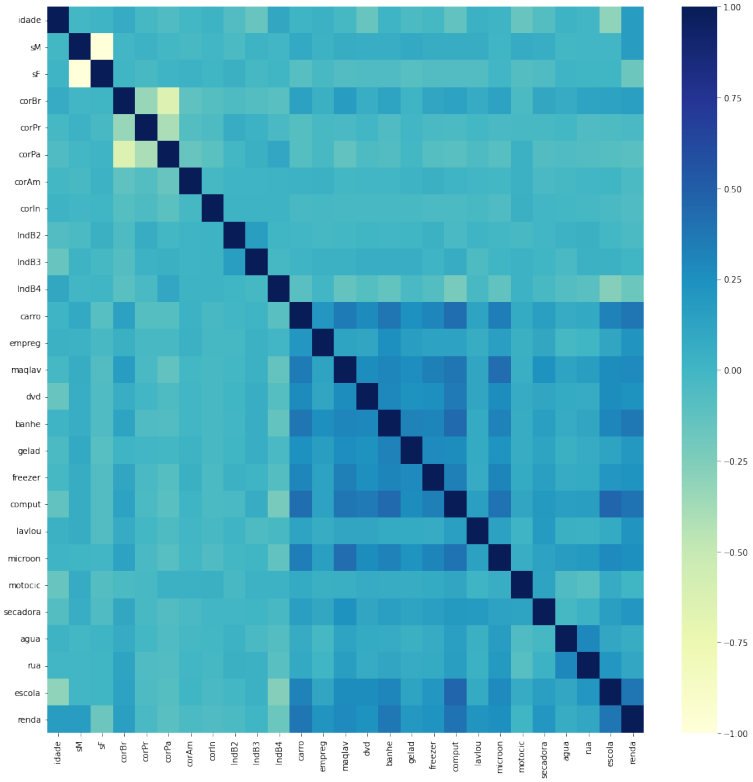


Fonte: Elaborado pela autora.

6.2 Análise dos Cenários

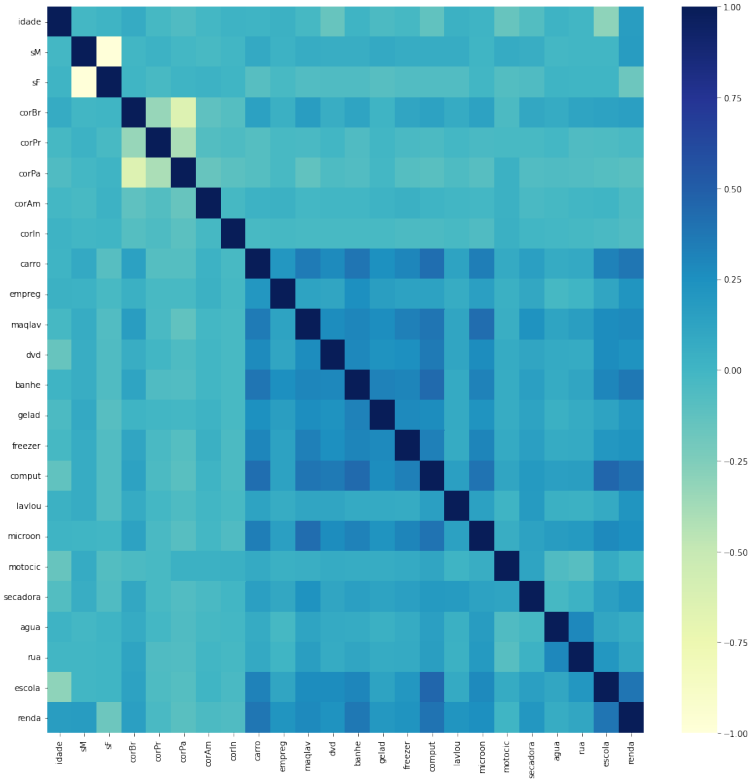
Na exploração dos possíveis cenários iniciais a serem utilizados nas simulações, são realizados testes utilizando os dados dentro do modelo SVM, com PCA, para análise do melhor cenário, baseado nos valores de acurácia. Dentro dessa perspectiva, são obtidos os valores 0,636 para o cenário 1, 0,666 para o cenário 2 e 0,613 para o cenário 3, ilustrados na Tabela 3, indicando que seguindo análise mais apurada do cenário 2 temos maior possibilidade de visualizar melhores resultados, já que nessa o valor de acurácia do cenário 2 se sobressaiu.

Figura 3 – Matriz de Correlação referente aos atributos do cenário 2



Fonte: Elaborado pela autora.

Figura 4 – Matriz de Correlação referente aos atributos do cenário 3



Fonte: Elaborado pela autora.

Tabela 3 – Acurácia média obtida a partir dos cenários iniciais

Cenários	**Acurácia**
Cenário 1	0,636
Cenário 2	0,666
Cenário 3	0,613

Fonte: Elaborado pela autora.

6.3 Transformação e mineração dos dados

Tendo como cenário escolhido o 2, ao serem aplicados os algoritmos de predição SVM, KNN e RF, observam-se os valores de acurária 0,666, 0,601, e 0,660, respectivamente. Tais resultados referem-se aos melhores valores de acurácia obtidos na execução de cada classificador com os valores de parâmetros propostos. O algoritmo SVM apresenta melhor resultado, reforçando sua eficiência em problemas com dimensionalidade significativa, como é o caso da base de dados utilizada neste trabalho. A partir de combinações de valores dos parâmetros *kernel* e C do classificador SVM, são coletadas medidas de acurácia, tendo como maior resultado obtido 0,667, referente à utilização do *kernel* Linear e C= 0,1, como mostrado na tabela 4.

Tabela 4 – Cenário 2 com PCA, aplicando no SVM

	Kernel			
C	Linear	Polinomial	RBF	Sigmoid
0,01	0,657	0,522	0,515	0,515
0,1	**0,667**	0,554	0,621	0,537
1	0,666	0,606	0,653	0,536
10	0,666	0,624	0,663	0,534
100	0,665	0,656	0,648	0,648

Fonte: Elaborado pela autora.

6.4 Melhor caso

Sob a perspectiva da utilização do método SVM no cenário 2 com os parâmetros citados anteriormente, são adicionados novos métodos dentro do roteiro já seguido do KDD, em prol da possibilidade de melhoria no valor de acurácia. Na etapa de pré-processamento, são executadas normalização e padronização dos dados em paralelo, combinados com as técnicas de redução de dimensionalidade, PCA, e seleção de atributos, SFS e SBS. Na Tabela 5, são mostrados os valores de acurácia reunidos a partir dessa abordagem de testes.

Baseado nessa análise, na utilização da abordagem de redução de dimensionalidade do algoritmo PCA, destaca-se, apesar da normalização e padronização dos dados, que o melhor resultado de acurácia média permanece na utilização dos dados não aplicados à esses métodos, em 0,667. Quanto à utilização da abordagem de seleção de atributos, para o algoritmo SFS, é obtida acurácia de 0,674, em dados normalizados. Para o algoritmo SBS identifica-se melhor resultado sem normalização ou padronização dos dados, com 0,672. A partir desses resultados, são identificadas diferenças mínimas nos valores, homologadas pela assertiva de que, por se tratar de uma base de dados numérica com escala de valores parecida, a normalização ou padronização dos dados não impacta significativamente no desempenho do classificador.

Tabela 5 – Simulações

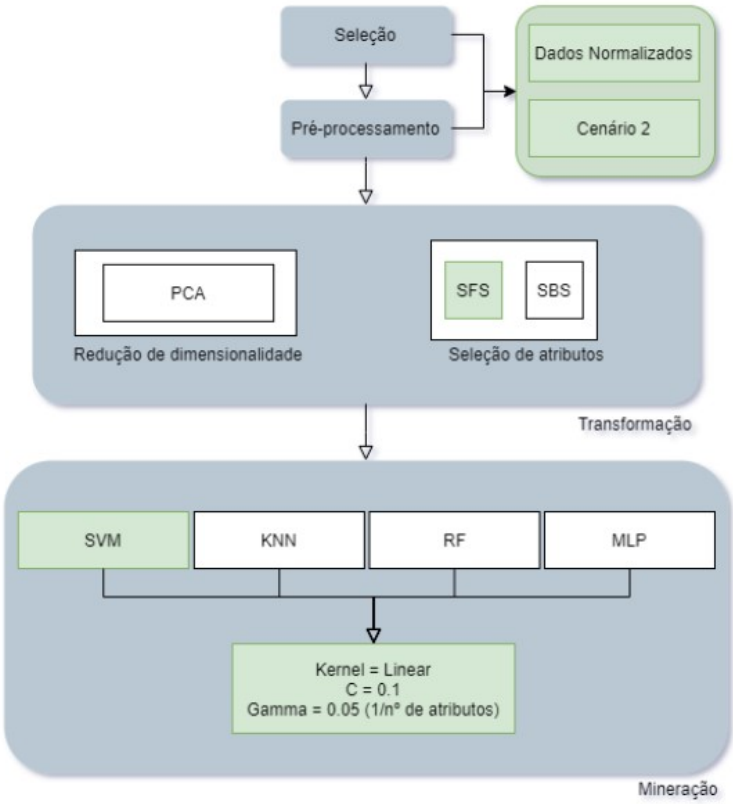
		Transformação + Mineração		
		PCA e SVM	SFS e SVM	SBS e SVM
Pré-processamento	Dados não normalizados e não padronizados	0,667	0,669	0,672
	Normalização	0,611	0,674	0,664
	Padronização z-score	0,665	0,672	0,670

Fonte: Elaborado pela autora.

Após análise das combinações de métodos, é identificado que o melhor roteiro de execução para a simulação seria a aplicação do SFS nos dados normalizados, em seguida executá-los em um SVM com os parâmetros escolhidos na Seção 6.3, como ilustrado no fluxograma da Figura 5. Nessa simulação, 7 atributos foram selecionados, são eles “idade”, “sM”, “sF”, “IndB2”, “IndB3”, “IndB4” e “escola”. Na seção 6.5 são apresentados melhores comentários sobre esses índices selecionados.

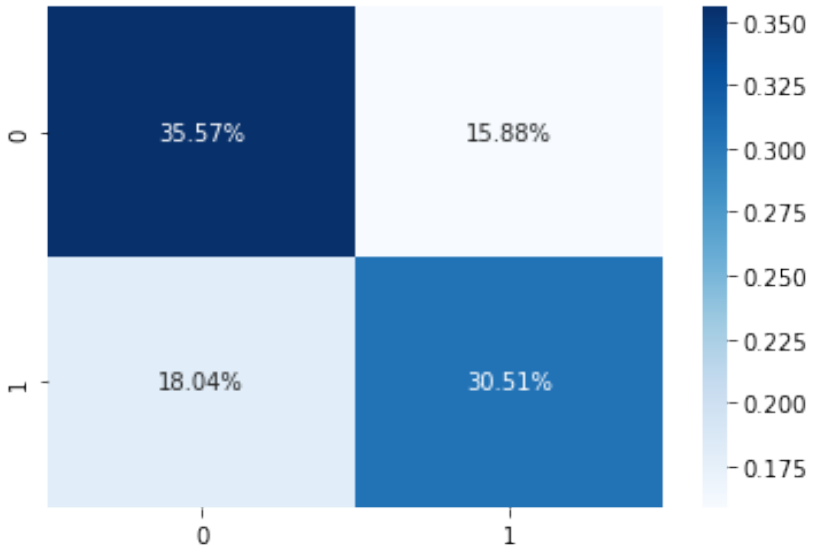
Como resultado da execução desse modelo, verifica-se acurácia média calculada de 0,674. Na Figura 6 é ilustrada em formato de matriz de confusão do modelo, onde se identificam 536 verdadeiros negativos, 279 falsos negativos, 625 verdadeiros positivos e 317 falsos positivos, em valores percentuais respectivos 30,51%, 15,88%, 35,57% e 18,04%. Analizando o resultado em termos de sensibilidade e especificidade, obtêm-se 0,69 e 0,63 respectivamente, considerado positivo pelo fato de o valor de sensibilidade ser maior que o valor de especificidade, pois indica maiores valores de verdadeiros positivos, ponto interessante no ponto de vista de modelos aplicados a dados que envolvem políticas públicas.

Figura 5 – Fluxograma do roteiro de etapas do modelo de melhor caso



Fonte: Elaborado pela autora.

Figura 6 – Matriz de confusão



Fonte: Elaborado pela autora.

6.5 Importância dos atributos

Após identificação do melhor modelo, algumas análises acima da importância dos atributos são feitas a seguir, considerando o coeficiente de Gini e a entropia relacionada ao ganho

de informação, com valores apresentados nas Tabelas 6 e 7, por ordem decrescente.

Tabela 6 – Coeficiente de Gini por atributos, em ordem decrescente de valores

Atributos	**Valores**
Índice B3	0,2585
idade	0,1319
Índice B2	0,1277
escola	0,0531
Índice B4	0,0413
sF	0,0403
dvd	0,0396
comput	0,0333
corPa	0,0325
renda	0,0318
...	...

Fonte: Elaborada pela autora.

Tabela 7 – Valores de entropia por atributos, em ordem decrescente de valores

Atributos	**Valores**
Índice B3	0,2460
idade	0,1724
Índice B2	0,1434
escola	0,0415
Índice B4	0,0357
comput	0,0354
dvd	0,0316
sM	0,0312
carro	0,0293
microon	0,0277
...	...

Fonte: Elaborada pela autora.

Realizando um comparativo, nota-se que os atributos “Índice B3”, “idade”, “Índice B2”, “escola” e “Índice B4” estão entre os que possuem os melhores valores no cenário 2, tanto de coeficiente de Gini como de entropia. Além disso, tais atributos estão entre os selecionados pelo algoritmo SFS no melhor modelo, evidenciando-os como mais relevantes para o modelo preditivo. O resultado condiz com o que é citado em (FÓRUM BRASILEIRO DE SEGURANÇA

PÚBLICA, 2018), onde expõe sobre tais fatores estarem associados à certas carências no âmbito civil e democrático, que impactam na construção de um cenário de medo e insegurança, já que dizem respeito a dados sociodemográficos, em idade e escolaridade, a tendência a posições autoritárias, a propensão a vitimização, e a própria chance de ser vítima de crime.

7 CONCLUSÃO E TRABALHOS FUTUROS

O presente trabalho teve como objetivo a identificação dos melhores modelos de análise de dados e técnicas de AM a serem aplicadas na previsão do grau relativo dos índices de predisposição ao medo do crime, relacionando esses níveis com informações sobre aspectos sociodemográficos e propensão à posições autoritárias do indivíduos. Para isso, foi necessária a realização de simulações de modelos preditivos de classificação com tratamento das informações dentro dos cenários propostos e a exploração do banco de dados, através da metodologia KDD.

A partir dos cenários propostos, foram encontrados melhores resultados no quadro onde os atributos referentes aos dados pessoais e socioeconômicos e aos índices sínteticos relacionados a chance de vitimização ao crime, ocorrência de vitimização e propensão a posições autoritárias são colocados como variáveis nas simulações. Tal observação contribui para a homologação da hipótese de que esses aspectos impactam na mensuração do medo do crime, pelos indivíduos ao qual foram coletadas as amostras com informações.

Sobre os modelos de previsão de predisposição ao medo do crime, a estratégia de utilização do classificador SVM com os dados normalizados, combinado com o algoritmo de análise de atributos SFS, se sobressaiu dentre as demais simulações, com o valor de acurácia 0,674 e com valores de sensibilidade e especificidade 0,69 e 0,63, respectivamente, verificando que a aplicação de métodos de AM é viável para esse tipo de análise.

Além disso, foi apurado que os atributos idade, escolaridade e índices voltados a predisposição à posições autoritárias e a chance e ocorrência de vitimização apresentam importância significativa para a previsão estimada, apontando-os como mais relevantes a serem consideradas.

Como trabalhos futuros, sugere-se propor aplicação das etapas de simulação em perspetivas de cenários diferentes da proposta, não só de entrada dos algoritmos, mas também de saída da previsão.

REFERÊNCIAS

ABEP Associação Brasileira de Empresas de Pesquisa. Critério brasil de classificação econômica. 2017.

ADORNO, T. W.; BRUNSWIK-FRENKEL, E.; LEVINSON, D. J.; SANFORD, R. N. The authoritarian personality. **Berkeley: The Norton Library.**, 1950.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. 1. ed. UK: Springer Science+Business Media, LLC, 2006. ISBN 0-387-31073-8.

BUSKIRK, T. D.; KIRCHNER, A.; ECK, A.; SIGNORINO, C. S. An introduction to machine learning methods for survey researchers. **Survey Practice**, American Association for Public Opinion Research, v. 11 (1), 2018.

CERIANI, L.; VERME, P. he origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. **The Journal of Economic Inequality**, v. 10, p. 421–443, 2012.

COHEN, J. A power primer. **Psychological Bulletin**, American Psychological Association, v. 121 (1), 1992.

CROCHIK, J. L. Personalidade autoritária e pesquisa empírica com a escala F: alguns estudos brasileiros. **Impulso**, v. 27(69), 2017.

DI, Z.; GONG, X.; SHI, J.; AHMED, H. O.; NANDI, A. K. Internet addiction disorder detection of chinese college students using several personality questionnaire data and support vector machine. **Addictive Behaviors Reports**, v. 10, p. 100200, 2019. ISSN 2352-8532.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, 1996. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>.

FILHO, T. L. L.; BARBOSA, V. N. M.; SEGUNDO, D. S. A.; JR., J. F. M.; JANNUZZI, P. M.; LIMA, R. S. Análises interseccionais a partir da raça e da classe: Medo do crime e autoritarismo no brasil. Psicologia: Ciência e Profissão, 2018.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. Medo da violência e o apoio ao autoritarismo no brasil: índice de propensão ao apoio a posições autoritárias. São Paulo, 2018.

G1. **50 milhões de brasileiros têm parente ou amigo assassinado, diz Datafolha**. 2017. Disponível em: <https://g1.globo.com/sao-paulo/noticia/50-milhoes-de-brasileiros-tem-parente-ou-amigo-assassinado-diz-datafolha.ghtml>. Acesso em: 24 fev. 2021.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2. ed. [S.l.]: O'Reilly Media, Inc, 2019.

HSU, C.-W.; CHANG, C.-C.; LIN, C.-J. **A Practical Guide to Support Vector Classification**. [S.l.], 2003.

KUBAT, M. **An Introduction to Machine Learning**. [S.l.]: Springer International Publishing AG, 2015. v. 2017.

MITCHELL, T. M. **Machine Learning**. 1. ed. USA: McGraw-Hill, Inc., 1997. ISBN 0070428077.

ROCHA, C. C.; PEREIRA, D. F.; MONTEIRO, A. F. A.; HENRIQUES, F. R. A preliminary study on the applied machine learning for detection of the predominant factor of big five personality test. **XI Computer on the Beach**, Anais do XI Computer on the Beach, v. 11 (1), 2020.

RUDIN, C.; WAGSTAFF, K. L. Machine learning for science and society. **Machine Learning**, Springer International Publishing AG, v. 95, p. 1–9, 2014.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. [S.l.]: Prentice Hall, 2010. v. 3.

SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A.; FONTENLA-ROMERO, O.; POLHILL, J. G.; CRAIG, T. Empirically-derived behavioral rules in agent-based models using decision trees learned from questionnaire data. In: ____. **Agent-Based Modeling of Sustainable Behaviors**. Cham: Springer International Publishing, 2017. p. 53–76. ISBN 978-3-319-46331-5.

ANEXO A – QUESTIONÁRIO APLICADO PELO FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA E INSTITUTO DATAFOLHA

Questionário Geral
Fórum Brasileiro de Segurança Pública
Data Folha

1. UF: ___
2. Idade: _____
3. Sexo: (1) Masculino (2) Feminino

Você diria que tem medo de...(**LEIA CADA ITEM) [ESTIMULADA E ÚNICA POR LINHA]**
P.1b **(P.1a = 1)** Se sim, muito medo ou pouco medo?

APLIQUE RODÍZIO	**Sim**	**Não**	**Muito Medo**	**Pouco Medo**	**Não sabe**
4. Ter sua residência invadida ou arrombada?	1	2	3	4	99
5. Ter objetos pessoais de valor tomados a força por outras pessoas em um roubo	1	2	3	4	99
6. Ter seu carro ou moto tomado de assalto ou furtados	1	2	3	4	99
7. Se envolver em brigas ou agressões físicas com outras pessoas	1	2	3	4	99
8. Morrer assassinado?	1	2	3	4	99
9. De ser sequestrado	1	2	3	4	99
10. De sofrer sequestro relâmpago	1	2	3	4	99
11. De ser vítima de agressão sexual	1	2	3	4	99
12. Ser vítima de uma fraude e perder quantia significativa de dinheiro	1	2	3	4	99
13. Receber uma ligação de bandidos exigindo	1	2	3	4	99
14. Ser vítima de violência por parte da Polícia Militar, aquela que executa o policiamento fardado e ostensivo nas ruas	1	2	3	4	99
15. Ser vítima de violência por parte da Polícia Civil, aquela que atua investigando crimes e registra ocorrência nas delegacias	1	2	3	4	99
16. Ter o celular furtado ou roubado	1	2	3	4	99
17. Ter os seus conteúdos pessoais divulgados na internet	1	2	3	4	99
18. Ter parentes envolvidos com drogas	1	2	3	4	99
19. Andar na vizinhança depois de anoitecer	1	2	3	4	99

Em uma escala de O a 10, onde 0 significa nenhuma chance de acontecer e 10 muita chance de acontecer, qual a chance de (**LEIA CADA ITEM) [ESTIMULADA E ÚNICA POR LINHA]**

APLIQUE RODÍZIO	**Anote**
20. Ter sua residência invadida ou arrombada no próximo mês?	
21. Ter objetos pessoais de valor tomados a força por outras pessoas em um roubo no próximo mês	
22. Ter seu carro ou moto tomado de assalto ou furtados no próximo mês	
23. Se envolver em brigas ou agressões físicas com outras pessoas no próximo mês	
24. Morrer assassinado no próximo mês	
25. De ser sequestrado no próximo mês	
26. De sofrer sequestro relâmpago no próximo mês	
27. De ser vítima de agressão sexual no próximo mês	
28. Ser vítima de uma fraude e perder quantia significativa de dinheiro no próximo mês	
29. Receber uma ligação de bandidos exigindo no próximo mês	

30. Ser vítima de violência por parte da Polícia Militar, aquela que executa o policiamento fardado e ostensivo nas ruas no próximo mês	
31. Ser vítima de violência por parte da Polícia Civil, aquela que atua investigando crimes e registra ocorrência nas delegacias no próximo mês	
32. Ter o celular furtado ou roubado no próximo mês	
33. Ter os seus conteúdos pessoais divulgados na internet no próximo mês	
34. Ter parentes envolvidos com drogas no próximo mês	
35. Andar na vizinhança depois de anoitecer no próximo mês	

No último mês : (**LEIA CADA ITEM) [ESTIMULADA E ÚNICA POR LINHA]**

APLIQUE RODÍZIO	**SIM**	**NAO**
36. Sua residência foi invadida ou arrombada	1	2
37. Ser roubado, assaltado ou furtado em casa, no transporte ou na escola/trabalho?	1	2
38. Seu carro ou moto tomado de assalto ou furtados	1	2
39. Se envolveu em brigas ou agressões físicas com outras pessoas	1	2
40. Foi sequestrado	1	2
41. Sofreu sequestro relâmpago	1	2
42. Foi vítima de agressão sexual	1	2
43. Foi vítima de uma fraude e perdeu quantia significativa de dinheiro	1	2
44. Recebeu uma ligação de bandidos exigindo dinheiro	1	2
45. Foi vítima de violência por parte da Polícia Militar, aquela que executa o policiamento fardado e ostensivo nas ruas	1	2
46. Foi vítima de violência por parte da Polícia Civil, aquela que atua investigando crimes e registra ocorrência nas delegacias	1	2
47. Teve o celular furtado ou roubado	1	2
48. Teve os seus conteúdos pessoais divulgados na internet	1	2
49. Teve parentes envolvidos com drogas	1	2
50. Andou na vizinhança depois de anoitecer	1	2

Vou ler algumas frases e gostaria que você me dissesse se concorda ou discorda de cada uma delas **(LEIA CADA ITEM)**. Você concorda ou discorda? **(SE CONCORDA OU DISCORDA)** Totalmente ou em parte? **(ESTIMULADA E ÚNICA)**

	APLIQUE O RODÍZIO	**CONCORDA TOTALMENTE**	**CONCORDA**	**CONCORDA EM PARTE**	**DISCORDA EM PARTE**	**DISCORDA**	**DISCORDA TOTALMENTE**	**NÃO SABE**
51	O que este país necessita, principalmente, antes de leis ou planos políticos, é de alguns líderes valentes, incansáveis e dedicados em quem o povo possa depositar a sua fé.	1	2	3	4	5	6	99
52	A maioria de nossos problemas sociais estaria resolvida se pudéssemos nos livrar das pessoas imorais, dos marginais e dos pervertidos.	1	2	3	4	5	6	99

53	A obediência e o respeito à autoridade são as principais virtudes que devemos ensinar as nossas crianças.	1	2	3	4	5	6	99
54	Os homens podem ser divididos em duas classes definidas: os fracos e os fortes.	1	2	3	4	5	6	99
55	Deve-se castigar sempre todo insulto à nossa honra.	1	2	3	4	5	6	99
56	A ciência tem o seu lugar, mas há muitas coisas importantes que a mente humana jamais poderá compreender.	1	2	3	4	5	6	99
57	Os crimes sexuais tais como o estupro ou ataques a crianças merecem mais que prisão; quem comete esses crimes deveria receber punição física publicamente ou receber um castigo pior.	1	2	3	4	5	6	99
58	Hoje em dia, as pessoas se intrometem cada vez mais em assuntos que deveriam ser somente pessoais e privados.	1	2	3	4	5	6	99
59	Um indivíduo de más maneiras, maus costumes e má educação dificilmente pode fazer amizade com pessoas decentes.	1	2	3	4	5	6	99
60	Se falássemos menos e trabalhássemos mais, todos estaríamos melhor.	1	2	3	4	5	6	99
61	Todos devemos ter fé absoluta em um poder sobrenatural, cujas decisões devemos acatar.	1	2	3	4	5	6	99
62	Não há nada pior do que uma pessoa que não sente profundo amor, gratidão e respeito por seus pais	1	2	3	4	5	6	99
63	Os homossexuais são quase criminosos e deveriam receber um castigo severo	1	2	3	4	5	6	99
64	Nenhuma pessoa decente, normal e em seu são juízo pensaria em ofender um amigo ou parente próximo	1	2	3	4	5	6	99
65	O policial é um guerreiro de Deus para impor a ordem e proteger as pessoas de bem.	1	2	3	4	5	6	99
66	Às vezes, os jovens têm ideias rebeldes que, com os anos, deverão superar para acalmar os seus pensamentos	1	2	3	4	5	6	99
67	Pobreza é consequência da falta de vontade de querer trabalhar	1	2	3	4	5	6	99

68. Qual sua cor?

(1) Branca

(2) Preta

(3) Parda

(4) Amarela

(5) Indígena

Agora, vou fazer mencionar alguns bens de consumo, e você me diz quantos:

69. Automóveis de passeio exclusivamente de uso particular? Quantos?

70. Empregados mensalistas, considerando apenas os que trabalham pelo menos 5 dias na semana? Quantos?

71. Máquinas de lavar roupas, excluindo tanquinho? Quantas?

72. Aparelho de DVD, incluindo qualquer outro dispositivo que leia DVD? Quantos?

73. Banheiros? Quantos?

74. Geladeira? Quantas?

75. Freezer independente ou aquele que faz parte da geladeira duplex? Quantos?

76. Microcomputador, considerando computadores de mesa, lap tops, notebooks e netbooks e excluindo tablets, palms ou smartphones? Quantos?

77. Máquina de lavar louças? Quantas?

78. Fornos de micro-ondas? Quantos?

79. Motocicletas, desconsiderando as que são utilizadas exclusivamente para fins profissionais? Quantas?

80. Máquina secadora de roupas? Quantas?

81. Até que ano da escola o chefe da família estudou?

(1) Analfabeto

(2) Primário ou Fundamental I Completo

(3) Ginasial ou Fundamental II Completo

(4) Colegial ou Ensino Médio Incompleto

(5) Colegial ou Ensino Médio Completo

(6) Superior Incompleto

(7) Superior Completo

(8) Pós-graduação

(97) Recuso

82. Renda Individual Mensal

(1)"Até 2 S.M."

(2) "Mais de 2 a 3 S.M"

(3)"Mais de 3 a 5 S.M"

(4) "Mais de 5 a 10 S.M"

(5) "Mais de 10 a 20 S.M"

(6) "Mais de 20 a 50 S.M"

(7) "Mais de 50 S.M"

(97) "Recusa"

(99) "Não sabe"

82. Voce mora em qual cidade? _____

ANEXO B – DESCRIÇÃO DOS DADOS

Tabela 8 – Medo do crime

Atributo	Descrição	Valores
p41aa	Ter sua residência invadida ou arrombada.	1 - Sim, 0 - Não
p41ab	Ter objetos pessoais de valor tomados a força por outras pessoas em um roubo ou assalto.	
p41ad	Se envolver em brigas ou agressões físicas com outras pessoas.	
p41ae	Morrer assassinado.	
p41af	De ser sequestrado.	
p41ag	De sofrer sequestro relâmpago.	
p41ah	De ser vítima de agressão sexual.	
p41ai	Ser vítima de uma fraude e perder quantia significativa de dinheiro.	
p41aj	Receber uma ligação de bandidos exigindo dinheiro.	
p41ak	Ser vítima de violência por parte da Polícia Militar, aquela que executa o policiamento fardado e ostensivo nas ruas.	
p41al	Ser vítima de violência por parte da Polícia Civil, aquela que atua investigando crimes e registra ocorrência nas delegacias.	
p41am	Ter o celular furtado ou roubado.	
p41an	Ter os seus conteúdos pessoais divulgados na internet.	
p41ao	Ter parentes envolvidos com drogas.	
p41ap	Andar na vizinhança depois de anoitecer.	

Fonte: Elaborado pela autora.

Tabela 9 – Escala de chances de ocorrência de vitimização por crime

Atributo	Descrição	Valores
p42a	Ter sua residência invadida ou arrombada no próximo mês.	0 (nenhuma chance de acontecer) - 10 (muita chance de acontecer)
p42b	Ter objetos pessoais de valor tomados a força por outras pessoas em um roubo ou assalto no próximo mês.	
p42d	Se envolver em brigas ou agressões físicas com outras pessoas no próximo mês.	
p42e	Morrer assassinado no próximo mês.	
p42f	De ser sequestrado no próximo mês.	
p42g	De sofrer sequestro relâmpago no próximo mês.	
p42h	De ser vítima de agressão sexual no próximo mês.	
p42i	Ser vítima de uma fraude e perder quantia significativa de dinheiro no próximo mês.	
p42j	Receber uma ligação de bandidos exigindo dinheiro no próximo mês.	
p42k	Ser vítima de violência por parte da Polícia Militar, aquela que executa o policiamento fardado e ostensivo nas ruas no próximo mês.	
p42l	Ser vítima de violência por parte da Polícia Civil, aquela que atua investigando crimes e registra ocorrência nas delegacias no próximo mês.	
p42m	Ter o celular furtado ou roubado no próximo mês.	
p42n	Ter os seus conteúdos pessoais divulgados na internet no próximo mês.	
p42o	Ter parentes envolvidos com drogas no próximo mês.	
p42p	Andar na vizinhança depois de anoitecer no próximo mês.	

Fonte: Elaborado pela autora

Tabela 10 – Escala de vitimização do crime

Atributo	Descrição	Valores
p43a	Sua residência foi invadida ou arrombada.	1 - Sim, 0 - Não
p43b	Você teve objetos pessoais de valor tomados a força por outras pessoas em um roubo ou assalto.	
p43d	Se envolveu em brigas ou agressões físicas com outras pessoas.	
p43f	Foi sequestrado.	
p43g	Sofreu sequestro relâmpago.	
p43h	Foi vítima de agressão sexual.	
p43i	Foi vítima de uma fraude e perdeu quantia significativa de dinheiro.	
p43j	Recebeu uma ligação de bandidos exigindo dinheiro.	
p43k	Foi vítima de violência por parte da Polícia Militar, aquela que executa o policiamento fardado e ostensivo nas ruas.	
p43l	Foi vítima de violência por parte da Polícia Civil, aquela que atua investigando crimes e registra ocorrência nas delegacias.	
p43m	Teve o celular furtado ou roubado.	
p43n	Teve os seus conteúdos pessoais divulgados na internet.	
p43o	Teve parentes envolvidos com drogas.	
p43p	Andou na vizinhança depois de anoitecer.	

Fonte: Elaborado pela autora

Tabela 11 – Assertivas relacionadas a propensão ao apoio a posições autoritárias

Atributo	Descrição	Valores
INV1p45a	O que este país necessita, principalmente, antes de leis ou planos políticos, é de alguns líderes valentes, incansáveis e dedicados em quem o povo possa depositar a sua fé.	1 (Discordo totalmente) - 6 (Concordo totalmente)
INV1p45b	A maioria de nossos problemas sociais estaria resolvida se pudéssemos nos livrar das pessoas imorais, dos marginais e dos pervertidos.	
INV1p45c	A obediência e o respeito à autoridade são as principais virtudes que devemos ensinar as nossas crianças.	
INV1p45d	Os homens podem ser divididos em duas classes definidas: os fracos e os fortes.	
INV1p45e	Deve-se castigar sempre todo insulto à nossa honra.	
INV1p45f	A ciência tem o seu lugar, mas há muitas coisas importantes que a mente humana jamais poderá compreender.	
INV1p45g	Os crimes sexuais tais como o estupro ou ataques a crianças merecem mais que prisão; quem comete esses crimes deveria receber punição física publicamente ou receber um castigo pior.	
INV1p45h	Hoje em dia, as pessoas se intrometem cada vez mais em assuntos que deveriam ser somente pessoais e privados.	
INV1p45i	Um indivíduo de más maneiras, maus costumes e má educação dificilmente pode fazer amizade com pessoas decentes.	
INV1p45j	Se falássemos menos e trabalhássemos mais, todos estaríamos melhor.	
INV1p45k	Todos devemos ter fé absoluta em um poder sobrenatural, cujas decisões devemos acatar.	
INV1p45l	Não há nada pior do que uma pessoa que não sente profundo amor, gratidão e respeito por seus pais.	
INV1p45m	Os homossexuais são quase criminosos e deveriam receber um castigo severo.	
INV1p45n	Nenhuma pessoa decente, normal e em seu são juízo pensaria em ofender um amigo ou parente próximo.	
INV1p45o	O policial é um guerreiro de Deus para impor a ordem e proteger as pessoas de bem.	
INV1p45p	Às vezes, os jovens têm ideias rebeldes que, com os anos, deverão superar para acalmar os seus pensamentos.	
INV1p45q	Pobreza é consequência da falta de vontade de querer trabalhar.	

Fonte: Elaborado pela autora

Tabela 12 – Dados Pessoais e sócio-econômicos

Atributo	Descrição	Valores
idade	Idade.	Valores inteiros
sf	Sexo Feminino.	binário
sm	Sexo Masculino.	binário
corBr	Cor Branca.	binário
corPr	Cor Preta.	binário
corPa	Cor Parda.	binário
corAm	Cor Amarela.	binário
corIn	Cor Indígena.	binário
carro	Automóveis de passeio exclusivamente de uso particular? Quantos?	0-0, 1-1, 2-2, 3-3, 4-4 ou mais
empreg	Empregados mensalistas, considerando apenas os que trabalham pelo menos 5 dias na semana? Quantos?	
maqlav	Máquinas de lavar roupas, excluindo tanquinho? Quantas?	
dvd	Aparelho de DVD, incluindo qualquer outro dispositivo que leia DVD? Quantos?	
banhe	Banheiros? Quantos?	
gelad	Geladeira? Quantas?	
freezer	Freezer independente ou aquele que faz parte da geladeira “duplex”? Quantos?	
comput	Microcomputador, considerando computadores de mesa, lap tops, notebooks e netbooks e excluindo tablets, palms ou smartphones? Quantos?	
lavlou	Máquina de lavar louças? Quantas?	
microon	Fornos de micro-ondas? Quantos?	
motocic	Motocicletas, desconsiderando as que são utilizadas exclusivamente para fins profissionais? Quantas?	
secadora	Máquina secadora de roupas? Quantas?	
agua	A água utilizada no seu domicílio é proveniente de rede geral de distribuição, poço, nascente ou de outro meio?	0-Não possui, 1-Outro meio, 2-Poço ou nascente, 3-Rede geral de distribuição
rua	Considerando o trecho da rua do seu domicílio, você diria que a rua é asfaltada, pavimentada, de terra ou cascalho?	0-Outro, 1-Terra ou cascalho, 2-Asfaltada, pavimentada
escola	Até que ano da escola o chefe da família estudou?	1-Analfabeto/ Primário / Fundamental I incompleto, 2-Primário ou Fundamental I completo/ Ginasial ou Fundamental II incompleto, 3-Ginasial ou Fundamental II completo, 4-Colegial ou Ensino Médio incompleto, 5-Colegial ou Ensino Médio completo, 6-Superior incompleto, 7-Superior completo, 8-Pós-graduação
renda	Renda Individual Mensal	1-Até 2 S.M, 2-Mais de 2 a 3 S.M, 3-Mais de 3 a 5 S.M, 4-Mais de 5 a 10 S.M, 5-Mais de 10 a 20 S.M, 6-Mais de 20 a 50 S.M, 7-Mais de 50 S.M

Fonte: Elaborado pela autora