
Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Suscetibilidade Quanto ao Uso da Internet pelo Público Infantojuvenil de Monte Carmelo

Franciele Cristina Espanhol Ferreira Alves



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Monte Carmelo - MG
2020

Franciele Cristina Espanhol Ferreira Alves

**Técnicas de Aprendizado de Máquina Aplicadas
na Previsão de Suscetibilidade Quanto ao Uso
da Internet pelo Público Infantojuvenil de
Monte Carmelo**

Trabalho de Conclusão de Curso apresentado à
Faculdade de Computação da Universidade Federal
de Uberlândia, Minas Gerais, como requisito exigido
parcial à obtenção do grau de Bacharel em Sistemas
de Informação.

Área de concentração: Sistemas de Informação

Orientador: Dra. Fernanda Maria da Cunha Santos

Monte Carmelo - MG

2020

Dedico este trabalho à minha família que com muito carinho não mediram esforços para que eu chegasse até esta etapa da minha vida.

Agradecimentos

Agradeço em primeiro lugar a Deus, por ser essencial em minha vida, autor do meu destino e meu guia. Aos meus familiares, pelo amor, carinho, compreensão e que sempre me deram forças para continuar a chegar no meu objetivo. A professora Fernanda Maria da Cunha Santos pela paciência na orientação e incentivo que tornaram possível a conclusão desta monografia. Aos membros da banca de defesa por aceitaram contribuir com a pesquisa. Enfim, a todos aqueles que de uma maneira ou de outra colaboraram para que este sonho pudesse ser realizado.

*“Grandes coisas fez o Senhor por nós, pelas quais estamos alegres.”
(Salmos 126:3)*

Resumo

A Internet oferece a conexão de bilhões de pessoas em todo o mundo por meio de várias formas de interações e organizações sociais. As redes sociais têm se tornado uma das principais formas de comunicação, que proporciona boas experiências e, contraditoriamente, preocupações para pais e professores de crianças e adolescentes. As redes sociais estão sendo utilizadas pelo público infantojuvenil demasiadamente, possibilitando gerar situações de riscos e vulnerabilidades à eles. Diante destes fatores, o objetivo deste trabalho foi empregar duas técnicas de Aprendizagem de Máquina para prever se o público infantojuvenil da cidade de Monte Carmelo está suscetível à alguma ameaça ou situação constrangedora na Internet. Esta previsão será feita ao analisar uma base de dados constituída de características pessoais que qualifiquem as formas de uso deste público em relação ao seu acesso às redes sociais da Internet. As técnicas de Aprendizado de Máquina escolhidas foram os algoritmos de Árvore de Decisão e Naive Bayes, ambos implementados na linguagem Python e pelas bibliotecas do WEKA. Nessas ferramentas computacionais, as técnicas de Aprendizado de Máquina apresentaram resultados satisfatórios, com acurácia acima de 90%.

Palavras-chave: Redes Sociais, Aprendizado de Máquina, Segurança Virtual, Infantojuvenil.

Lista de ilustrações

Figura 1 – Árvore de decisão para o problema de classificação de mamíferos. . . .	17
Figura 2 – Fluxograma da metodologia para classificar se há ou não susceptibilidade de ameaças na forma de uso as redes sociais.	22
Figura 3 – Relação entre a idade dos entrevistados e as redes sociais que mais utilizam.	23
Figura 4 – Quantidade de alunos que possuem Internet em casa.	24
Figura 5 – Meio eletrônico mais utilizado.	24
Figura 6 – Relação entre idade e se passaram por alguma situação constrangedora nas redes sociais.	25
Figura 7 – Tela do WEKA.	26

Lista de tabelas

Tabela 1 – Matriz de confusão dos resultados gerados pela árvore de decisão no WEKA.	29
Tabela 2 – Matriz de confusão dos resultados gerados pelo algoritmo Naive Bayes no WEKA.	29
Tabela 3 – Grau de acurácia dos classificadores.	30

Lista de siglas

AM Aprendizado de Máquinas

AD Árvore de Decisão

HMM Hidden Markov Models

RB Rede Bayesiana

Sumário

1	INTRODUÇÃO	11
1.1	Motivação	12
1.2	Objetivos da Pesquisa	13
1.3	Contribuições	14
1.4	Organização da Monografia	14
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	Aprendizagem de Máquina	15
2.1.1	Árvore de Decisão	16
2.1.2	Aprendizado Bayesiano	18
2.2	Trabalhos Correlatos	18
3	EXPERIMENTOS E ANÁLISE DOS RESULTADOS	21
3.1	Método para a Avaliação	21
3.1.1	Criação da Base de dados	21
3.1.2	WEKA	25
3.1.3	As Técnicas de Aprendizagem de Máquinas	26
3.2	Experimentos	28
3.3	Avaliação dos Resultados	30
4	CONSIDERAÇÕES FINAIS	31
4.1	Principais Contribuições	31
4.2	Trabalhos Futuros	32
	REFERÊNCIAS	33

APÊNDICES

35

APÊNDICE A	–	36
A.1	Questionário aplicado aos alunos das escolas de Monte Carmelo- MG	36
A.2	Código na linguagem Python da árvore de decisão	38
A.3	Código na linguagem Python do Naive Bayes	39

CAPÍTULO **1**

Introdução

A sociedade vive em constante modificação. A Internet e as tecnologias estão cada vez mais presentes na vida das pessoas, conseqüentemente, crianças começam cada vez mais cedo a lidar com elas (RODRIGUES, 2015).

O uso imprudente das inovações tecnológicas pode colocar em risco a segurança e a saúde de seus usuários. Para Machado e Mochi (2013) pode ocorrer dano emocional, diminuição de autoestima ou perturbação no desenvolvimento. Nesse contexto, os pais e responsáveis surgem como orientadores e mediadores da relação que os seus filhos devem estabelecer com as diferentes tecnologias. Como exemplo, a utilização da Internet, principalmente, o uso das redes sociais pelo público infantojuvenil é um tema que vem gerando grande preocupação entre pais e educadores.

Crianças e adolescentes passam a maior parte do tempo em redes sociais, interagindo com pessoas conhecidas ou que nunca viram pessoalmente, sem terem noção das vulnerabilidades a que estão expostos. Ou seja, uma vez que os dados pessoais estão disponíveis na rede de computadores, é possível visualizar o perfil e saber onde pessoas vivem e o que fazem, podendo originar situações de perseguições, chantagens e roubo de identidade. Pessoas mal-intencionadas estabelecem contato com crianças por meio de conversas em salas de bate-papo, jogos compartilhados, mensagens instantâneas e e-mail.

Em um mundo globalmente conectado à Internet e com uma sociedade rica em informação, é preciso haver o discernimento e equilíbrio das informações expostas, identificando comportamentos em ambientes digitais e promovendo tomadas de decisões críticas (BERSON; BERSON, 2005). A busca por softwares que realizam filtros em redes sociais não é suficientemente eficaz para gerir o comportamento dos jovens perante o mundo digital, sendo preciso a conscientização e atenção dos usuários da falta de segurança na Internet (FLEMING et al., 2006).

Para tentar suprir a falta de segurança na Internet e visando o monitoramento do comportamento de crianças e adolescentes quando estão no mundo virtual, alguns autores estão utilizando os algoritmos de Aprendizagem de Máquina para tentar detectar ameaças e vulnerabilidades (JÚNIOR et al., 2015; SANTIN P. L. ; FREITAS, 2012; SANTOS;

GUEDES, 2019).

Nas últimas décadas, o Aprendizado de Máquinas (AM) (do inglês *Machine Learning*) tornou-se um dos pilares da tecnologia da informação, destacando algoritmos e técnicas computacionais que são capazes de aprender a partir de dados existentes. Isto é, com a crescente quantidade de dados se tornando disponível, há boas razões para acreditar que a análise inteligente de dados se tornará ainda mais difundido como ingrediente necessário para o progresso tecnológico (SMOLA; VISHWANATHAN, 2008). O AM tem uma gama extremamente ampla de aplicações. Por exemplo, tem grande contribuição em processamento de linguagem natural, motores de busca, diagnósticos médicos, bioinformática, reconhecimento de fala, reconhecimento de escrita, visão computacional e locomoção de robôs e sistemas de previsão (AMORIM; BARONE; MANSUR, 2008).

Objetivou-se para este trabalho, o uso de técnicas de Aprendizagem de Máquina para prever se o público infantojuvenil está suscetível à alguma ameaça ou situação constrangedora na Internet. Esta previsão será feita ao analisar uma base de dados constituída de características pessoais e de dados que qualifiquem as formas de uso deste público em relação ao seu acesso à redes sociais da Internet.

1.1 Motivação

A segurança nas redes sociais não é um assunto trivial, principalmente, quando se refere aos usuários com faixa etária infantojuvenil. A preocupação de pais e responsáveis estão em que seus filhos não fiquem expostos à situações de risco. Vários estudos e reportagens foram divulgados nos meios de comunicação informando os malefícios que a exposição das redes sociais podem trazer para os usuários das mesmas.

O Jornal EM FOCO¹ descreve os riscos que existem para crianças que acessam a rede como pedofilia, pornografia e crimes virtuais, e estão se tornando cada vez mais comuns. A vulnerabilidade do público infantojuvenil é cada vez maior, pois acreditam que estão seguros por estarem dentro de suas residências, e assim compartilham informações pessoais e mantém conversas com pessoas estranhas por meio das redes sociais.

Em uma reportagem exibida pelo programa Profissão Repórter da emissora Rede Globo² destaca como a Internet oferece riscos à crianças e adolescentes. Bullying nas redes sociais, vícios em jogos online, vídeos de desafios que podem ferir ou matar são apontados como riscos para esse público. A reportagem cita um garoto de 13 anos que participou de um jogo perigoso estimulado pela Web e acabou com quase 40% do corpo queimado quando tentava fazer o truque das mãos de fogo. Segundo a reportagem a

¹ <https://emfoco.anchieta.br/2018/04/05/criancas-e-adolescentes-na-internet-o-perigo-pode-estar-em-casa/>

² <https://g1.globo.com/profissao-reporter/noticia/2018/11/29/internet-pode-oferecer-riscos-para-criancas-e-adolescentes.ghtml>

garrafa de álcool explodiu na mão dele. No momento do acidente, ele estava sozinho em casa. O adolescente fez 13 cirurgias, três enxertos e ficou 65 dias internado.

Em texto exibido pela Agência Brasil³ psicólogos apontam que o uso excessivo da Internet em redes sociais e a sensação de segurança gerada pela comunicação via celular destacam o aumento de casos de sexting, devido à falta de instrução sexual e a erotização precoce.

Segundo a Revista ÉPOCA⁴ o assédio a menores de idade em redes sociais e em aplicativos ocorre em tempo integral. *"Uma pesquisa obtida com exclusividade por ÉPOCA mostra que ainda são poucos os pais e responsáveis por crianças no Brasil que impõem regras de uso na Internet para seus filhos"*.

Consequentemente, questiona se poderíamos prever ou classificar um padrão a partir do modo de uso dos usuários nas redes sociais, principalmente, do público infantojuvenil, utilizando um dos algoritmos propostos em AM. A partir desta previsão, poderíamos alertar pais e responsáveis por um perigo eminente à seu filho, que fica exposto a um mundo até então desconhecido e formado por distintos tipos de personalidades (PEREIRA, 2009).

Aliás, o aumento e o surgimento de novos problemas que ocorrem na Internet com o público infantojuvenil pode estar relacionado com a escassez de ferramentas computacionais propostas para a segurança virtual de crianças e adolescentes.

1.2 Objetivos da Pesquisa

Devido ao uso intenso, tanto os sites quanto os aplicativos das redes sociais reúnem materiais suficientes para subsidiar previsões e classificar padrões que caracterizam os seus usuários em larga escala (FREITAS et al., 2008). As técnicas de AM podem utilizar esta vasta quantidade de dados para aprender e gerar conclusões diante os questionamentos e tomadas de decisões.

Diante deste cenário, objetivou analisar e comparar o uso de técnicas de AM numa base de dados para prever se o público infantojuvenil está suscetível à alguma situação constrangedora na Internet. As informações contidas na base de dados foi construída por um estudo aplicado a um grupo de usuarios dos 9 aos 14 anos que utilizam as redes sociais, e que descrevem o comportamento que crianças e adolescentes da cidade de Monte Carmelo usam as redes sociais. Os métodos de AM utilizados neste trabalho foram Árvore de Decisão (AD) e a Rede Bayesiana (RB).

³ <http://agenciabrasil.ebc.com.br/educacao/noticia/2016-07/adolescentes-os-riscos-do-uso-excessivo-da-internet>

⁴ <https://epoca.globo.com/vida/experiencias-digitais/noticia/2015/10/rede-social-nao-e-lugar-para-crianca.html>

1.3 Contribuições

A realização deste trabalho visa destacar o comportamento do público infantojuvenil da cidade de Monte Carmelo em relação ao modo de uso nas redes sociais. Isto será realizado por métodos de AM capazes de analisar e identificar um padrão.

1.4 Organização da Monografia

O trabalho foi estruturado em 4 Capítulos. O Capítulo 2 apresenta o estado da arte, no qual o tema é fortalecido pela análise e detalhamento do referencial teórico; o Capítulo 3 descreve os algoritmos de AM implementados pelo WEKA e pela linguagem de programação Python, além da descrição da base de dados. O Capítulo 4 exibe os resultados atingidos pelos algoritmos AD e RB. Finalizando, o último Capítulo contém as considerações finais do trabalho.

Fundamentação Teórica

A Inteligência Computacional é uma área da Ciência da Computação que procura demonstrar pelos princípios da natureza, a criação de sistemas inteligentes que reproduzem funções do comportamento humano como o aprendizado, percepção, raciocínio, evolução e adaptação. O objetivo destes sistemas computacionais é atuar no auxílio de processos lotados em diferentes áreas de atuação, como a análise de dados e classificação de padrões na medicina, engenharia, ciências e negócios.

Nas próximas subseções foram apresentadas, resumidamente, teorias da Inteligência Computacional que foram estudadas e implementadas neste trabalho. Além disso, foi descrito as principais referências bibliográficas que trouxeram ideias e subsídios para o desenvolvimento desse trabalho.

2.1 Aprendizagem de Máquina

Para Mitchell (1997), AM é uma área de pesquisa da Inteligência Computacional que tem a finalidade de estudar métodos capazes de extrair conceitos com base em um modelo de informações. Um sistema de aprendizado é um programa de computador que possui capacidade de tomar decisões com base em experiências acumuladas por meio de soluções bem-sucedidas de problemas anteriores. Em outras palavras, o sistema de aprendizado tem a capacidade de tomar decisão por já ter tido experiência acumulada em casos de sucesso (MONARD; BARANAUSKAS, 2003c).

Existem duas formas de adquirir conhecimento pelos algoritmos propostos pelas técnicas de AM: supervisionado e não-supervisionado. Lorena e Carvalho (2003) relata que no aprendizado supervisionado tem-se a representação de um professor externo, onde o algoritmo de aprendizagem de máquina é treinado a partir de conjuntos de exemplos rotulados. Para Batista (2003), o aprendizado supervisionado consiste em fornecer ao sistema de aprendizado um conjunto de exemplos $E = \{E_1, E_2, \dots, E_n\}$, sendo que cada exemplo E_i , que está contido em E , possui um rótulo associado. Esse rótulo define a classe a qual o exemplo pertence. Assim, um exemplo E_i é uma tupla $E_i = (X_i, Y_i)$, onde X_i é um vetor

de valores que representam as características ou atributos do exemplo E_i . Y_i é o valor da classe desse exemplo, que pode ser um dado qualitativo que assume um conjunto de valores discretos.

Uma característica de todos os paradigmas de algoritmos supervisionados é que o conceito a ser aprendido deve estar relacionado com exemplos já observados, e cada exemplo deve estar rotulado com a classe a qual pertence. Já no algoritmo não-supervisionado, não há instâncias rotuladas da função a ser aprendida (LORENA; CARVALHO, 2003). Para Batista (2003), é fornecido ao sistema de aprendizado um conjunto de exemplos E , onde cada exemplo se dá somente de vetores X não incluindo especificação sobre a classe Y . Assim, o modelo é construído por regularidades, formando clusters dos exemplos com atribuições similares. Segundo Monard e Baranauskas (2003a), no aprendizado não-supervisionado, o algoritmo faz a análise de um modelo e tenta agrupá-los de alguma forma.

Existem algumas técnicas de AM que são implementados em sistemas computacionais que atuam na classificação e previsão de dados tanto em trabalhos científicos quanto em atividades rotineiras em diferentes áreas de atuação. Essas técnicas são: redes neurais artificiais, algoritmos genéticos, árvores de decisão, *Support Vector Machine* (SVMs), sistemas fuzzy, classificadores Bayesianos e outras (LORENA; CARVALHO, 2003).

Um sistema computacional que possa atuar em situações de incertezas deve ser capaz de atribuir níveis de confiabilidade para todas as sentenças em sua base de conhecimento, e ainda, estabelecer relações entre as sentenças. Lidar com falta de informação significa lidar com incertezas, e tais situações podem ser representadas por grafos que retratam as relações causais entre eventos. As redes bayesianas oferecem uma abordagem para o raciocínio probabilístico que engloba teoria de grafos, para o estabelecimento das relações entre sentenças e ainda, teoria de probabilidades, para a atribuição de níveis de confiabilidade (MARQUES R. L.; DUTRA, 2002).

De forma semelhante, uma árvore de decisão pode ser construídas a partir de um determinado conjunto de atributos, além de existir, exponencialmente, várias árvores que representam um mesmo conjunto (TAN et al., 2009).

As definições das técnicas de árvore de decisão e de aprendizado Bayesiano serão descritas sucintamente nas próximas subseções.

2.1.1 Árvore de Decisão

Árvore de decisão é uma estrutura de dados recursiva, onde um nó folha corresponde a uma classe (um rótulo) ou um nó de decisão (nó interno) que contém um teste sobre algum atributo. Em cada nó interno, possui exatamente uma aresta chegando e duas ou mais saindo para uma subárvore. Cada subárvore tem a mesma estrutura que a árvore (MONARD; BARANAUSKAS, 2003b).

A árvore é representada como um conjunto de regras. Cada regra tem início na raiz da árvore e caminha até uma de suas folhas. Aplica-se a condição de teste do nó raiz ao registro e segue a ramificação apropriada baseada no resultado do teste. Isto direcionará a um nó interno, que executará um novo teste resultando numa aresta direcionada a um nó folha ou a outro nó interno. Se for encaminhado a um nó folha, o rótulo da classe associado ao nó folha é atribuída ao registro. A Figura 2.1.1 mostra o caminho na árvore de decisão que preve o rótulo da classe de um flamingo.

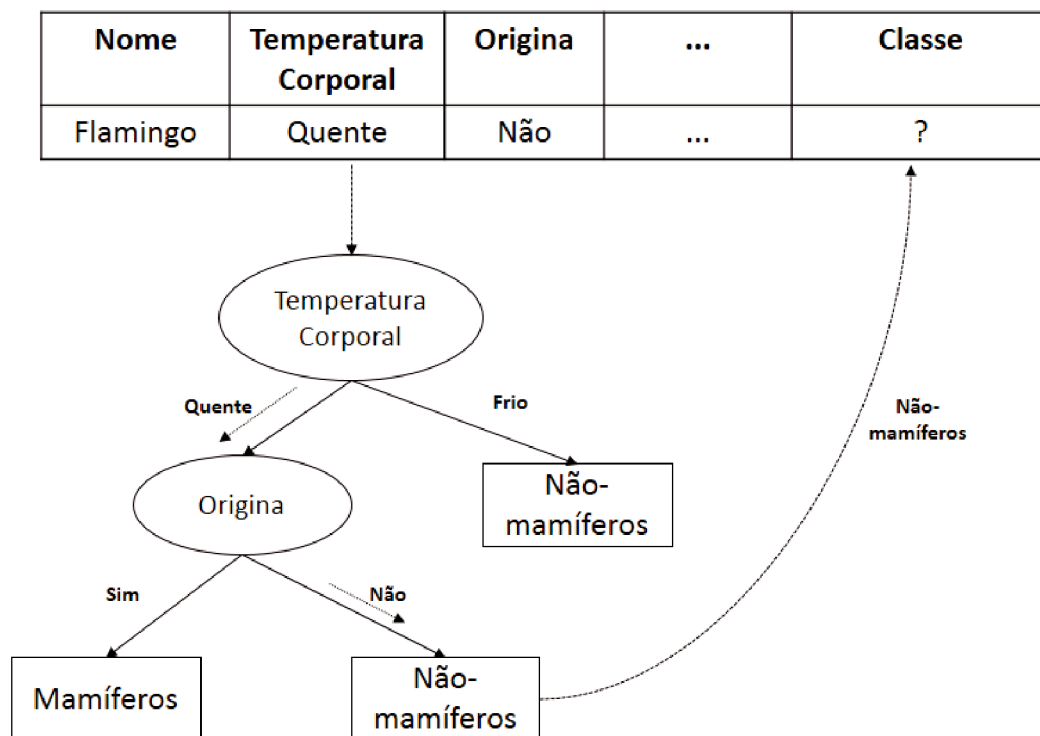


Figura 1 – Árvore de decisão para o problema de classificação de mamíferos.

Fonte: Retirado do livro Tan et al. (2009, p. 179).

Deve-se considera o uso de árvores de decisão em situações que:

- ❑ As instâncias são descritas por pares atributo-valor;
- ❑ A função objeto ou função alvo possui valor discreto;
- ❑ Pode ser necessário hipótese disjuntas;
- ❑ Os exemplos de treino poderão ter erro ou estarem incompletos.

Estas árvores são treinadas de acordo com um conjunto de treino e, posteriormente, outros exemplos são classificados de acordo com essa mesma árvore. Para a construção destas árvores são usados algoritmos como o ID3, ASSISTANT e C4.5.

2.1.2 Aprendizado Bayesiano

Em algumas situações, o rótulo da classe de um registro da base de dados de teste não pode ser previsto com certeza, devido a alguns atributos com ruídos ou da presença de fatores de confusão que afetam a classificação deste registro (TAN et al., 2009). Nestes casos, modelos estatísticos podem encontrar uma boa aproximação do conceito induzido. Dentre os métodos estatísticos, pode-se citar o Aprendizado Bayesiano, que usa uma maneira probabilística e o conhecimento prévio do problema, combinado com os exemplos de treinamento, para saber qual a porcentagem final de uma hipótese (MONARD; BARANAUSKAS, 2003a).

Redes Bayesianas constituem um modelo gráfico que representa de forma simples as relações de causalidade das variáveis de um sistema. Uma Rede Bayesiana consiste do seguinte:

- ❑ Um conjunto de variáveis e um conjunto de arcos ligando as variáveis;
- ❑ Cada variável possui um conjunto limitado de estados mutuamente exclusivos;
- ❑ As variáveis e arcos formam um grafo dirigido sem ciclos;
- ❑ Para cada variável A que possui como pais $B_1, \dots, B_n$, existe uma tabela $P(A | B_1, \dots, B_n)$.

Repare que, caso A não possua um pai, a tabela de probabilidades é reduzida para uma probabilidade incondicional $P(A)$. Uma vez definida a topologia da rede, basta especificar as probabilidades dos nós que participam em dependências diretas, e utilizar estas para computar as demais probabilidades que se deseje.

A probabilidade condicional $P(a|b) = x$, pode ser interpretada como: "dado o evento b , a probabilidade do evento a é x ". Por exemplo, $P(\text{Cárie} | \text{Dor}) = 0.8$, indica que caso um paciente esteja com dor de dente e nenhuma outra informação esteja disponível, então, a probabilidade do paciente ter uma cárie é 0.8. Portanto, a regra de Bayes é definida por: $P(b|a) = P(a|b)P(b)/P(a)$.

2.2 Trabalhos Correlatos

A interatividade fornecida pelos recursos das redes sociais traz ao público infantojuvenil um local onde encontram respostas para as suas dúvidas e curiosidades (SANTIN P. L. ; FREITAS, 2012). Assim, será descrito a seguir os principais trabalhos que subsidiaram o estudo para o desenvolvimento deste trabalho.

O artigo de Santin P. L. ; Freitas (2012) propõe um método para análise de mensagens instantâneas de texto, realizadas em salas de bate-papo na Internet, visando a modelagem

do perfil do agressor e da vítima separadamente. Esta modelagem é baseada na identificação dos estágios do diálogo entre um agressor e sua possível vítima. Esta abordagem permite detectar a correlação entre os estágios de cada perfil com intuito de melhorar a precisão da detecção de aliciamento e ainda, de maneira análoga, inferir a probabilidade da exposição das vítimas e ao risco do abuso sexual. O trabalho analisa uma base de dados que contém conversas de salas de bate-papo entre duas pessoas, podendo ser uma vítima e a outra o agressor. Como segunda fase do trabalho, houve o pré-processamento do texto das transcrições das conversações, separando as linhas da conversa em enunciados do agressor e enunciados da vítima. Posteriormente, os modelos das falas do agressor e da vítima foram definidos usando as técnicas estocásticas dos Modelos Escondidos de Markov (*Hidden Markov Models (HMM)*) para a modelagem de aliciamento sexual de menores. Com o HMM pode-se obter a probabilidade da ocorrência (susceptibilidade) do aliciamento sem ter todos os estágios da sequência de observação, pois não se sabe a priori o tamanho que uma conversação pode atingir.

O estudo desenvolvido por Santos e Guedes (2019) objetivou identificar predadores sexuais brasileiros por meio de Redes Neurais Convolucionais ao analisar conversas provenientes de provas criminais disponibilizadas recentemente ao público. O conjunto de dados preliminar foi disponibilizado pelo Ministério Público Federal de São Paulo (MPF-SP) em parceria com o Centro Universitário da Fundação Educacional Inaciana (FEI) para realizar a análise textual e detectar a ação de predadores sexuais na Internet por meio de Redes Neurais Convolucionais. Testou-se vários cenários durante a reprodução dos experimentos, e o modelo atingiu taxa de erro zero em até 50 épocas.

No trabalho apresentado por Júnior et al. (2015) foi desenvolvida uma ferramenta inteligente com a utilização de técnicas de Mineração de Dados e Sistemas Multiagentes com a capacidade de classificar o nível de exposição de crianças à ataques dentro da rede social, com base em suas interações no Facebook. O software desenvolvido para combater esses riscos utilizou da plataforma multiagente Jade (Java Agent Development Framework) e da ferramenta de mineração de dados WEKA. O algoritmo utilizado para classificação e aplicação dos testes, foi o J48, elegido por representar os dados em uma árvore binária, facilitando a visualização e interpretação de como o modelo trabalha. Os resultados desse experimento mostram os detalhes sobre recursos utilizados pelas crianças dentro da rede social Facebook e a forma de utilização desses recursos, se estão sendo aproveitados de forma sensível à ameaças virtuais ou não.

Esses trabalhos citados tem relação com o presente trabalho no que diz respeito ao uso de técnicas computacionais com o intuito de identificar aliciadores, predadores, ou seja a susceptibilidade de crianças e adolescentes quando estão expostas na rede. O que diferencia a pesquisa em relação aos trabalhos de Júnior et al. (2015), Santin P. L. ; Freitas (2012), Santos e Guedes (2019) foi a base de dados composta por uma pesquisa feita através de questionários aplicados às escolas com alunos de 9 a 14 anos de idade. E também foram

utilizadas técnicas de AM como a árvore de decisão e o Naive Bayes para tentar classificar se a amostra pesquisada são suscetíveis ou não à ameaças.

Experimentos e Análise dos Resultados

Neste capítulo serão apresentadas os atributos e como foi criada a base de dados utilizada neste trabalho, bem como os experimentos realizados com a mesma. Ademais, para analisar a eficiência dos métodos computacionais de AM e comparar os resultados obtidos, utilizou-se das bibliotecas disponíveis pela ferramenta WEKA e implementou as técnicas de AM na linguagem Python.

3.1 Método para a Avaliação

3.1.1 Criação da Base de dados

A metodologia do presente trabalho seguiu a sequência de atividades apresentadas no organograma da Figura 2.

A base de dados foi criada com informações obtidas por crianças e adolescentes que residem na cidade de Monte Carmelo, no Estado de Minas Gerais. A página do site QEDU (<https://qedu.org.br/cidade/2246-montecarmelo/censoescolar?year=2018dependence=0localization=>) traz informações do número de estudante por série de ensino. As séries pesquisadas foram do 5º ao 9º ano que compreende alunos de 9 a 14 anos, que é o foco da pesquisa. Segundo o site QEDU, o município de Monte Carmelo possui 3182 alunos matriculados. A pesquisa foi feita com uma amostra que totalizaram 498 pessoas, que corresponde a 16% dos alunos de escolas da rede pública e da rede privada, sendo indivíduos do sexo masculino e feminino. Para realização da pesquisa, foi feito um pedido de autorização aos diretores das escolas para aplicação dos questionários. As escolas pesquisadas foram: Escola Estadual Clara Chaves com 67 alunos; Colégio Alpha COC com 110 alunos; Escola Estadual Coronel Virgílio Rosa com 102 alunos; Escola Estadual Dona Sindá com 93 alunos; e Escola Estadual Gregoriano Canedo com 126 alunos. Desta quantidade descrita, 22,09% refere-se à alunos da rede particular e o restante da rede pública de ensino.

Os alunos preencheram o questionário impresso sem a identificação dos mesmos, e as instruções passadas foram a importância da sinceridade e veracidade nas respostas. O

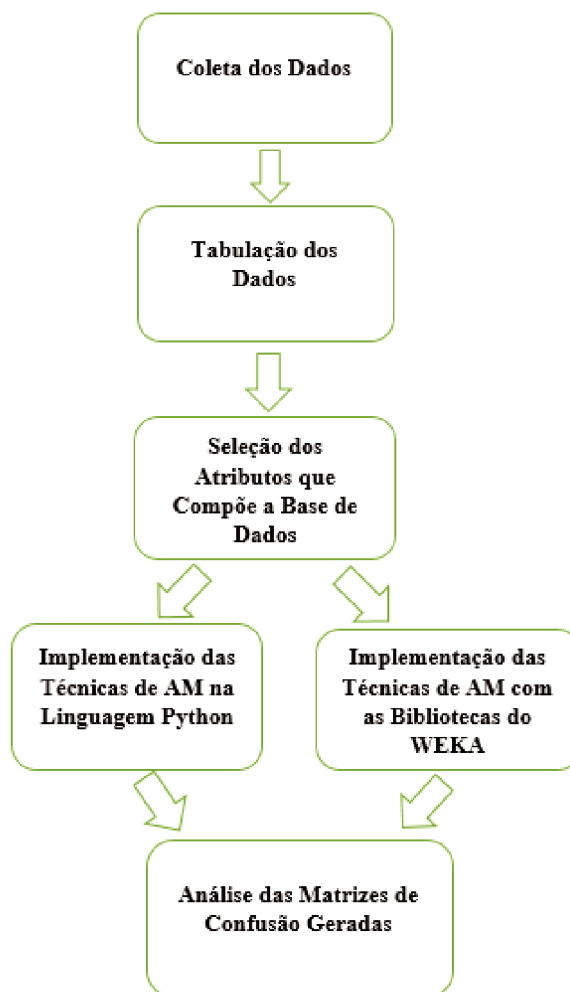


Figura 2 – Fluxograma da metodologia para classificar se há ou não susceptibilidade de ameaças na forma de uso as redes sociais.

tempo estimado para o preenchimento foi cerca de quinze minutos por sala de aula. O questionário entregue aos alunos está anexado no Apêndice A.1.

Após a coleta dos dados nas cinco escolas, foi realizada a análise das informações contidas nas respostas dos questionários, organizando e tabulando os dados em planilhas eletrônicas. O software utilizado para organizar as respostas foi o PSPP, um software livre direcionado à cálculos estatísticos (www.gnu.org/software/pspp/get.html) que teve a finalidade de demonstrar os dados de estatística descritiva. Assim, foi possível identificar o perfil dos alunos de acordo com a sua idade, escolaridade, quais redes sociais que utilizam, a frequência de uso nas redes sociais e outros dados. Esses dados puderam ser representados por gráficos, os quais serão exibidos a seguir para tornar mais legível sua interpretação.

A Figura 3 exibe a relação entre a idade dos entrevistados e a rede social que mais utilizam. Dentre os 240 entrevistados com idade entre 9 a 10 anos, 73,5% utilizam o Youtube, e em segundo lugar, com 11,3%, o Facebook. Isso pode levantar a questão de que o conteúdo que as crianças e os adolescentes querem assistir, sem horário predeterminado,

justifica a alta audiência do Youtube e a substituição dos programas de televisão por esta rede social. Em contrapartida, a rede social menos utilizada por essa faixa de idade é o Twitter, com 0,5%, abaixo do Instagram, com 6,9% dos alunos pesquisados.

Pré-adolescentes com idade entre 11 e 12 anos, totalizaram 128 alunos pesquisados, o que corresponde a 25,8% do total da amostra analisada. Estes pré-adolescentes também utilizam com mais frequência o canal do Youtube, com um total de 47,7%, seguido pelo Facebook com 21,9% .

A faixa etária dos 13 aos 14 anos, totaliza 163 alunos e corresponde a 32,8% do conjunto total. Estes adolescentes utilizam em primeiro lugar o Facebook, com 39,3%, seguido do canal Youtube com 27% e em seguida a rede social Instagram com 15,3% dos entrevistados.

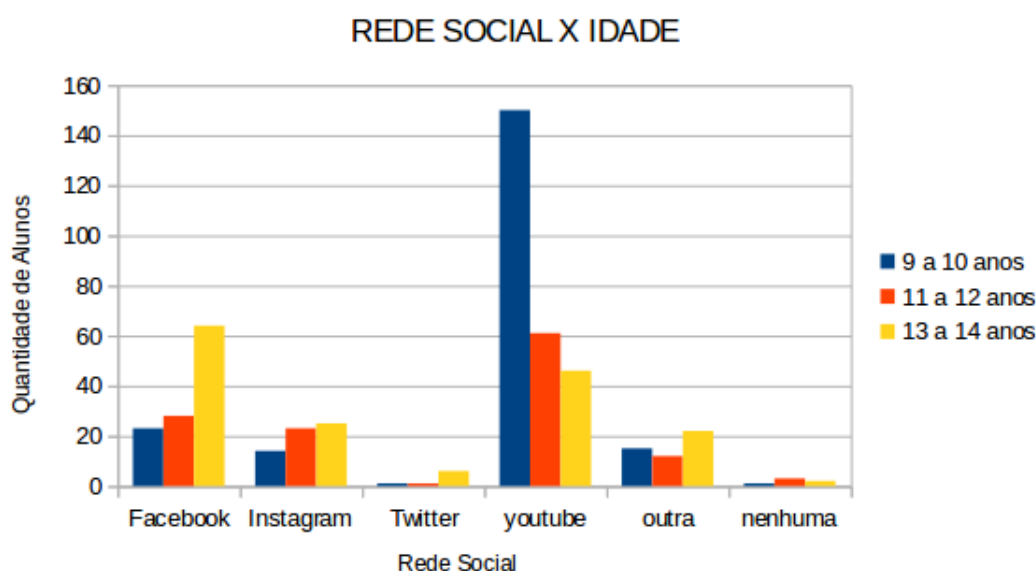


Figura 3 – Relação entre a idade dos entrevistados e as redes sociais que mais utilizam.

A Figura 4 mostra a quantidade de alunos que possuem Internet em casa. Ao analisar a figura 4 destaca que dentre os alunos de 9 aos 14 anos, 88,76% tem acesso a Internet. Esses dados demonstram que as crianças de hoje têm crescido em um ambiente com maior infraestrutura tecnológica e com acesso a informação do que a que seus pais vivenciaram. Assim, tais fatos enfatizam que a Internet faz parte do cotidiano das pessoas.

A Figura 5 exhibe qual o meio eletrônico que os alunos envolvidos na pesquisa utilizam para acessar as redes sociais. Os resultados obtidos, segundo a Figura 5, foram que o acesso as redes sociais pelas crianças de 9 a 10 anos são: 60,9% usam smartphone; 40,6% usam computador e smartphone; e 37,1% apenas o computador. O restante não possui acesso. De acordo com a faixa etária de 11 a 12 anos, 31,6% utilizam smartphone, 24,3% utilizam computador e smartphone e, 15,6% utilizam somente o computador. No que diz

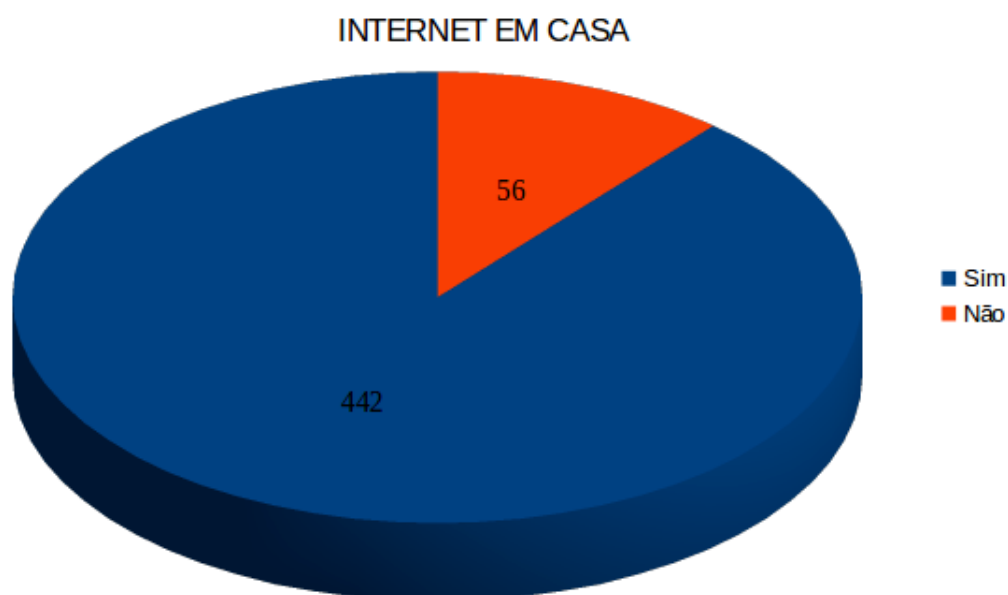


Figura 4 – Quantidade de alunos que possuem Internet em casa.

respeito aos entrevistados de 13 a 14 anos, 37,9% utiliza o smartphone, 27,7% utilizam ambos, computador e smartphone e, 23,4% utilizam computador.

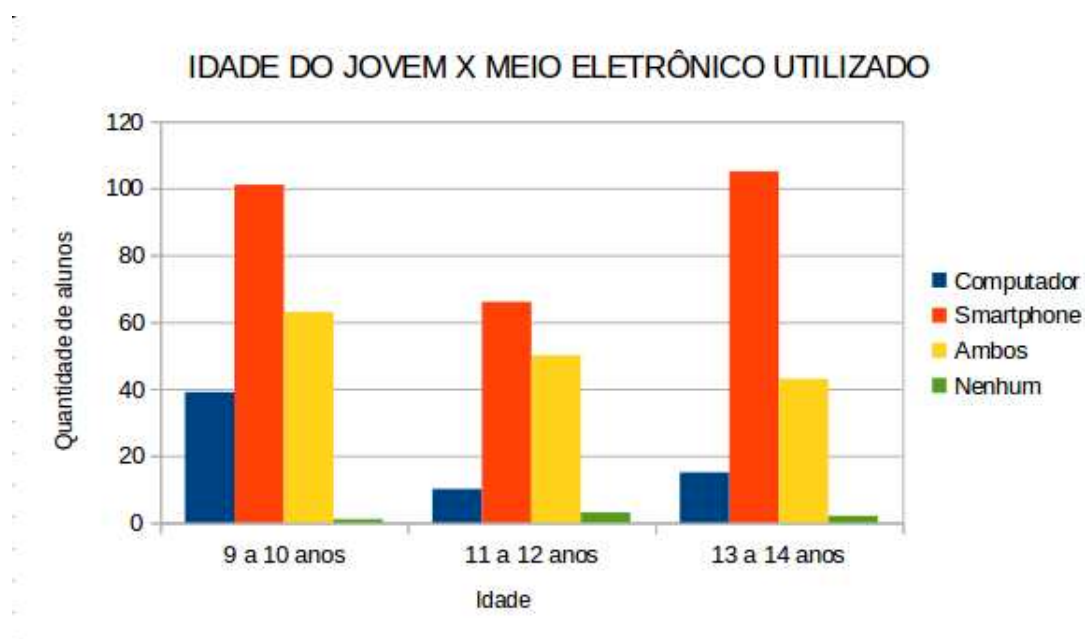


Figura 5 – Meio eletrônico mais utilizado.

A Figura 6 exibe os resultados gerados ao observar se os entrevistados já passaram por alguma situação constrangedora nas redes sociais.

Para atender os objetivos deste estudo, foram analisados e filtrados alguns dados gerados pela entrevista com os alunos da cidade de Monte Carmelo. São eles: idade, a rede social que mais utiliza, frequência de acesso, tempo de interação contínua, controle de

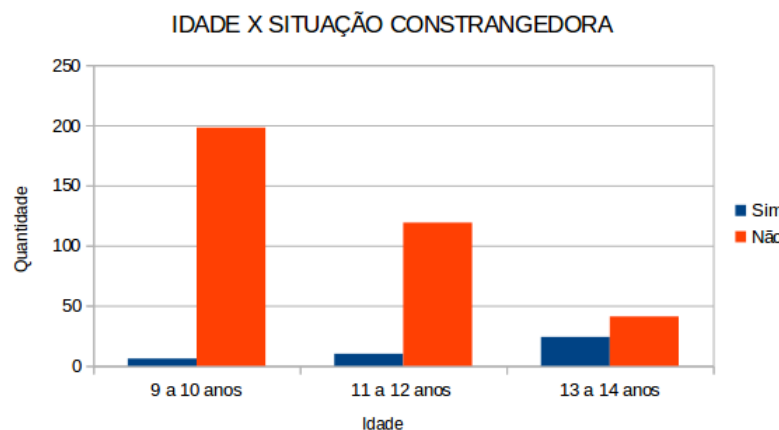


Figura 6 – Relação entre idade e se passaram por alguma situação constrangedora nas redes sociais.

acesso feito pelos pais. Estes atributos compõem a base de dados, a qual foi analisada pelas técnicas de AM que identificaram se os entrevistados são susceptíveis ou não à alguma situação constrangedora pelas redes sociais.

As técnicas de AM testadas na base de dados foram a AD e RB. Estas duas técnicas foram implementadas utilizando o WEKA, com o intuito de entender as variáveis necessárias para a construção destes algoritmos e de gerar resultados rápidos para interpretá-los. Após o uso do WEKA, implementou, novamente, uma AD e uma RB na linguagem Python, com o propósito de comparar os resultados gerados entre as duas ferramentas computacionais.

3.1.2 WEKA

O WEKA possui uma coleção de algoritmos de Aprendizagem de Máquina escrito na linguagem de programação Java, e que foi criado por pesquisadores da Universidade de Waikato, Nova Zelândia (SCHMITT, 2013). Esta ferramenta consiste de vários programas, o que denominaremos de bibliotecas, dos principais algoritmos e técnicas de AM disseminados, tais como redes bayesianas, máquina de vetores de suporte, árvores de decisão, redes neurais artificiais, entre outras (SCHMITT, 2013). Para poder utilizar destas bibliotecas, é necessário definir dentro da base de dados, o conjunto de treinamento e o conjunto de testes, sendo necessário a base de dados estar num formato aceitável, ou seja, o formato ARFF (Attribute Relation File Format). Para tal, a base de dados construída neste trabalho foi editada no programa PSPP e modificada no Microsoft Excel, para que os dados fossem salvos na extensão CSV, o que também é uma extensão aceita pelo WEKA.

Na própria ferramenta WEKA, transformou-se o arquivo de extensão CSV para a extensão ARFF, acessando no menu *"Tools » ARFFVIEWER"*. Cada linha da base de

dados é uma instância, formada por atributos que estão organizados nas setes primeiras colunas. A última coluna é a classe, ou seja, o rótulo que se deseja alcançar com as técnicas de AM. Ou seja, a última coluna contém a classe que determina se o público infantojuvenil já passou ou não por alguma situação constrangedora ao usar as redes sociais. A Figura 7 mostra a tela da ferramenta WEKA.

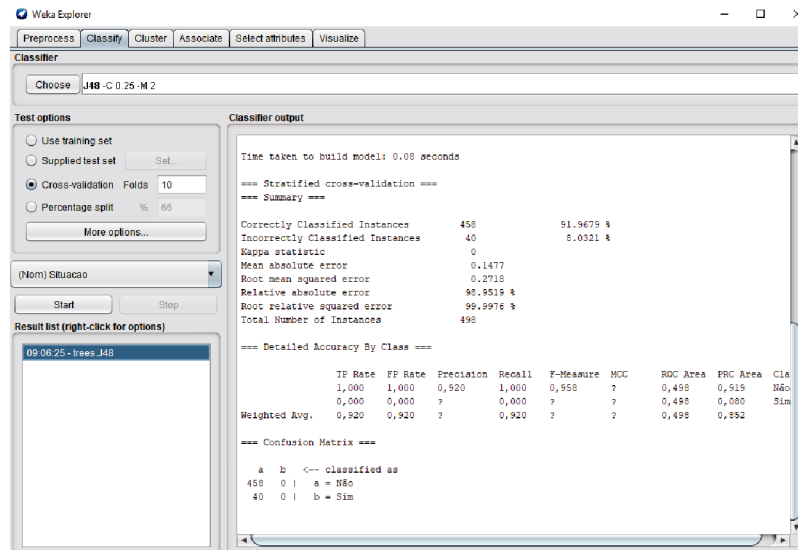


Figura 7 – Tela do WEKA.

3.1.3 As Técnicas de Aprendizagem de Máquinas

Com os dados em formato ARFF, buscou-se a escolha dos classificadores. A ferramenta Weka oferece vários classificadores: árvores de decisão, redes neurais, SVMs (*Support Vector Machines*), k-NN (k-Nearest Neighbors) e Naïve Bayes. Foi feita a opção pelas árvores de decisão e pelo Naive Bayes, devido às situações de incertezas que a base de dados caracteriza-se.

Ademais, a escolha pelo algoritmo de árvore de decisão se deu pelo fato de ser um método de aprendizado de máquina supervisionado não-paramétrico (que significa que não assume nada sobre os dados de antemão), muito utilizados em tarefas de classificação e regressão. A árvore de decisão é de fácil implementação e também não requer nenhum conhecimento estatístico para sua interpretação. São instáveis, pois pequenas alterações produzem novas árvores.

Por outro lado, a opção pelo algoritmo Naive Bayes deve-se pelo algoritmo ser muito simples e rápido, e possuir um desempenho relativamente maior do que outros classificadores. Ele só precisa de um pequeno número de dados de teste para fazer classificações com uma boa precisão. O algoritmo Naive Bayes também desconsidera a correlação entre as variáveis (features). Assim, por exemplo, se uma fruta é considerada uma laranja e se ela for de cor “laranja”, “redonda” e possuir “aproximadamente 10cm de diâmetro”,

o algoritmo não vai levar em consideração a correlação entre esses fatores. Ou seja, ele tratará cada atributo de forma independente.

Na implementação da árvore de decisão, utilizou o algoritmo J48. Este algoritmo surgiu da remodelagem do algoritmo C4.5, que foi escrito na linguagem C para a linguagem Java (WITTEN et al., 2016), e tem o objetivo de gerar uma árvore de decisão baseada no conjunto de dados de treinamento, e usa o modelo para classificar as instâncias do conjunto de teste. Para o autor Vieira et al. (2018), o J48 é um algoritmo de código aberto que usa a abordagem “dividir para conquistar” na montagem da árvore, baseando-se no ganho de informação identificado por meio da redução de entropia. Isto é, quanta informação um atributo traz, para então selecionar a separação ótima.

Desta forma, um problema complexo é decomposto em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe (WITTEN et al., 2016). O algoritmo J48 pode ser usado para atributos contínuos e discretos bem como para valores categóricos e ausentes.

Já o Naive Bayes é baseado em métodos bayesianos, que trata de um algoritmo de classificação supervisionada. Possui a capacidade de produção de estimativas de probabilidade ao invés de simples classificações. Ou seja, para cada rótulo de classe, o classificador pode gerar uma estimativa de um novo objeto pertencer à mesma.

O Weka possui algumas opções de separação da base de dados para realizar o treinamento e teste dos algoritmos selecionados. São eles:

- ❑ *Use training set*: para construir o modelo desejado, simplesmente usar o conjunto de dados que foi fornecido no arquivo ARFF;
- ❑ *Supplied test set*: fornece um conjunto diferente de dados para construir o modelo;
- ❑ *Cross-validation* : constroi um modelo baseado em subconjuntos dos dados fornecidos e calcula sua média para criar um modelo final;
- ❑ *Percentage split*: toma um subconjunto percentual dos dados fornecidos para construir um modelo final.

A validação cruzada promove através de iterações um aumento das comparações entre os poucos elementos da base de dados, incrementando o treinamento e levando a um resultado mais satisfatório (WITTEN et al., 2016). Para o presente trabalho foi utilizado a validação cruzada com 70% para o conjunto de treinamento e 30% para o conjunto de teste.

Para comparar os resultados obtidos pelas bibliotecas do WEKA, foi implementado a árvore de decisão, bem como a rede Bayesiana na linguagem Python, com intuito de entender a lógica dos algoritmos propostos e analisar suas eficiências. Python é uma linguagem de programação de alto nível, dinâmica e com propósitos gerais. A linguagem

foi concebida no final dos anos 1980 e seu principal autor é Guido van Rossum. É muito eficiente, pois consegue-se fazer mais coisas com menos linhas de código. A sintaxe é simples e consistente, e suas bibliotecas são amplas e abrangentes.

Os resultados gerados pelos algoritmos serão exibidos usando a matriz de confusão, que ilustra a precisão da solução para um problema de classificação. A Matriz de confusão é uma matriz de valores reais e valores preditos pelo seu classificador. Em outras palavras, é uma forma intuitiva de saber como seu classificador está se comportando.

Ademais, será utilizado o cálculo da Acurácia, que consiste na proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo (CLESIO, 2014). Ela é dada pela seguinte fórmula:

$$ACURACIA = \frac{TOTALDEACERTOS}{TOTALDEDAOSNOCONJUNTO}$$

$$ACURACIA = \frac{(VP + VN)}{(P + N)},$$

onde VP corresponde ao verdadeiro positivo; VN corresponde ao verdadeiro negativo; P são os positivos e N negativos.

3.2 Experimentos

A finalidade do presente estudo foi a construção de um sistema para classificar se crianças e adolescentes que fazem o uso das redes sociais da Internet já foram alvos de situações constrangedoras. Ou seja, construir um classificador que em sua fase de treinamento, identifique nos atributos informados um padrão que possa indicar quais usuários infantojuvenis estarão em situação de constrangimento ou não.

A razão pela qual foi decidido usar a plataforma Weka é que além de ser uma ferramenta simples e intuitiva, através dela é possível testar diferentes algoritmos de AM para saber qual funcionará melhor para o sistema em questão. E também testar algumas técnicas de pré-processamento de dados e ver como o modelo se comportou.

Para tal, foi implementado na linguagem Python a árvore de Decisão com o código que pode ser visto no Apendice A.2. Já as bibliotecas e o código usado na implementação do Naive Bayes no Python pode ser visualizado no Apendice A.3.

Os resultados gerados pela árvore de decisão, implementado na linguagem de programação Python, gerou 90,67% de acurácia do modelo. Já o algoritmo Naive Bayes, apresentou 91.66% de acurácia.

Por conseguinte, obteve-se no WEKA os seguintes resultados usando a árvore de decisão com o algoritmo J48. O total de instâncias classificadas corretas na árvore de decisão foram 458, que correspondem a 91.96%. Já as instâncias classificadas incorretas foram 40, totalizando 8.03%. A matriz de confusão destes resultados está exibida na Tabela 1:

Assim, a Tabela 1 mostra que:

Não Passou	Passou	
458	0	Não Passou
40	0	Passou

Tabela 1 – Matriz de confusão dos resultados gerados pela árvore de decisão no WEKA.

- ❑ O modelo classificou 458 instâncias que representa os alunos que não passaram por constrangimento e que realmente não passaram (verdadeiros/positivos);
- ❑ O modelo classificou 0 instâncias que indica a quantidade de alunos que passaram por constrangimento e que na verdade não passaram (falsos/positivos);
- ❑ O modelo classificou 40 instâncias que representam alunos que não passaram por constrangimento, mas que na verdade passaram por constrangimento (falsos/negativos);
- ❑ O modelo classificou 0 instâncias que indicam alunos que passaram por constrangimento e que realmente passaram por constrangimento (verdadeiros/negativos).

Por conseguinte, foi testado no WEKA o classificador Naive Bayes. O total de instâncias classificadas corretas foram 453 que corresponde a 90.9639% de acertos. Já as instâncias classificadas como incorretas chegam ao número de 45, totalizando 9.0361% de erros. Esses resultados podem ser visto na matriz de confusão apresentada na Tabela 2:

Não Passou	Passou	
449	9	Não Passou
36	4	Passou

Tabela 2 – Matriz de confusão dos resultados gerados pelo algoritmo Naive Bayes no WEKA.

Assim, a tabela 3 mostra que:

- ❑ O modelo classificou 449 instâncias como alunos que não passaram por constrangimento e que realmente não passaram (verdadeiros/positivos);
- ❑ O modelo classificou 9 instâncias que indicam a quantidade de alunos que passaram por constrangimento e que na verdade não passaram (falsos/positivos);
- ❑ O modelo classificou 36 instâncias como alunos que não passou por constrangimento, mas que na verdade passaram (falsos/negativos);
- ❑ O modelo classificou 4 instâncias como alunos que passaram por constrangimento e que realmente passaram (verdadeiros/negativos).

A acurácia obtida pela árvore de decisão implementada no WEKA foi de 91.95%. Enquanto que a acurácia obtida pelo algoritmo Naive Bayes no WEKA foi de 90.96%.

3.3 Avaliação dos Resultados

Os resultados obtidos com os métodos de AM tanto da biblioteca WEKA, quanto os implementados pela linguagem Python, apresentaram resultados semelhantes ao comparar os valores da acurácia em cada caso. Esses valores podem ser confirmados pelos números mostrados na Tabela 3.

	WEKA	Python
Árvore de Decisão	91.95%	90,67%
Algoritmo Nayve Bayes	90.96%	91.66%

Tabela 3 – Grau de acurácia dos classificadores.

Os resultados apresentados na Tabela 2 mostram porcentagens de acertos significativos para ambos os algoritmos de AM. No entanto, é importante destacar que a quantidade de instâncias da base de dados foi pequena, o que limitou ambos os algoritmos de AM durante as fases de treinamento e teste.

Considerações Finais

Com a crescente propagação da Internet e sua intensa utilização pelos mais variados públicos da sociedade surgem conflitos e problemas sociais a serem enfrentadas. A exposição das crianças e adolescentes nas redes sociais podem colocá-las em risco, que vão desde passar por uma situação constrangedora até se tornarem alvo ao alcance de pedófilos.

Durante a realização da enquete nas escolas da cidade de Monte Carmelo, alguns professores destacaram a relevância deste assunto nos tempos atuais. Uma professora relatou que viu o rendimento de sua sala de aula cair, justamente, pelos alunos passarem várias horas jogando na Internet ou assistindo vídeos, o que foi confirmado por depoimentos dos próprios pais em reunião escolar. Outra professora reconheceu que ela mesma não tem como controlar o acesso de seu filho de 11 anos à Internet, devido a quantidade de tempo gasto por ele na frente do computador.

A pesquisa realizada mostrou que há um grande número de crianças e adolescentes cada vez mais inseridos na tecnologia digital. A adoção do uso de redes sociais por parte desses jovens, aliada à facilidade do acesso aos dispositivos eletrônicos, têm tornado um cenário preocupante no que diz respeito a informações, exposição e amizades que são trocadas quando estão online.

As políticas de segurança das redes sociais têm feito sua parte ao restringir acessos aos menores de treze anos. Entretanto, o que se pode notar que, mesmo em número menor, há a reincidência de possuírem contas falsas nesses sites. Uma vez cadastrados, passam a utilizar as redes sociais de forma irresponsável.

4.1 Principais Contribuições

O trabalho desenvolvido procurou demonstrar pelas técnicas de AM que é possível identificar e classificar através de características relacionadas à forma de uso das redes sociais, se o público infantojuvenil passou ou não por situações constrangedoras em uma base de dados específica. Desta forma, os resultados alcançados neste trabalho são um subsídio ao desenvolvimento de estudos mais minuciosos que enfatizam metodologias ca-

pazes de identificar possíveis vulnerabilidades na Internet, priorizando ações de medidas protetivas, principalmente à crianças e adolescentes.

O estudo mostra grande desafio aos pais e responsáveis à fiscalização do conteúdo encontrado no mundo virtual, pois é um ambiente em constante transformação. É preciso a implementação de políticas públicas de forma a prevenir e conscientizar a todos os desenvolvedores de conteúdo digital, sobre a segurança das crianças e adolescentes.

4.2 Trabalhos Futuros

Como trabalhos futuros, sugere melhorar a base de dados, fazendo um levantamento detalhado dos comportamentos dos alunos na mesma faixa etária analisada. Este aperfeiçoamento, poderia atrelar a participação de outros profissionais como psicólogos, pedagogos, desenvolvedores de sistemas para Web e outros, para melhor abrangência dos dados a serem levantados.

Ademais, O método desenvolvido neste trabalho pode ser aperfeiçoado aumentando as instâncias e os atributos da base de dados, o que tornaria viável testar e analisar outras técnicas definidas pelo aprendizado de máquina.

Sugere também a criação de uma interface visual, onde qualquer pessoa, por meio de um aplicativo poderá escolher entre opções os valores dos atributos e como saída obterá a resposta se a criança encontra-se suscetível ou não a riscos na Internet.

Outro estudo futuro proposto seria a busca mais detalhada de trabalhos semelhantes já realizados em outros países, visando trazer novas ideias e melhorias para a segurança das redes sociais e para os aplicativos Web destinados ao público infantojuvenil brasileiro.

Referências

- AMORIM, M. J.; BARONE, D.; MANSUR, A. U. Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2008. v. 1, n. 1, p. 666–674. Citado na página 12.
- BATISTA, G. E. d. A. P. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003. Citado 2 vezes nas páginas 15 e 16.
- BERSON, I. R.; BERSON, M. J. Challenging online behaviors of youth: Findings from a comparative analysis of young people in the united states and new zealand. **Social Science Computer Review**, Sage Publications Sage CA: Thousand Oaks, CA, v. 23, n. 1, p. 29–38, 2005. Citado na página 11.
- CLESIO, F. Métricas de avaliação de modelos de classificação/predição. **Brasil:[sn]**, 2014. Citado na página 28.
- FLEMING, M. J. et al. Safety in cyberspace: Adolescents' safety and exposure online. **Youth & Society**, Sage Publications Sage CA: Thousand Oaks, CA, v. 38, n. 2, p. 135–154, 2006. Citado na página 11.
- FREITAS, C. M. D. S. et al. Extração de conhecimento e análise visual de redes sociais. In: **Anais do XXVIII Congresso da Sociedade Brasileira de Computação**. [S.l.: s.n.], 2008. p. 106–120. Citado na página 13.
- JÚNIOR, M. S. R. F. et al. Análise comportamental para proteção da criança nas redes sociais por meio de mineração de interações e sistemas multiagentes. In: **Anais do XII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.: s.n.], 2015. Citado 3 vezes nas páginas 11, 12 e 19.
- LORENA, A. C.; CARVALHO, A. C. P. d. L. F. **Introdução às máquinas de vetores suporte (support vector machines)**. [S.l.]: ICMC-USP, 2003. Citado 2 vezes nas páginas 15 e 16.
- MACHADO, I. V.; MOCHI, T. d. F. G. Da proteção da integridade psicológica no âmbito familiar brasileiro: Novas perspectivas de compreensão e intervenção. **Espaço Jurídico Journal of Law [EJJL]**, v. 14, n. 2, p. 387–416, 2013. Citado na página 11.

- MARQUES R. L.; DUTRA, I. Redes bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. **Coppe Sistemas—Universidade Federal do Rio de Janeiro**, p. 1–22, 2002. Citado na página 16.
- MITCHELL, T. M. Does machine learning really work? **AI magazine**, v. 18, n. 3, p. 11–11, 1997. Citado na página 15.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: **Sistemas Inteligentes Fundamentos e Aplicações**. 1. ed. Barueri-SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204-168. Citado 2 vezes nas páginas 16 e 18.
- _____. Indução de regras e árvores de decisão. **Sistemas Inteligentes. Rezende, SO Editora Manole Ltda**, p. 115–140, 2003. Citado na página 16.
- _____. Indução de regras e Árvores de decisão. In: **Sistemas Inteligentes - Fundamentos e Aplicações**. 1. ed. Barueri-SP: Manole Ltda, 2003. p. 115–140. ISBN 85-204-168. Citado na página 15.
- PEREIRA, S. E. F. N. **Redes sociais de adolescentes em contexto de vulnerabilidade social e sua relação com os riscos de envolvimento com o tráfico de drogas**. Tese (Doutorado) — Universidade de Brasília, 2009. Citado na página 13.
- RODRIGUES, L. M. **As crianças e os media: análise de discursos dirigidos aos pais**. Tese (Doutorado), 2015. Citado na página 11.
- SANTIN P. L. ; FREITAS, C. O. A. . P. E. . S. A. Modelagem de aliciamento de menores em mensagens instantâneas de texto. In: **Anais do XII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais**. [S.l.: s.n.], 2012. p. 288–301. Citado 4 vezes nas páginas 11, 12, 18 e 19.
- SANTOS, L.; GUEDES, G. Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na internet. In: **Anais do VIII Brazilian Workshop on Social Network Analysis and Mining**. Porto Alegre, RS, Brasil: SBC, 2019. p. 143–154. ISSN 2595-6094. Disponível em: <<https://sol.sbc.org.br/index.php/brasnam/article/view/6556>>. Citado 3 vezes nas páginas 11, 12 e 19.
- SCHMITT, V. F. Uma análise comparativa de técnicas de aprendizagem de máquina para prever a popularidade de postagens no facebook. 2013. Citado na página 25.
- SMOLA, A.; VISHWANATHAN, S. **Introduction to Machine Learning**. Cambridge, UK: Syndicate of the University of Cambridge, 2008. Citado na página 12.
- TAN, P. et al. **Introdução ao datamining: mineração de dados**. Ciencia Moderna, 2009. ISBN 9788573937619. Disponível em: <<https://books.google.com.br/books?id=69d6PgAACAAJ>>. Citado 3 vezes nas páginas 16, 17 e 18.
- VIEIRA, E. M. de A. et al. Avaliação da performance do algoritmo j48 para construção de modelos baseados em árvores de decisão. **Revista Brasileira de Computação Aplicada**, v. 10, n. 2, p. 80–90, 2018. Citado na página 27.
- WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016. Citado na página 27.

Apêndices

APÊNDICE **A**

A.1 Questionário aplicado aos alunos das escolas de Monte Carmelo-MG



**UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO**

QUESTIONÁRIO PARA LEVANTAMENTO DE DADOS DE CRIANÇAS E ADOLESCENTES NAS REDES SOCIAIS PARA PESQUISA DE TRABALHO DE CONCLUSÃO DE CURSO.

- 1) Qual a sua idade?
 - ☐ de 9 a 11 anos
 - ☐ de 11 a 14 anos
 - ☐ de 13 a 14 anos
- 2) Possui Internet em casa?
 - ☐ sim
 - ☐ não
- 3) Qual a rede social que mais utiliza?
 - ☐ Facebook
 - ☐ Instagram
 - ☐ Twitter
 - ☐ Youtube
 - ☐ Outra:
- 4) Qual o meio utilizado para entrar nas redes sociais? (Se forem ambos, marque as duas opções)
 - ☐ computador
 - ☐ Smarthphone
 - ☐ ambos
- 5) Com que frequência utiliza as redes sociais?
 - ☐ menos de 1 hora por dia

☐ De 1 hora a 2 Horas por dia

☐ Mais de 2 horas por semana

6) Quando acessa sites de redes sociais (como Facebook Youtube, instagram ou Twitter), quanto tempo, aproximadamente, você passa publicando ou conversando com outras pessoas?

☐ 15 minutos (contínuo)

☐ 30 minutos (contínuo)

☐ 1 a 2 horas (contínuo)

☐ 3 a 4 horas (contínuo)

7) Qual é o nível de escolaridade do seu pai?

☐ Da 1a à 4a série do Ensino Fundamental (antigo primário)

☐ Da 5a à 8a série do Ensino Fundamental (antigo ginásio)

☐ Ensino Médio (antigo 2o grau)

☐ Ensino Superior

☐ Especialização

☐ Não estudou

☐ Não sei

8) Qual é o nível de escolaridade da sua mãe?

☐ Da 1a à 4a série do Ensino Fundamental (antigo primário)

☐ Da 5a à 8a série do Ensino Fundamental (antigo ginásio)

☐ Ensino Médio (antigo 2o grau)

☐ Ensino Superior

☐ Especialização

☐ Não estudou

☐ Não sei

9) Quanto é, aproximadamente, a renda familiar mensal? (Marque apenas uma resposta)

☐ Nenhuma renda.

☐ Até 1 salário mínimo.

☐ De 1 a 3 salários mínimos.

☐ De 3 a 6 salários mínimos.

☐ De 6 a 9 salários mínimos.

☐ De 9 a 12 salários mínimos.

☐ De 12 a 15 salários mínimos.

☐ Mais de 15 salários mínimos.

10) Como você classifica o seu conhecimento de Informática (Acesso e pesquisa em site, Uso de aplicativos como editores de texto, editores de apresentações, aplicativos de edição de imagem, abrir e salvar arquivos em diretórios, ...)?

☐ Muito bom.

☐ Bom.

☐ Ruim.

☐ Muito ruim.

11) Seus pais ou responsáveis tem o controle de quem você conversa ou conteúdo que acessa quando está online?

☐ sim

☐ não

12) Em que você acha que as redes sociais influenciam na sua vida?

13) Você já passou por alguma situação que o deixasse constrangido quando esteve online?

☐ sim. Qual:

☐ não.

A.2 Código na linguagem Python da árvore de decisão

import pandas as pd from sklearn.tree # importa a biblioteca sklearn da árvore de decisão

```
import DecisionTreeClassifier
```

```
from sklearn.model_selection
```

```
import train_test_split
```

```
from sklearn import metrics
```

```
import numpy as np
```

```
df_base = pd.read_csv('sample_data/Dados1.CSV')
```

```
df_base.head()
```

```
df_base.info()
```

```
x_train, x_test, y_train, y_test = train_test_split
```

```
(df_base.drop('situacao',axis=1),df_base['situacao'],test_size=0.3)
```

```
x_train.shape, x_test.shape
```

```
y_train.shape, y_test.shape
```

```
clf = DecisionTreeClassifier()
```

```
clf = clf.fit(x_train,y_train)
```

```
clf.feature_importances_
```

```
for feature,importancia in zip(df_base.columns,clf.feature_importances_):
```

```
print(':.format(feature,importancia))
```

```
resultado = clf.predict(x_test)
```

```
resultado
```

```
print(metrics.classification_report(y_test,resultado))
```

A.3 Código na linguagem Python do Naive Bayes

```

import numpy as np
import pandas as pd
import sklearn
from sklearn.naive_bayes
import GaussianNB
from sklearn.model_selection
import cross_validate
import train_test_split
from sklearn.metrics
import accuracy_score
df_base = pd.read_csv('sample_data/Dados1.CSV')
//carrega a base de dados
df_base.shape
df_base.head()
atributos = ['idade', 'rede_social', 'meio_utilizado',
'frequencia', 'tempo_interacao', 'controle_acesso'] //seleciona as variáveis(feature se-
lection
atrib_prev = ['situacao']//variável a ser prevista
x = df_base[atributos].values
y = df_base[atrib_prev].values //cria os objetos
split_test_size = 0.30 //define a taxa de split
x_treino, x_teste, y_treino, y_teste = train_test_split
(df_base.drop('situacao',axis=1),df_base['situacao'],test_size=0.3) //dados de treino
e de teste
modelo_v1 = GaussianNB() //criando o modelo preditivo
modelo_v1.fit(x_treino, y_treino.ravel()) //treinando o modelo
nb_predict_train = modelo_v1.predict(x_treino) print("Exatidão
(Accuracy): 0:4f".format(metrics.accuracy_score(y_treino,
nb_predict_train) * 100))
print() //verifica a exatidão do modelo nos dados de treino
nb_predict_test = modelo_v1.predict(x_teste) print("Exatidão
(Accuracy): 0:4f".format(metrics.accuracy_score(y_teste, nb_predict_test)* 100))
print()//verifica a exatidão do modelo nos dados de teste
print("confusion matrix") //mostra matriz de confusão
print("0".format(metrics.confusion_matrix(y_teste,
nb_predict_test, labels = [1,0])))
print() print("classification report")
print(metrics.classification_report(y_teste,nb_predict_test, labels =[1,0] ))

```