Utilização de aprendizagem de máquina para a identificação de dependência em aparelhos celulares com foco em casos que possam causar reprovação e evasão

Gabriel Santos[1], Felipe Gonçalves dos Santos[1], Aline Rocha[1] e Thiago Reis da Silva[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Piauí – IFPI – Campus Corrente

[2]Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – IFMA – Campus São João dos Patos

gabrielsbti10@gmail.com, felipe.santos@ifpi.edu.br, aline.rocha@ifpi.edu.br, thiago.reis@ifma.edu.br

Abstract. *This work introduces the problems caused by the excessive use of cell phones in modern society and shows that they can cause school problems to students who use it too much. For this, we integrated technology with the use of an artificial intelligence area, machine learning, with data from a specific database. For the proposed task, the Naive Bayes, AdaBoost, SVM, Bagging and Random Forest classifiers were used. At the end of the tests, the SVM classifier presented the best performance in the general context.*

Resumo. *Este trabalho introduzir os problemas causados pelo uso excessivo dos celulares na sociedade moderna, e mostrar que eles podem causar problemas escolares aos alunos que o usa de forma demasiada. Para isto fizemos uma integração da tecnologia com o uso de uma área da inteligência artificial, o aprendizado de máquina, com dados de uma base de dados específica. Para a tarefa proposta foram usados os classificadores Naive Bayes, AdaBoost, SVM, Bagging e Random Forest, ao final dos testes o classificador SVM apresentou o melhor desempenho no contexto geral.*

1. Introdução

A evolução tecnológica no mundo, e particularmente no Brasil, vem sendo importante para o desenvolvimento econômico, cultural e social. Em 2013 menos da metade das residências tinham acesso à internet, por meio de incentivos fiscais o Brasil buscou a inclusão digital, como resultado 70,5% dos domicílios já contam com o uso da internet, sendo 69% por meio de um celular (IBGE, 2017), que grande parte do seu uso é direcionado às redes sociais.

Neste contexto, o uso em excesso de aparelhos telefônicos pode causar problemas tanto físicos quanto emocionais, tais como depressão, ansiedade, impulsividade até a diminuição da capacidade social, entre os problemas físicos se destacam dores de cabeça, nos braços, no pescoço e as contraturas dos músculos dorsais. Além destes, vale ressaltar os problemas em ambientes escolares, onde eles podem ser responsáveis pela queda de desempenho dos alunos que fazem uso de forma demasiada dele.

Através destes dados, pode-se observar como o uso de *smartphones* podem ser prejudiciais aos alunos, pois o mesmo causa diversas distrações a quem os possuem, gerando uma certa dependência, ocasionando assim uma possível queda no desempenho acadêmico dos estudantes. Sendo assim, este artigo busca através do Aprendizado de Máquina (*Machine Learning*), realizar o diagnóstico da dependência por aparelhos *mobiles* através de um conjunto de pesquisas desenvolvidas junto a psicóloga em no Instituto Federal de Educação, Ciência e Tecnologia do Piauí – Campus Corrente.

Aprendizagem de Máquina pode ser definida como uma subárea da Ciência da Computação e da Estatística que estuda e desenvolve algoritmos e técnicas que aprimoram o seu desempenho e acurácia segundo uma medida de erro e com a experiência que é dada por conjuntos de dados de treinamento (MITCHELL, 1997).

Sendo assim, para reportar o trabalho realizado, as demais seções deste artigo encontram-se organizadas da seguinte forma. Na Seção 2 são reportados os trabalhos correlatos a esta pesquisa. Já a Seção 3 aborda os materiais e métodos, a Seção seguinte os resultados e discussões e, por fim, na Seção 5, são apresentadas as considerações finais.

2. Trabalhos correlatos

No estudo de Elden *et. al.*, (2013), foi proposto o desenvolvimento de um novo algoritmo chamado Ada-GA, que utiliza AdaBoost *Ensemble* com um algoritmo genético para construir um modelo preditivo de classificação de desempenho de estudantes do ensino superior. O Ada-GA é um algoritmo de *boosting* que objetiva otimizar a quantidade de classificadores fracos e seus pesos usando algoritmo genético para tal, com o intuito de melhorar o desempenho do comitê de classificadores. Após os vários experimentos utilizando algoritmos como J48, *Naive Update*, KMeans para compor o comitê com o AdaBoost, os autores apresentaram uma tabela de comparação de resultados, onde o AdaBoost teve um percentual de precisão preditiva de 81,85% e o modelo proposto apresentou 82,07% de precisão.

Na pesquisa de Rosales (2017), propôs um projeto no qual foi desenvolvido um *framework* capaz de determinar o grau da severidade de experiências relatadas do *bullying* com o limite de até 140 caracteres, em primeiro momento era classificado com *Support Vector Machine* (SVM) um classificador binário supervisionado, logo após era atribuído o desenvolvimento do sistema de Lógica Fuzzy. Para melhoria dos SVM, eram rotulados múltiplos textos para aumentar a acurácia e precisão das classificações. Com a precisão de 95,42% de acerto na predição, considerado que se precisava de melhora nas classificações do SVM.

No estudo de Manhães *et. al*., (2011), foram utilizadas técnicas de mineração de dados para extrair informações de evasão/retenção da base de dados do SIG@ de estudantes dos cursos de graduação da Universidade Federal do Rio de Janeiro, identificando três classes distintas de alunos. A pesquisa avaliou um conjunto de seis algoritmos de mineração de dados do software Weka: Árvore de decisão (J48 e *Simple Chart*), SVM, Estatístico (*Naive Bayes*), Redes neurais artificiais (MLP), *Ensemble* (AdaBoost), os quais apresentaram uma acurácia entre 70% a 86%. Ao final eles adotaram o *Naive Bayes* que, apesar de não ter sido o melhor classificador, apresentou um rendimento global que atendia aos objetivos da pesquisa.

3. Materiais e métodos

O aprendizado de máquina é um ramo da inteligência artificial que tem como objetivo desenvolver técnicas capazes de ensinar ao computador a aprender e/ou desempenhar determinada tarefa de forma melhor a partir das próprias experiências (MITCHELL, 1997). Dentre as várias abordagens existentes de aprendizado de máquina, utilizemos no estudo a preditiva também conhecida como supervisionada.

A tecnologia de aprendizagem de máquinas pode ser categorizada como: supervisionada ou não supervisionada. Quanto ao algoritmo supervisionado este recebe exemplos de entrada e saídas e a partir disso "aprende" uma regra que mapeia as entradas e saídas. Já no não supervisionado, não são fornecidos exemplos de saída, somente de entrada, então é o próprio algoritmo que define os agrupamentos e tipos de saídas a serem devolvidos (MITCHELL, 1997). Existem diversos tipos de algoritmos de aprendizagem de máquina cada um com uma metodologia e finalidade diferente, por exemplo, *Naive Bayes*, *AdaBoost*, dentre outros. O funcionamento de um algoritmo de aprendizagem de máquina supervisionado pode ser entendido conforme ilustra a Figura 1.

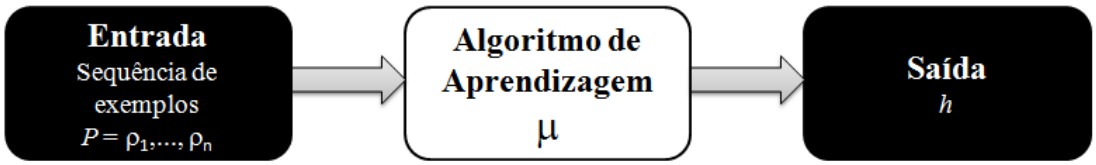

Figura 1: Aprendizado de máquina supervisionado.

3.1. Abordagem Proposta

A abordagem proposta é ilustrada na Figura 2. Para atingir o objetivo do estudo a abordagem inclui as seguintes etapas de: coleta de dados e classificação apresentadas a seguir.

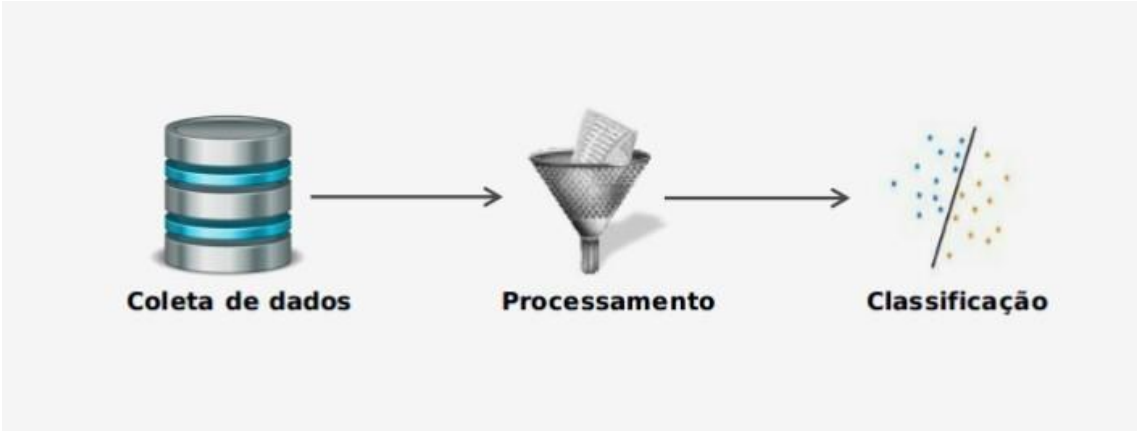

Figura 2: Abordagem proposta.

3.2. Coleta de dados

Os dados usados nesta pesquisa foram recolhidos em 11 turmas do ensino médio integrado ao técnico de IFPI – Campus Corrente, através de um questionário elaborado conjuntamente com a psicóloga da instituição contendo 12 questões relacionadas ao

tema da pesquisa. Ao final obtivemos um total de 334 alunos, sendo 171 do sexo masculino e 163 do sexo feminino.

3.3. Classificação

Nesta etapa submetemos os dados ao algoritmo de aprendizagem máquina. Segundo Paula *et. al.*, (2020), uma característica importante de técnicas supervisionadas é que elas apresentam fundamentalmente o mesmo comportamento: dado um conjunto de dados de treinamento, composto por instâncias que são formadas por vários atributos previamente rotulados (polaridades) nas classes de interesse, o modelo aprende características (base de treino) de cada classe e poderá ser usado para classificar outras amostras (base de teste). Os classificadores adotados foram o *Multinomial Naive Bayes*, *AdaBoostClassifier*, *Bagging*, Máquinas de Vetores Suporte (SVM) e o *Random Forest*.

3.4. Tecnologias utilizadas

Todas as implementações foram desenvolvidas em Python 3. Os algoritmos implementados fazem parte de uma biblioteca *open source* de aprendizado de máquina do Python chamada scikit-learn.

3.5. Validação

Para seleção do melhor conjunto de parâmetros do algoritmo de aprendizado de máquina foram utilizadas as métricas de precisão, revocação e o F-Measure. Nas subseções a seguir discutiremos sobre os critérios citados e sua aplicação, utilizado as seguintes abreviações: TP que são os verdadeiros positivos, FN que representa os falsos negativos e FP que representa os falsos positivos.

3.5.1 Precisão

A precisão é análise, dos que foram classificados como corretas, quantos efetivamente eram de corretas, ou seja, quantos foram classificados correntemente. A fórmula para se obter a precisão de um classificador é apresentado na Figura 3.

$$precisão = \frac{TP}{TP + FP}$$

Figura 3 – Fórmula para se obter a precisão de um classificador.

3.5.2 Revocação

É a proporção de todos os itens relevantes em uma coleção particular ou banco de dados que a busca foi capaz de recuperar. A fórmula para se obter a revocação de um classificador é ilustrado na Figura 4.

$$revocação = \frac{TP}{TP + FN}$$

Figura 4 – Fórmula para se obter a revocação de um classificador.

3.5.3 F-Measure

Essa métrica combina a precisão e a revocação de modo a trazer um número único que indique a qualidade geral do modelo. Sua fórmula é demonstrada na Figura 5.

$$F - Measure = \frac{2 * (precisão * revocação)}{precisão + revocação}$$

Figura 5 – Fórmula para se definir qual o melhor classificador.

4. Resultados e discussões

Após analisarmos os dados coletados pela pesquisa nos algoritmos de aprendizagem de máquina pré-selecionados, obtivemos os resultados apresentados nas Tabelas 1 e 2 respectivamente. Na Tabela 1, temos os dados em que os algoritmos tinham como função avaliar se aquele indivíduo em questão possui uma certa dependência de aparelhos *mobiles*. Na Tabela 2, temos os dados quando os algoritmos tinham como objetivo classificar se a pessoa em questão não apresentava dependência em dispositivos *mobiles*.

Tabela 1 – Resultados dos primeiros anos: caso em que possuem uma possível dependência.

	Precisão	**Revocação**	***F-Measure***	**Taxa de Acerto**
Naive Bayes	0,75	0,115	0,199	56,12
AdaBoost	0,875	0,777	0,823	84,12
SVM	1,0	0,814	0,897	91,22
Bagging	0,777	0,518	0,621	78,94
Random Forest	0,894	0,629	0,738	85,96

Na Tabela 1 temos os dados dos alunos das turmas de primeiro ano do ensino médio, ao analisar os resultados obtidos podemos observar que o algoritmo SVM, foi o que apresentou o melhor resultado no contexto geral, com uma taxa de acerto e *F-Measure* de 91,22% e 0,897 respectivamente. Já nos testes realizados com as turmas de segundo e terceiro ano, o SVM foi novamente o que apresentou os melhores resultados, mas dessa vez tivemos o *Random Forest* apresentando resultados semelhantes ao classificar os alunos de terceiro ano, estes apresentaram uma taxa de acerto de 80%, nos dois casos. Em contrapartida, os piores resultados vieram do algoritmo *Naive Bayes* nas turmas de primeiro e segundo anos com as seguintes taxa de acertos 56,12% e 53,33%, já nos terceiros anos os piores resultados ficaram com os classificadores *AdaBoost* e *Bagging* com uma percentagem de acerto 53,33%.

Tabela 2 – Resultados dos terceiros anos: caso em que não são classificados com dependência.

	Precisão	**Revocação**	***F-Measure***	**Taxa de Acerto**
Naive Bayes	0,6	0,6	0,6	60,0
AdaBoost	0,714	0,4	0,363	53,3
SVM	0,714	1,0	0,833	80,0
Bagging	0,333	0,4	0,336	53,3
Random Forest	0,428	0,6	0,499	80,0

Na Tabela 2 temos os dados dos alunos das turmas de terceiro ano do ensino médio, ao averiguar as informações obtidas podemos constatar que algoritmo que apresentou o melhor desempenho foi o SVM, com uma taxa de acerto e *F-Measure* de 80% e 0,833 respectivamente, vale ressaltar que o classificador *Random Forest* apresentou uma taxa de acerto igual ao SVM, mas na métrica *de F-Measure* pode-se notar que o SVM, obteve um melhor desempenho no geral. O pior desempenho ficou com os algoritmos *AdaBoost* e *Bagging*. Nos testes realizados com as turmas de primeiro e segundo ano, o algoritmo SVM, novamente foi o que apresentou mais apto para a tarefa. Em contraparte os piores resultados vieram pelo lado do algoritmo *Naive Bayes* com taxas baixas em relação aos outros classificadores.

Um fato que pode ser notado após a análise dos resultados foi que os algoritmos apresentaram uma queda de desempenho quando as turmas analisadas eram dos segundo e terceiro anos, isso se deve tanto pelo número de dados para treino serem mais escassos. Outra análise que apresentaremos nesta pesquisa é a porcentagem de homens e mulheres que foram avaliadas com uma possível dependência por dispositivos *mobiles* e a percentagem dos que foram avaliados na situação oposta, ou seja, os que não possuem dependência. Os dados foram divididos de acordo com as turmas de primeiro, segundo e terceiro ano, os dados estão expostos nos Gráficos 1, 2 e 3.

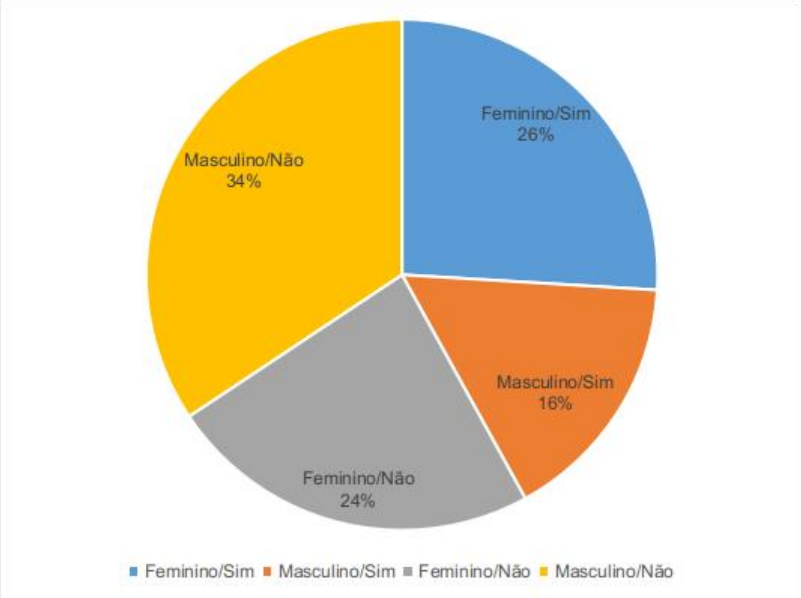

Gráfico 1 – Dados dos alunos do primeiro ano do ensino médio.

Para a realização da classificação do aluno nos perfis abordados na pesquisa, foram levados em conta os padrões que haviam sido discutidos antes da aplicação do questionário com a psicóloga da instituição. Os padrões para a classificação da possível dependência foi: das 12 questões presentes no questionário, ficou definido que se o aluno respondesse, sim em 5 questões, este já se encaixaria no grupo de possível dependência a dispositivos *mobiles*.

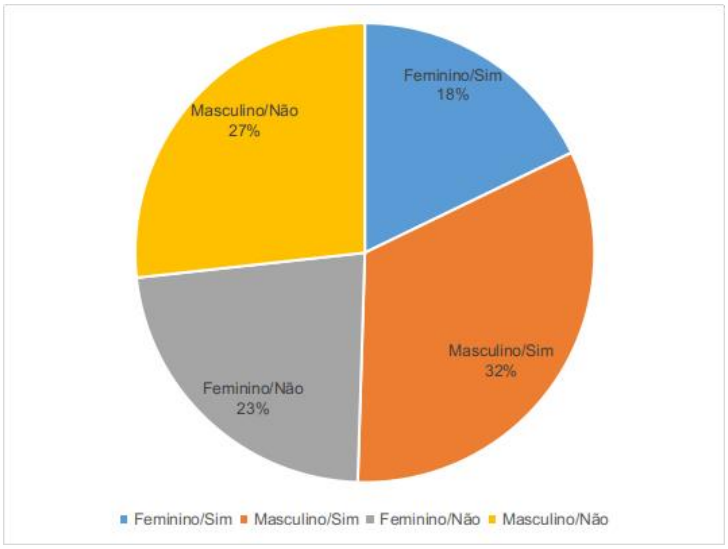

Gráfico 2 – Dados dos alunos do segundo ano do ensino médio.

Nos gráficos 1 e 2, onde estão sendo representados os dados dos alunos de primeiros e segundos anos, o total de pessoas que foram avaliadas como dependentes do *smartphone* e as quais não foram avaliadas como dependentes foi relativamente igual, o fato um tanto quanto preocupante pois boa parte dos alunos se dedicam ao celular como uma prioridade maior que o estudo, isso não apenas prejudica nos resultados das avaliações como também na adesão no mercado profissional para formação técnica. O uso do *smartphone* é ainda mais alarmante nos terceiros anos (ver Gráfico 3), pois os alunos já estão próximos de serem formados na instituição e a um passo do mercado de trabalho, o uso do aparelho é ainda mais crítico, visto que 68% dos alunos destas turmas foram considerados com uma certa dependência a dispositivos *mobiles*.

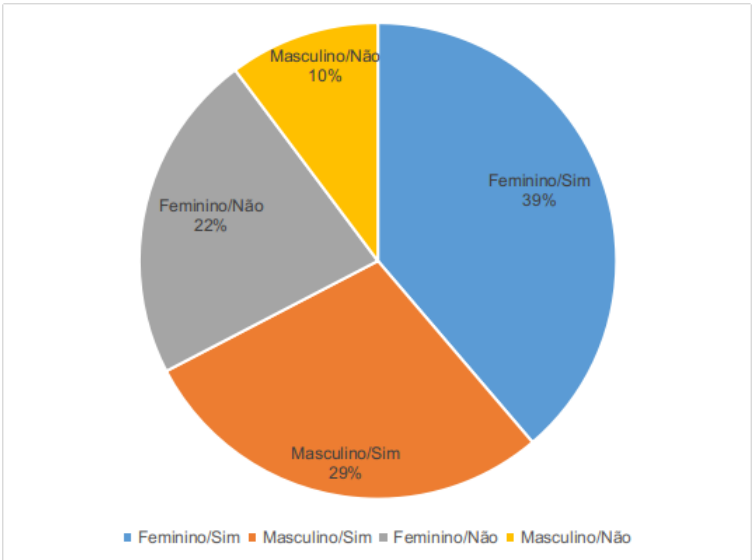

Gráfico 3 – Dados dos alunos do terceiro ano do ensino médio.

5. Considerações finais.

Após todos os teste realizados podemos concluir que o algoritmo SVM, foi o que apresentou os melhores índices entre os cinco classificadores testados, e ele ainda foi o

melhor em todos os cenários testados. Na contramão do SVM, o *Naive Bayes* foi quem apresentou o pior classificado em dois dos três cenários testados. Nos cenários restantes os algoritmos *AdaBoost* e *Bagging* apresentaram a pior colocação.

Outro fator interessante é a integração da tecnologia como uma possível forma sanar um problema que ocorre no ambiente acadêmico, mas também pode impactar de forma positiva na vida da pessoa que aceita que possui uma certa sujeição a aparelhos celulares, pois o uso excessivo do mesmo causa impactos tanto físicos quanto mentais.

Como trabalhos futuros pretendemos aplicar este questionário em outras escolas da região e melhorar ainda mais a nossa base de dados, para que possamos melhorar mais os níveis de precisão dos algoritmos para que os testes possam ser ainda mais confiáveis em todos os cenários possíveis.

References

Breiman, L. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

Breiman, L. Bagging predictors. Machine learning, v. 24, n. 2, p. 123-140, 1996.

Elden, A. S.; Moustafa, M. A.; Harb, H. M.; Emara, A. H. AdaBoost ensemble with simple genetic algorithm for student prediction model. In: International Journal of Computer Science & Information Technology, v. 5.2, p.73. 2013.

IBGE. Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/rendimento-despesa-e-consumo/9127-pesquisa-nacional-por-amostra-de-domicilios.html?edicao=10500&t=resultados>. Acesso em: 20/05/2020.

Mitchell, T. Machine Learning. McGraw Hill, 1997.

Paula, H.; Souza, B.; Nakamura, F.; Nakamura, E. Quantificando a Importância de Emojis e Emoticons para Identificação de Polaridade em Avaliações Online. In: Anais do XIV Simpósio Brasileiro de Sistemas Colaborativos, pp. 228-239, 2020.

Rosales, P. C. Framework para identificação da severidade de bullying baseado em machine learning e lógica fuzzy. Dissertação de Mestrado. Universidade Estadual de Campinas – UNICAMP. 2017.

Santos, A. C. M. Aprendizado de máquina aplicado ao diagnóstico de Dengue. In: Anais do Encontro Nacional de Inteligência Artificial e Computacional, pp. 697-708, 2016.

Soares, R. G. F. Uso de meta-aprendizado para a seleção e ordenação de algoritmos de agrupamento aplicados a dados de expressão gênica. Dissertação de Mestrado. Universidade Federal de Pernambuco – UFPE. 2008.

Zhang, Z.; Xie, X. Research on AdaBoost. M1 with Random Forest. In: 2nd International Conference on Computer Engineering and Technology (ICCET), 2010.