http://journal.unoeste.br/index.php/ce/index
DOI: 10.5747/ce.2020.v12.n3.e327
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 19/08/2020 Revisado: 01/09/2020 Aceito: 08/09/2020

APLICAÇÃO DE MACHINE LEARNING NA IDENTIFICAÇÃO DE E-MAILS COMO SPAM

MACHINE LEARNING APPLICATION TO IDENTIFY E-MAILS AS SPAM

Michelle Tais Garcia Furuya, Danielle Elis Garcia Furuya

Universidade do Oeste Paulista – UNOESTE, Presidente Prudente, SP. E-mail: michellegfuruya@gmail.com

RESUMO - O serviço de e-mail é uma das principais ferramentas utilizadas nos dias de hoje e é um exemplo de que a tecnologia facilita a troca de informações. Por outro lado, um dos maiores empecilhos enfrentados pelos serviços de e-mail corresponde ao spam, nome dado à mensagem não solicitada recebida por um usuário. A aplicação de aprendizado de máquina (machine learning) vem ganhando destaque nos últimos anos como alternativa para identificação eficiente de spam. Nessa área, diferentes algoritmos podem ser avaliados para identificar qual apresenta melhor desempenho. O objetivo deste estudo consiste em identificar a capacidade dos algoritmos de aprendizado de máquina em classificar corretamente os e-mails e identificar também qual algoritmo obteve maior acurácia. A base de dados utilizada foi retirada da plataforma Kaggle e os dados foram processados pelo software Orange com quatro algoritmos: Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Naive Bayes (NB). A divisão dos dados em treino e teste considerou 80% dos dados para treinamento e 20% para teste. Os resultados evidenciam que o Random Forest foi o algoritmo com melhor desempenho com acurácia de 99%.
Palavras-chave: classificação; algoritmos; acurácia.

ABSTRACT – The e-mail service is one of the main tools used today and is an example that technology facilitates the exchange of information. On the other hand, one of the biggest obstacles faced by e-mail services is spam, the name given to the unsolicited message received by a user. The machine learning application has been gaining prominence in recent years as an alternative for efficient identification of spam. In this area, different algorithms can be evaluated to identify which one has the best performance. The aim of the study is to identify the ability of machine learning algorithms to correctly classify e-mails and also to identify which algorithm obtained the greatest accuracy. The database used was taken from the Kaggle platform and the data were processed by the Orange software with four algorithms: Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Naive Bayes (NB). The division of data in training and testing considers 80% of the data for

training and 20% for testing. The results show that Random Forest was the best performing algorithm with 99% accuracy.
Keywords: classification; algorithms; accuracy.

1. INTRODUÇÃO

O e-mail é um meio de comunicação essencial, especialmente nas últimas décadas em que o acesso à Internet foi facilitado e ampliado. O uso da tecnologia traz inúmeros benefícios, mas pode também trazer problemas. O serviço de e-mail, por exemplo, tem o spam como um de seus problemas mais conhecidos. Entende-se por spam qualquer mensagem indesejada cujo envio não foi solicitado pelo usuário que a recebe. É comum identificar um spam como sendo um e-mail padrão enviado sem permissão para uma grande quantidade de pessoas (BASSIOUNI; EL-DAHSHAN, 2018).

As consequências de um spam podem incluir problemas de segurança de dados e exposição a conteúdos impróprios (FARIS, *et al.,* 2018). Além disso, o recebimento contínuo de e-mail de spam pode prejudicar o armazenamento do servidor, a largura de banda e o tempo do usuário. (DADA *et al*., 2019). Outro ponto a ser destacado é que o spam acaba degradando a confiabilidade dos demais e-mails a partir do momento em que o usuário passa a apresentar dúvidas quanto à autenticidade de uma mensagem (ROY; VISWANATHAM, 2016). Por isso a necessidade de separar o spam dos demais e-mails se torna essencial.

Nos últimos anos diversas técnicas estão sendo aplicadas com o intuito de detectar antecipadamente o spam. Um meio de solucionar o problema do spam é o uso de filtros. A filtragem é responsável por reorganizar os e-mails com base em padrões definidos. Segundo Faris *et al.* (2018), um modelo de filtro é a engenharia do conhecimento que funciona por meio de um conjunto de regras específico estabelecido pelo usuário ou empresa que o utiliza. As regras baseiam-se na análise do conteúdo da mensagem, podendo identificar um spam por meio de palavras-chave por exemplo. Entretanto esse método é, muitas vezes, considerado inadequado por necessitar de constantes atualizações, o que além de demandar tempo não garante resultados eficientes. Por isso as técnicas de machine learning (ML) aparecem como alternativa para identificação eficiente de spam, pois não necessitam que regras sejam especificadas e são mais adaptáveis às possíveis mudanças e atualizações. Tal característica é de suma importância visto que os responsáveis pelo envio de spam tendem a modificar as técnicas de envio constantemente como meio de evitar os filtros. As técnicas de ML podem ser aplicadas em diversas áreas e podem contribuir na identificação e solução de diferentes questões.

Como complementação Dada *et al.* (2019), também afirmam que as duas principais abordagens utilizadas são a engenharia do conhecimento e o aprendizado de máquina (Machine Learning). Em concordância com o autor citado anteriormente, Dada *et al.* (2019) afirma que a abordagem de aprendizado de máquina demonstrou ser mais eficiente em relação à engenharia do conhecimento por trabalhar com amostras de treino constituídas, no caso, por mensagens de e-mail pré-classificadas.

A utilização de algoritmos de aprendizado de máquina (Machine Learning - ML) é uma alternativa de otimizar os processos de classificação de dados.

O aprendizado de máquina (ML) é uma subárea da inteligência artificial. A partir da experiência adquirida com um conjunto de dados, o ML verifica a capacidade de diferentes algoritmos aprenderem, ou seja, de melhorar o desempenho (MITCHELL, 1997). A função dos algoritmos é tomar

decisões inteligentes a partir de um conjunto de dados. Para isso é necessário que os algoritmos aprendam a reconhecer padrões complexos (HAN; KAMBER, 2006).

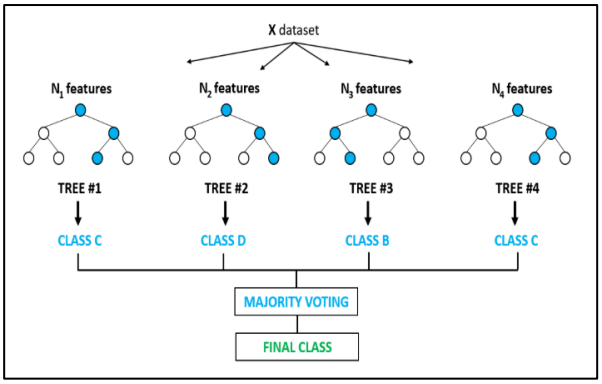
Para se definir um modelo, um conjunto de dados de treinamento é utilizado. Esses dados de treinamento correspondem aos dados cuja classe (rótulo) a qual pertencem já é conhecida. O modelo escolhido é então usado para prever rótulos desconhecidos em outros dados (HAN; KAMBER, 2006).

Os principais serviços de e-mail como Gmail, Outlook e Yahoo utilizam diferentes técnicas de aprendizado de máquina (ML) para resolver a ameaça do spam. Por meio do emprego de ML é possível não somente identificar e-mail de spam, mas também se adaptar a situações variadas. Em outras palavras as técnicas de ML são capazes de gerar novas regras com base no que foi aprendido. Desse modo os novos meios criados para burlar a filtragem também podem ser identificados. O modelo de ML adotado pelo Google, por exemplo, chega a apresentar 99,9% de precisão no processo de detecção e filtragem de spam (DADA *et al.* 2019).

Estudos relacionados ao tema empregam diferentes algoritmos para processar os dados. Quatro dos principais algoritmos de ML aplicados em diversos estudos são o Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Naive Bayes (NB).

Dentre os algoritmos mais recentes está o Random Forest. As vantagens desse algoritmo incluem a possibilidade de lidar com milhares de dados de entrada e aproximação das características mais importantes no processo de classificação. Além disso o algoritmo é capaz de calcular o valor aproximado de dados ausentes com o intuito de manter a precisão. De modo geral o Random Forest apresenta menos erros de classificação e maior acurácia em relação ao Decision Tree (outro algoritmo de mesma estrutura). Já em comparação com o Support Vector Machine (SVM) seus resultados são iguais ou até melhores dependendo do caso (BASSIOUNI; EL-DAHSHAN, 2018; DADA *et al.*, 2019). Outra característica do algoritmo é que este modelo apresenta flexibilidade e rapidez durante a execução. A Figura 1 mostra um exemplo de como o algoritmo RF funciona.

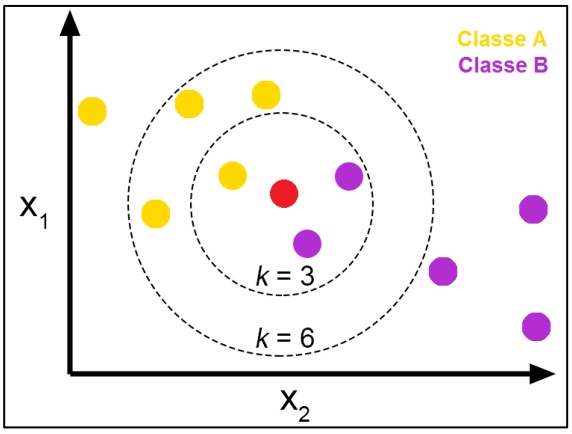
Figura 1. Modelo do algoritmo Random Forest Classifier.



Fonte: Global Software Support (2018).

O método de classificação K-Nearest Neighbors (KNN) baseia-se nas amostras de treinamento mais próximas (BASSIOUNI; EL-DAHSHAN, 2018). Esse algoritmo pode ser usado para classificação ou regressão. O resultado é uma associação de classes onde a classe mais comum entre os vizinhos próximos corresponde à classe cujo objeto em análise pertencerá (KARTHICK, 2017). A Figura 2 mostra um exemplo do algoritmo KNN.

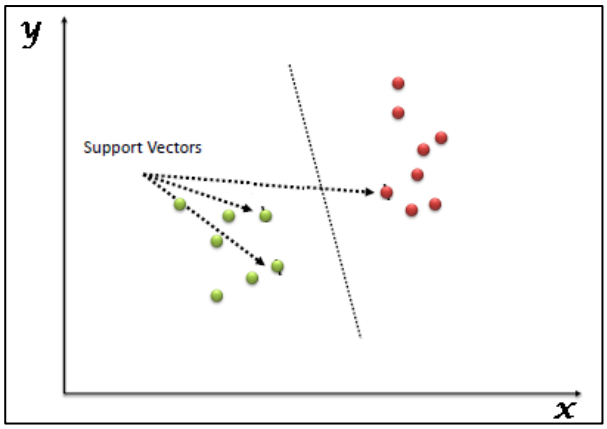
Figura 2. Modelo do algoritmo KNN Classifier.



Fonte: José. I (2018).

Support Vector Machine (SVM) é um algoritmo cuja classificação em classe ou grupo específico é possível por meio da identificação de padrões (DADA *et al.* 2019). A capacidade de adaptação a diversos tipos de dados é uma das vantagens do SVM. A figura 3 mostra um modelo desse algoritmo.

Figura 3. Modelo do algoritmo Support Vector Machine.

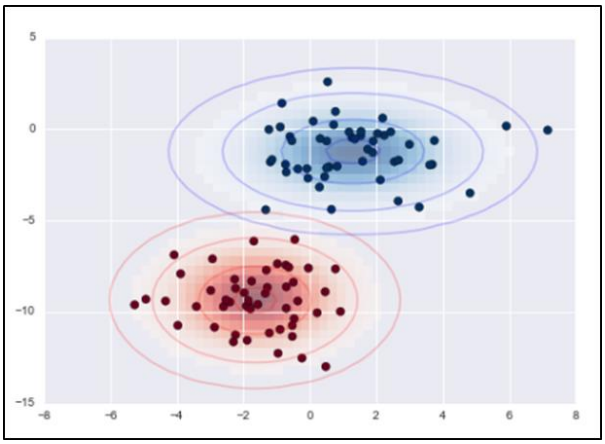


Fonte: Ray (2017).

Roy e Viswanatham (2016), testaram a performance do SVM na identificação de e-mail de spam e concluíram que o algoritmo apresenta alta precisão. Na mesma linha Bassiouni e El-Dahshan (2018), concluíram que o SVM pode apresentar desempenho maior ou igual a outros algoritmos com a vantagem de utilizar menos dados de treinamento para atingir o resultado desejado.

Segundo VanderPlas (2016) o modelo Naive Bayes é um algoritmo de classificação extremamente rápido e simples que geralmente é adequado para conjuntos de dados de dimensões muito altas. Por ser rápido e ter poucos parâmetros ajustáveis, se torna útil para um problema de classificação. A Figura 4 mostra um modelo do algoritmo.

Figura 4. Modelo do algoritmo Naive Bayes Classifier.



Fonte: VanderPlas, (2016).

Devido à quantidade de dados existente atualmente, a utilização de ferramentas que auxiliem no processamento desses dados se torna fundamental. Uma alternativa é o uso de softwares que auxiliam a análise de dados por meio de um conjunto de métodos e algoritmos. Dessa forma os dados são transformados em informação. Diversos softwares estão disponíveis para uso e podem ser utilizados para classificação de dados com auxílio dos algoritmos de aprendizado de máquina. O software Orange é um exemplo e corresponde a um software de programação visual. Por meio dele é feito a mineração de dados, ou seja, processamento de um grande volume de dados em busca de padrões como regras de associação. O processamento é feito por algoritmos de aprendizado de máquina que possibilitam que a análise de dados seja feita posteriormente (NAIK; SAMANT, 2016).

Já o Kaggle é uma plataforma que disponibiliza bases de dados sobre diversos temas. É comum a plataforma ser utilizada para competições de Data Science, sejam elas acadêmicas ou não. Os datasets disponíveis correspondem a dados sobre assuntos diversos em que o usuário pode utilizá-los para treinar as habilidades em machine learning e análise de dados. Constatou-se uma base de dados disponibilizada no Kaggle denominada “Email Spam Classification Dataset CSV” descrita como “CSV file containing spam/not spam

information about 5172 emails" com informações para identificar se um e-mail é ou não spam.

O objetivo deste estudo, portanto, foi identificar a capacidade dos algoritmos de aprendizado de máquina em classificar corretamente os e-mails em spam e não spam e identificar também qual algoritmo, dentre o Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Naive Bayes (NB), obteve maior acurácia.

2. METODOLOGIA

A base de dados obtida no Kaggle tem o título "Email Spam Classification Dataset CSV", apresenta 5172 exemplos de e-mails e 3000 atributos que podem contribuir para um e-mail ser classificado ou não como spam. Os 3000 atributos correspondem a diversas palavras que podem estar inclusas na mensagem.

Após o acesso a base de dados realizou-se a divisão dos dados em treino e teste considerando 80% dos dados para treinamento e 20% para teste (Tabela 1). Após a divisão realizou-se o treinamento dos algoritmos no software Orange. Foram realizados testes com quatro algoritmos de Machine Learning: Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Naive Bayes (NB).

Quadro 1. Quantidade de dados para treino e teste

	QUANTIDADE DE DADOS
TREINO (80%)	4138
TESTE (20%)	1034
TOTAL	5172

Fonte Os autores.

O teste foi programado para repetir dez vezes o desempenho de cada algoritmo para posteriormente verificar a acurácia de cada um, a métrica adotada para verificar a precisão dos algoritmos foi a F1 measure. As repetições nos testes com algoritmos são recomendadas para evitar possíveis erros por parte do algoritmo, pois em alguns casos o algoritmo pode apresentar alta acurácia somente na classificação dos dados utilizados para amostras de treino. Em casos como esse, na aplicação do algoritmo em dados de teste, ou seja, dados diferentes do treino, a acurácia do algoritmo tende a diminuir significativamente. Portanto, as repetições auxiliam em um treinamento eficiente do algoritmo.

Por meio do teste foi possível identificar o algoritmo com melhor desempenho para a base de dados selecionada. A Figura 5 mostra o fluxograma com as etapas para a realização do estudo.

Figura 5. Etapas para verificação do melhor algoritmo de classificação.

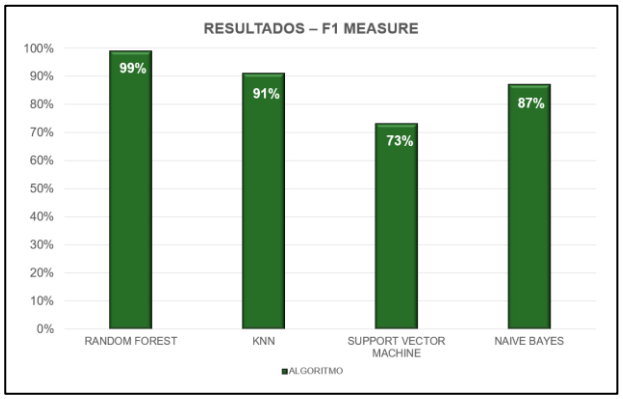
Fonte: Os autores (2020).

3. RESULTADOS

Os algoritmos Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Naive Bayes foram treinados com a base de dados do Kaggle.

O Gráfico 1 mostra os resultados obtidos por cada algoritmo na classificação de e-mails como spam na métrica F1.

Gráfico 1. Valor de F1 obtido por cada algoritmo de Machine Learning.



Fonte: Os autores..

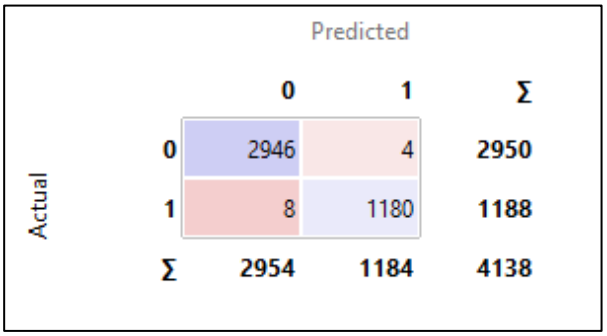
Os quatro algoritmos utilizados neste estudo apresentaram resultados com uma diferença significativa:

Random Forest → 99%
K-Nearest Neighbors → 91%
Support Vector Machine → 73%
Naive Bayes → 87%

Apesar de todos os algoritmos apresentarem acurácia alta, o algoritmo Random Forest obteve o melhor desempenho na classificação de e-mails como spam. A Figura 6 mostra a matriz de confusão correspondente a classificação do algoritmo Random Forest com a quantidade de dados. A Figura 7 mostra a matriz de confusão com a porcentagem (%) de cada classe obtida pelo algoritmo.

Segundo Freitas *et al.* (2007) a matriz de desempenho, conhecida como Matriz de Confusão faz uma análise consistente do comportamento do classificador. Essa matriz fornece uma representação de desempenho quantitativa para cada classificador em termos de reconhecimento de classe.

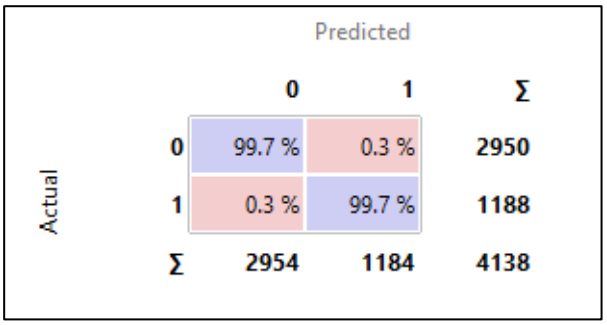
Figura 6. Matriz de Confusão do algoritmo RF gerada no software Orange com a quantidade de dados classificados para cada classe.



Fonte: Os autores..

Há duas classes consideradas na classificação dos dados: classe 0 e classe 1. Neste caso, a classe 1 corresponde aos e-mails classificados como spam; e a classe 0 corresponde aos demais e-mails (e-mail comum). Pela matriz de confusão do algoritmo Random Forest é possível observar que o algoritmo obteve apenas quatro erros em relação aos dados originais para a classe 0, ou seja, o algoritmo foi capaz de classificar corretamente (verdadeiro negativo) 2946 amostras como e-mail comum em um total de 2950 presentes na base. Em relação a classe 1, o algoritmo foi capaz de classificar corretamente (verdadeiro positivo) 1180 amostras da base original, desta forma, somente 8 dados que correspondem a spam foram classificados como e-mail comum.

Figura 7. Matriz de Confusão do algoritmo RF gerada no software Orange com a porcentagem dos dados classificados para cada classe.



Fonte: Os autores .

4. DISCUSSÃO

Com base nos estudos realizados nos últimos anos verifica-se que a área de aprendizado de máquina vem ganhando grande destaque em segmentos variados para fins também diversos. O auxílio proporcionado pelos algoritmos alcança resultados mais precisos do que técnicas mais tradicionais. É possível perceber que grande parte dos estudos científicos envolvendo a área do aprendizado de máquina são trabalhos recentes, o que enfatiza mais ainda a crescente adoção de métodos envolvendo inteligência artificial.

Nesse contexto, o algoritmo Random Forest, em específico, tem alcançado excelentes resultados, visto que estudos cujo foco seja a comparação de desempenho entre diferentes algoritmos tem apontado que o Random Forest apresenta a melhor ou uma das melhores métricas. Os algoritmos baseados em árvore de decisão têm mostrado altas acurácias em aplicações de diversas áreas.

O trabalho de Hu *et al.* (2019), por exemplo, consiste em analisar uma base de dados do Kaggle para classificação de malware. Para estudar o desempenho de diferentes algoritmos, a pesquisa utilizou dez algoritmos incluindo Random Forest (RF), Support Vector Machine (SVM) e Naive Bayes (NB). Assim como o presente estudo, os autores também constataram que o algoritmo Random Forest obteve melhores resultados por apresentar as melhores métricas de desempenho. Apesar do Random Forest apresentar resultados elevados, os autores citam o fato de nenhum algoritmo atingir 100% em todas as métricas de desempenho.

Já Bassiouni e El-Dahshan (2018), também avaliaram o desempenho de dez algoritmos, porém utilizaram o software WEKA. O tema estudado também se refere à identificação de e-mails como spam. Dentre os algoritmos analisados estão os quatro utilizados no presente estudo. Após o processamento dos dados o trabalho concluiu que a maior acurácia, de 95,45%, também foi alcançada pelo algoritmo Random Forest.

Subasi *et al.* (2018), realizaram uma avaliação sobre o melhor método de filtragem de e-mail de spam com base em algoritmos de árvore de decisão. O resultado aponta que o Random Forest juntamente com o Rotation Forest foram os algoritmos que alcançaram melhor resultado. A conclusão desse trabalho evidencia a relevância do Random Forest mostrando que o algoritmo não somente se destaca em comparação a algoritmos com outra estrutura, mas também em relação aos baseados em árvore de decisão.

5. CONSIDERAÇÕES FINAIS

Verificou-se que a utilização de técnicas de ML é útil na classificação de e-mails como spam. Para a classificação de uma base de dados que apresentam informações sobre a presença de spam nos e-mails com a aplicação de diferentes algoritmos de machine learning no software Orange, o algoritmo RF apresenta o melhor desempenho em relação aos demais algoritmos com uma acurácia de 99%. Apesar do algoritmo RF obter a maior acurácia, os outros algoritmos, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e Naive Bayes (NB), apresentaram uma boa acurácia (91%, 73% e 87% respectivamente). Dessa forma, os quatro algoritmos testados neste estudo (RF, KNN, SVM e NB) podem ser aplicados em casos similares.

REFERÊNCIAS

BASSIOUNI, M.; ALI, M.; EL-DAHSHAN, E. A. (2018). Ham and Spam E-Mails Classification Using Machine Learning Techniques. **Journal of Applied Security Research,** v. 13, n. 3, p. 315–331. 2018 https://doi.org/10.1080/19361610.2018.1463136

DADA, E. G.; BASSI, J. S.; CHIROMA, H.; ABDULHAMID, S. M.; ADETUNMBI, A. O.;AJIBUWA, O. E. Machine learning for email spam filtering: review, approaches and

open research problems. **Heliyon,** v. 5, n. 6, p. e01802. 2019. https://doi.org/10.1016/j.heliyon.2019.e01802

FARIS, H.; AL-ZOUBI, A. M.; HEIDARI, A. A.; ALJARAH, I.; MAFARJA, M.; HASSONAH, M. A.; FUJITA, H. An Intelligent System for Spam Detection and Identification of the most Relevant Features based on Evolutionary Random Weight Networks. **Information Fusion**, v. 48, p. 67-83, Aug. 2018.https://doi.org/10.1016/j.inffus.2018.08.002

FREITAS, C.O.A.; DE CARVALHO, J. M.; OLIVEIRA, J.; AIRES, S.B.K.; SABOURIN, R. (2007) Confusion Matrix Disagreement for Multiple Classifiers. *In*: RUEDA, L.; MERY, D.; KITTLER, J. (eds) **Progress in Pattern Recognition, Image Analysis and Applications.** CIARP 2007. Lecture Notes in Computer Science, v. 4756. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-76725-1_41.

GLOBAL SOFTWARE SUPPORT. **Random Forest Classifier – Machine Learning**, 2018. Disponível em: https://www.globalsoftwaresupport.com/random-forest-classifier-bagging-machine-learning/. Acesso em: 22 jun. 2020.

HAN, J. D.; KAMBER, M. **Data Mining Concept and Tehniques.** San Fransisco: Morgan Kauffman, 2006.

HU, Y.-H. F.; ALI, A.; HSIEH, C.-C. G.; WILLIAMS, A. **Machine Learning Techniques for Classifying Malicious API Calls and N-Grams in Kaggle Data-set.** 2019 SoutheastCon. https://doi.org/10.1109/SoutheastCon42311.2019.9020353

JOSÉ. I. **KNN (K-Nearest Neighbors)**, 2018. Disponível em: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d. Acesso em: 22 jun. 2020.

KARTHICK, S. Semi Supervised Hierarchy Forest Clustering and KNN Based Metric Learning Technique for Machine Learning System. **Journal of Advanced Research in Dynamical and Control Systems**. v. 9. Sp– 18. 2017.

MITCHELL, T. M. **Machine Learning.** [*S.l.*]: McGraw-Hill, 1997.

NAIK, A.; SAMANT, L. Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. **Procedia Computer Science**, v. 85, p. 662–668. 2016. https://doi.org/10.1016/j.procs.2016.05.251

RAY., S. **Understanding Support Vector Machine algorithm from examples (along with code)**, 2017. Disponível em: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. Acesso em: 22 jun. 2020.

ROY, S. S.;VISWANATHAM, V. M. Classifying Spam Emails Using Artificial Intelligent Techniques. **International Journal of Engineering Research in Africa**, v. 22, p. 152–161, 2016.https://doi.org/10.4028/www.scientific.net/JERA.22.152

SUBASI, A.; ALZAHRANI, S.; ALJUHANI, A.; ALJEDANI, M. Comparison of Decision Tree Algorithms for Spam E-mail Filtering. *In:* INTERNATIONAL CONFERENCE ON COMPUTER APPLICATIONS & INFORMATION SECURITY (ICCAIS). 1., 2018. https://doi.org/10.1109/CAIS.2018.8442016

VANDERPLAS, J. **Python Data Science Handbook.** O'Reilly and Associates, 2016.