# Otimização Automática de Classificadores para Auxiliar no Diagnóstico da Depressão

**Renata Cristina Santana**[1], **Thiago Henriques N. de Lima**[1], **Saulo A. P. Pinto**[1], **Luis E. Zárate**[1], **Cristiane N. Nobre**[1]

[1] Instituto de Ciências Exatas e Informática
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brasil

{renata.cris.santana, thiagohenriqueslima}@gmail.com,

{saulo, zarate, nobre}@pucminas.br

***Abstract.*** *Depression is a disease that affects millions of people around the world. It is complexly related to biological, social, and psychological factors. This paper presents preliminary results of using the Auto-WEKA package to automatically find classifiers that perform well to predict whether a person has developed depression or not. Classifiers combined with automatic reduction of attributes such as Random Forest, Best Fit and Correlation-based Subset Selection achieved 100% hit, sensitivity, and specificity rates, indicating that they are a promising alternative to assist medical professionals in the diagnosis of depression.*

***Resumo.*** *A depressão é uma doença que atinge milhões de pessoas em todo o mundo. Ela é relacionada de modo complexo a fatores biológicos, sociais e psicológicos. Este trabalho apresenta resultados preliminares da utilização do pacote Auto-WEKA para encontrar automaticamente classificadores que tenham bom desempenho para predizer se uma pessoa desenvolveu depressão ou não. Classificadores combinados com redução automática de atributos como Random Forest, Best Fit e Correlation-based Subset Selection, alcançaram taxas de acerto, sensibilidade e especificidade de 100%, o que indica que são uma alternativa promissora para auxiliar profissionais da área médica no processo de diagnóstico da depressão.*

## 1. Introdução

Segundo a Organização Mundial da Saúde (OMS), a depressão é um transtorno mental comum. Mais de trezentos milhões de pessoas de todas as idades sofrem com essa doença, sendo que as mulheres são mais susceptíveis que os homens (Silva et al., 2014). A depressão resulta de uma complexa interação de fatores sociais, psicológicos e biológicos. As pessoas que passaram por eventos adversos na vida, como o desemprego, falecimento de entes queridos ou trauma psicológico são mais propensas a desenvolver depressão, além das inter-relações entre depressão e saúde física.

Nos últimos anos, a área de Aprendizagem de Máquina (*Machine Learning*) tem alcançado resultados impressionantes em problemas difíceis como reconhecimento de fala, reconhecimento de objetos em imagens e vídeos e processamento de linguagem natural (Ravì et al., 2017), bem como em problemas da Bioinformática, como o diagnóstico

e prognóstico de doenças (Ravì et al., 2017). De especial interesse, é a predição da ocorrência da depressão, doença que acomete grande número de pessoas[1].

Dada a extensão dos problemas causados pela depressão, técnicas que permitam predizer, com alta confiabilidade, se determinada pessoa pode desenvolver a depressão, tornam-se importantes para auxiliar no diagnóstico desta doença. Uma abordagem promissora está na utilização de técnicas de aprendizagem de máquina. Em alguns problemas elas melhoraram as taxas de acerto para níveis sobre-humanos. É o caso do reconhecimento de objetos em imagens, cuja taxa de acerto é de 99,77% (Ravì et al., 2017; Witten et al., 2016). Cada método de aprendizagem possui um conjunto próprio de *hiperparâmetros* (Kottoff et al., 2016; Witten et al., 2016) que devem ser sintonizados. Assim, além de selecionar o método de aprendizagem entre as dezenas disponíveis, deve-se selecionar, também, um conjunto adequado de parâmetros para cada método, tarefa extremamente árdua para iniciantes em Aprendizagem de Máquina, quiçá para leigos, principalmente no caso de conjuntos de dados grandes com muitos atributos (Kottoff et al., 2016). Logo, optou-se por usar o pacote Auto-WEKA (Kottoff et al., 2016) para WEKA (*Waikato Environment for Knowledge Analysis*) (Witten et al., 2016) que foi projetado para otimizar a combinação "método de aprendizagem+hiperparâmetros".

Dessa forma, o objetivo maior deste trabalho é automatizar o processo de escolha do melhor classificador com seus hiperparâmetros a fim de auxiliar no diagnóstico da depressão a partir de dados previamente coletados que informam se uma pessoa desenvolveu ou não a depressão nos últimos doze meses, conforme definido na Subseção 3.1. Aqui, são apresentados resultados da seleção de métodos de classificação e de seleção de características (atributos) feita pelo Auto-WEKA.

A Seção 2 descreve brevemente conceitos utilizados no trabalho. A Seção 3 apresenta uma descrição dos dados e do processamento executado sobre os mesmos. Já a Seção 4 apresenta os principais resultados e uma discussão sobre os mesmos. Conclusões e direções futuras são detalhados na Seção 5.

## 2. Conceitos

### 2.1. Auto-WEKA

O pacote Auto-WEKA (Kottoff et al., 2016) para WEKA (Witten et al., 2016), versão 2.6, está disponível para ser instalado no Gerenciador de Pacotes (*Package Manager*) da Versão 3.8 do WEKA. Após instalado, ele pode ser utilizado diretamente pela aba homônima ou como um dos classificadores na pasta "*Functions*". O Auto-WEKA faz uma otimização bayesiana da combinação método+hiperparâmetros, chamada de uma *configuração*. Ele gera várias destas configurações que são avaliadas no conjunto de dados utilizando-se a validação cruzada de 10 dobras (Kohavi, 1995). O Auto-WEKA disponibiliza métodos de seleção e avaliação de atributos, classificadores e regressores. No total, são quarenta métodos suportados e que podem ser combinados uns com os outros, considerando valores de seus parâmetros para formar as configurações a serem avaliadas (Kottoff et al., 2016).

[1] *Depression*. *World Health Organization*. http://www.who.int/mediacentre/factsheets/fs369/en/

### 2.2. Métricas de avaliação do modelo

A sensibilidade é a taxa de verdadeiros positivos e indica o número de positivos que foram corretamente classificados como tal, isto é, a taxa de instâncias da classe "Sim" que foram classificadas como "Sim" pelo classificador. Em termos médicos, indica a porcentagem de pessoas acometidas por certa "doença" que foram diagnosticadas corretamente e não foram deixadas sem diagnóstico correto. Já a especificidade indica que a condição a ser diagnosticada é realmente verdadeira e não é outra condição detectada erroneamente. Ou seja, alta especificidade indica que raramente outra condição é diagnosticada no lugar da condição-alvo (Altman e Bland, 1994).

No contexto da Aprendizagem de Máquina, a Estatística Kappa é a proporção na qual um classificador indica corretamente a classe considerando a diferença entre um classificador perfeito, que não erra, e um classificador randômico, que escolhe a classe "aleatoriamente". Assim, a estatística divide a diferença entre o número de acertos do classificador e o classificador randômico pela diferença entre o número de acertos do classificador perfeito e o classificador randômico (Witten et al., 2016). Logo, o maior valor de Kappa é 100% (classificador perfeito) e o menor é zero (classificador aleatório).

## 3. Métodos

### 3.1. Base de Dados

Os dados utilizados neste trabalho foram coletados e/ou "produzidos" pela Pesquisa de Saúde da Comunidade Canadense (*Canadian Community Health Survey* – CCHS) no ano de 2014, totalizando 1129 atributos com dados disponibilizados ao público, sob requisição. Estes atributos incluem dados de informações socio-demográficas, econômicas, comportamentais e de saúde da população amostrada (Canada, 2015). Entre os atributos, existem, também aqueles "produzidos" ou "derivados" de outros atributos por meio de cálculos ou agrupamento (Canada, 2016). É o caso do módulo de depressão.

As questões que compõem o módulo de depressão do CCHS são baseadas em um formulário longo da escala CIDI (*Composite International Diagnostic Interview*), que foi desenvolvida no final dos anos 1980 e início dos 1990 (WHO, 1993).

Neste trabalho, o atributo (ou variável) utilizado para gerar a saída esperada ("classe") para a tarefa de predição é o DPSDPP (*Depression Scale - Probability of Caseness to Respondents*). Os valores deste atributo representam probabilidades de a pessoa respondente do questionário ter sido diagnosticada como tendo depressão nos últimos doze meses. A probabilidade é calculada com base no trabalho de Kessler, Mroczek e colegas, conforme (Canada, 2016). São atribuídas probabilidades iguais a 0, 0,05, 0,25, 0,50, 0,80 e 0,90. Para este trabalho, visto que apresenta resultados preliminares, foram consideradas apenas instâncias cujo valor do atributo DPSDPP é 0 ou 0,9, representados pelas classes "Não" (para probabilidade de não ocorrência de depressão, ou seja, igual a 0) e "Sim" (para "alta" probabilidade de ocorrência de depressão, ou seja, igual a 0,9), respectivamente. Evidentemente, como os valores de DPSDPP são derivados e relacionados a outros três atributos (DPSDMT, o último mês que a pessoa se sentiu deprimida, DPSDSF, indicador do nível de depressão que avalia se a pessoa ficou triste ou deprimida por duas semanas ou mais no ano anterior, e DPSDWK, número de semanas que se sentiu deprimido), estes foram excluídos.

Assim, das 63522 instâncias do conjunto de dados do ano 2014, foram eliminadas 42508 cujos entrevistados não responderam às perguntas do módulo de depressão. As demais 21014 possuíam probabilidades atribuídas. O passo seguinte descartou as instâncias com probabilidades iguais a 0,05, 0,25, 0,50 e 0,80, resultando 18246 instâncias. Destas, como apenas 1104 tinham probabilidade igual a 0,9 (ou seja, pertencentes à classe "Sim") e as demais, probabilidade igual a 0 (ou seja, pertencentes à classe "Não"), foi feito o balanceamento do número de instâncias nas duas classes.

Para cada execução do Auto-WEKA, um conjunto de dados com 2208 instâncias foi utilizado. Todas as 1104 instâncias da Classe "Sim" foram incluídas em todos os conjuntos. As 1104 instâncias da Classe "Não" foram selecionadas randomicamente (sem repetição) considerando distribuição de probabilidade uniforme. Foram gerados trinta arquivos no formato ARFF (*Attribute Relation File Format*), um dos formatos aceitos pelo WEKA, com os trinta conjuntos de dados, a fim de aumentar a qualidade e confiabilidade dos resultados (Subseção 3.2).

### 3.2. Processamento

Alguns testes iniciais foram executados apenas com o objetivo de avaliar a influência dos parâmetros disponibilizados na interface gráfica do Auto-WEKA. Os dois parâmetros mais comumente alterados são o tempo de execução (*timeLimit*), em minutos, e a memória máxima disponível, em megabytes, (*memLimit*), cujos valores-padrão são 15 e 1024, respectivamente. Além disso, é recomendado que o tempo de execução seja de "algumas horas" para que milhares de configurações sejam avaliadas (Kottoff et al., 2016). Como o Auto-WEKA provê processamento paralelo com múltiplas linhas de execução (*threads*), o número de execuções em paralelo ("*parallelRuns*") foi ajustado para 4, o que aumentou quase quatro vezes o número de configurações avaliadas no mesmo tempo de execução. Entretanto, ao ser aumentado o tempo de execução para "várias horas" (2, 4, 6, 8 e 12 horas foram testados, inicialmente), a taxa de acertos diminuiu em relação àquela dos testes com quinze minutos. Como em nenhum dos testes de quinze minutos o Auto-WEKA reportou uso de seleção de atributos, foi necessário ampliar o escopo dos testes.

Dessa forma, a fim de ratificar que o desempenho com 15 minutos seria o melhor e ampliar a possibilidade de o Auto-WEKA encontrar métodos de seleção de atributos, foi decidido que testes com trinta arquivos de dados e com 15, 30, 45, 60 e 480 minutos seriam executados. Em seguida, novos testes, com conjuntos de dados formados apenas com os dados dos atributos selecionados seriam executados.

## 4. Resultados e Discussões

Como o produto final do trabalho deve ser usado por profissionais da área da saúde, todos os testes executados foram hospedados em uma máquina com uma configuração de não muito alto custo, um *notebook* DELL com processador Intel Core i5, 2.2GHz, com 8GB de RAM, rodando Windows 10 sob carga interativa. A Tabela 1 sumariza o número de configurações para cada tempo de execução testado. No total, foram 315 horas de testes e 22248 configurações avaliadas.

**Tabela 1. Número de configurações testadas pelo Auto-WEKA para encontrar o melhor classificador.**

| Tempo de execução (min) | 15 | 30 | 45 | 60 | 480 |
|---|---|---|---|---|---|
| **Média** | 98 | 116 | 136 | 148 | 243 |
| **Mínimo** | 87 | 86 | 86 | 91 | 180 |
| **Máximo** | 115 | 144 | 166 | 188 | 310 |
| **Total** | 2940 | 3480 | 4081 | 4453 | 7294 |

A Tabela 2 apresenta os valores médios, mínimos e máximos das execuções realizadas para os trinta arquivos de dados. As taxas de acerto são extremamente altas, principalmente para o tempo de 15 minutos, com média de 98,7% de acertos e com pouca variação: o desvio-padrão é de apenas 0,26%. Ainda que as taxas de acerto somente sejam questionáveis, esse comportamento se mantém para a estatística Kappa, sensibilidade, e especificidade (Seção 2). A especificidade de cada classificador é maior que a sensibilidade, o que indica que os classificadores deixam de classificar outra desordem em vez da depressão mais do que deixam de diagnosticar depressão (Tabela 2), ainda que em taxas muito baixas. Todos os classificadores têm alto valor de Kappa, indicando desempenho longe do esperado para "o acaso" e muito mais próximos ao do classificador perfeito.

Ao contrário do recomendado por Kottoff et al. (2016), os melhores resultados foram alcançados com o menor tempo de processamento (15 minutos). Em geral, o desempenho nas quatro métricas decresce com o aumento do tempo, o que pode indicar uma super adaptação de modelos aos dados, o que deve ser investigado. Entretanto há exceções a este comportamento. Uma delas foi utilizada para a seleção de atributos: os valores para as quatro métricas avaliadas mantiveram-se constantes, e maiores que 95%, independente do tempo.

**Tabela 2. Taxa de acerto, estatística Kappa, sensibilidade e especificidade para o melhor classificador encontrado pelo Auto-WEKA.**

| **Taxa de Acerto** | | | | | | **Estatística Kappa** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Tempo | 15 | 30 | 45 | 60 | 480 | 15 | 30 | 45 | 60 | 480 |
| Média | 98,71 | 96,37 | 92,25 | 90,94 | 87,29 | 97,41 | 92,72 | 84,50 | 81,89 | 74,57 |
| Mínimo | 97,64 | 81,52 | 81,25 | 81,25 | 81,52 | 95,29 | 63,04 | 62,50 | 62,50 | 63,04 |
| Máximo | 99,05 | 99,05 | 99,05 | 99,05 | 98,60 | 98,10 | 97,92 | 98,10 | 98,10 | 97,19 |
| **Sensibilidade** | | | | | | **Especificidade** | | | | |
| Tempo | 15 | 30 | 45 | 60 | 480 | 15 | 30 | 45 | 60 | 480 |
| Média | 97,94 | 95,31 | 90,75 | 89,31 | 84,60 | 99,47 | 97,42 | 93,75 | 92,58 | 89,97 |
| Mínimo | 96,65 | 77,63 | 77,63 | 77,63 | 77,63 | 98,64 | 85,42 | 83,79 | 83,79 | 85,42 |
| Máximo | 98,73 | 98,73 | 98,46 | 98,46 | 97,74 | 99,82 | 99,82 | 99,82 | 99,82 | 99,46 |

Em todos os testes de quinze minutos, o melhor classificador foi uma *RandomForest* (Witten et al., 2016), sem nenhuma aplicação de métodos de seleção de atributos. Outro classificador que retornou bons resultados foi o *Support Vector Machine* (SVM) com 91,80% de acertos. Em todos os testes em que métodos de seleção foram utilizados, o desempenho diminuiu, com exceção daquele do arquivo 24: método de busca *BestFit* com seleção de subconjuntos correlacionados ao atributo de saída *CfsSubsetEval*) (Witten et al., 2016). Logo, estes métodos foram escolhidos para a seleção de atributos.

Trinta atributos foram selecionados e utilizados até o tempo de redação deste artigo (arquivos 5 e 24). Eles estão relacionados, conforme esperado, a fatores biológicos (19 atributos), sociais (7 atributos) e psicológicos (4 atributos) [2]. O Auto-WEKA, alcançou taxas de acerto de 98,60% e 100%, para os arquivos 5 e 24, respectivamente, igualando e superando os resultados já apresentados.

[2] *Depression. World Health Organization.* http://www.who.int/mediacentre/factsheets/fs369/en/

## 5. Conclusões e Trabalhos Futuros

Os resultados apresentados, ainda que preliminares, são promissores. Em problemas complexos de classificação com milhares de atributos e de instâncias no conjunto de dados, resultados da ordem de 99% a 100% de acerto, sensibilidade e especificidade são difíceis de se alcançar com métodos tradicionais de Aprendizagem de Máquina, sendo possíveis apenas com técnicas de aprendizagem profunda (*deep learning*) (Ravì et al., 2017; Witten et al., 2016). O Auto-WEKA apresenta-se como uma alternativa para explorar, de forma mais efetiva, o poder dos métodos de classificação, ao otimizar a escolha dos métodos com seus conjuntos de parâmetros. Métodos tradicionais como as *RandomForest* associados aos de seleção de atributos apresentaram resultados notáveis em conjunto com o Auto-WEKA, o que indica que são uma alternativa promissora para auxiliar no processo de diagnóstico da depressão. Nos testes executados, foi observado que o Auto-WEKA pode melhorar os resultados com sucessivas execuções. Isso abre a possibilidade de otimizar o desempenho para novos conjuntos de dados e de atributos selecionados, ampliando ainda mais o uso dessa ferramenta em aplicações que exigem resultados precisos, como no diagnóstico de doenças como a depressão.

## Referências


D. G. Altman e J. M. Bland. Statistics notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308(6943):1552, 1994. ISSN 0959-8138. doi: 10.1136/bmj.308.6943.1552. URL `http://www.bmj.com/content/308/6943/1552`.

M. T. Silva, T. F. Galvao, S. S. Martins, e M. G. Pereira. Prevalence of depression morbidity among brazilian adults: a systematic review and meta-analysis. *Revista Brasileira de Psiquiatria*, 36(3):262–270, 2014. URL `https://www.ncbi.nlm.nih.gov/pubmed/25119639`.

L. Kottoff, C. Thornton, H. H. Hoos, F. Hutter, e K. Leyton-Brown. Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka. *Journal of Machine Learning*, 17:1–5, 2016. URL `https://www.cs.ubc.ca/labs/beta/Projects/autoweka/papers/16-599.pdf`.

I. H. Witten, E. Frank, M. A. Hall, e C. J. Pal. *Data Mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 4th edition, 2016.

D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, e G.-Z. Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2017.

S. Canada. *Canadian Community Health Survey (CCHS) - Annual Component - User guide 2014 and 2013-2014 Microdata files*, 2015.

S. Canada. *Canadian Community Health Survey (CCHS) - Annual Component (2014) - Public Use Microdata File - Derived Variable (DV) Specifications*, 2016.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8. URL `http://dl.acm.org/citation.cfm?id=1643031.1643047`.

WHO. Composite international diagnostic interview, version 1.1. In *Composite International Diagnostic Interview (CIDI), Version 1.1*. World Health Organization, 1993.