Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Thiago da Silva Alves

Inteligência Artificial no ciclo de crédito

Curitiba
2019

Thiago da Silva Alves

Inteligência Artificial no ciclo de crédito

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Luiz Eduardo Soares de Oliveira

Curitiba
2019

Inteligência Artificial no ciclo de crédito

Classificação de clientes inadimplentes utilizando métodos de *machine learning*

Thiago da Silva Alves[1]

[1]thiago.est@gmail.com

Resumo

Clientes inadimplentes representam um dos maiores riscos às instituições financeiras, dado seu potencial de provocar prejuízo. Somado a isso, está o esforço para minimizar esse risco, que consome recurso com o objetivo de repará-lo. Naturalmente, uma relação de custo e benefício onde o correto destino do recurso transforma o prejuízo em retorno financeiro. Nesse contexto, o objetivo deste trabalho é explorar o uso de algoritmos de inteligência artificial baseados em aprendizado supervisionado (*machine learning*) para classificar clientes inadimplentes. Serão descritas as etapas desde a construção da base da dados analítica (ABT), seleção de variáveis utilizando algoritmo genético (AG) até estratégia de aprendizado considerando custo do erro de classificação, reamostragem e medidas de avaliação ($Precision$, $Recall$ e $F_1 Score$) para conjutos de dados desbalanceados.

Palavras-chave: classificação supervisionada, algoritmo genético, reamostragem, métricas de avaliação, dados desbalanceados

Abstract

Overdue customers are one of the biggest threats to financial institutions, given their potencial to cause losses. Farther, is the effort to minimize this risk, wich expend resources to recover it. Clearly a cost benefit ratio, where correctly allocate resources turns losses into profits. In this context, the goal of this work is explore artificial intelligence algorithms based on supervised machine learning to classify overdue customers. Will describe steps from analytical base table (ABT) building, variable selection with genetic algorithm (GA) to learning strategy with misclassification cost, resample and evaluate metrics ($Precision$, $Recall$ and $F_1 Score$) for unbalanced data.

Keywords: supervised classification, genetic algorithm, resample, evaluate metrics, unbalanced data

1. Introdução

Todas as instituições financeiras que oferecem produtos de crédito aos seus clientes têm que tomar decisões que envolvem risco de perda financeira. Essas decisões se estendem desde a prospecção até a possível inadimplência do cliente, nesse contexto está o ciclo de crédito, constituído pelas seguintes etapas:

- Prospecção: decidir qual produto oferecer para cada cliente.
- Análise: decidir a liberação do crédito para o cliente.
- Gestão: monitorar indicadores dos produtos concedidos aos clientes.
- Cobrança: localizar, contactar e recuperar o crédito inadimplente.

Em cada uma dessas etapas uma grande quantidade de dados do cliente é adquirida para sustentar as decisões, essas tornam-se variáveis em modelos estatísticos de mensuração de risco que são amplamente utilizados por instituições financeiras no mundo todo.

A etapa de cobrança, tem desafios adicionais além de classificar os clientes em possíveis pagadores ou não pagadores. Localizar e contactar o cliente no momento certo decidindo qual o meio de abordagem é o mais adequado, são exemplos desses desafios que em geral são tratados em estratégias definidas com base em obter a máxima recuperação (financeira) gastando o mínimo necessário.

Diante desse cenário, surge a oportunidade de aplicação de métodos de aprendizado supervisionados (ver sessão 4) para classificar clientes inadimplentes

em pagadores e não pagadores, utilizando os dados adquiridos durante o relacionamento do cliente com a instituição financeira em todas as etapas do ciclo de crédito. Dessa oportunidade, os classificadores surgem como uma alternativa aos modelos de risco tradicionais e serão o objeto de estudo deste trabalho.

Especificamente o objetivo principal deste estudo é explorar o uso e avaliar o desempenho de classificadores baseados em aprendizagem de máquina. Para isso será necessário transformar uma grande quantidade de dados em características que explicam o comportamento do cliente (2), selecionar as que mais contribuem para o aprendizado (3), abordar o problema de desbalanceamento entre as classes (4.3) e por fim, escolher medidas de avaliação que permitam considerar o custo do erro de classificação na decisão final (4.5). Desta maneira, iremos mostrar a possibilidade de introduzir conceitos de inteligência artificial na tomada de decisões de risco em instituições financeiras.

2. Dados

Os dados utilizados foram cedidos por uma instituição financeira de grande porte no mercado brasileiro, sob condição de manutenção do sigilo de todas as informações sensíveis dos clientes e completa anonimização dos nomes das variáveis. Correspondem a uma amostra de 220 mil clientes inadimplentes há no máximo 30 dias (tempo em atraso) nos produtos conta corrente com limite e empréstimos pessoais. E, 672 variáveis extraídas de diversas fontes (sistemas) e momentos da jornada do cliente durante o ciclo de crédito.

2.1. Origem

Diariamente os sistemas responsáveis por suportar a operação dos produtos de crédito, identificam se o cliente está inadimplente e o enviam para os sitemas de cobrança, que análogamente pode ser comparado com um estoque, onde as entradas são as ocorrências de inadimplência e as saídas são os pagamentos.

2.2. Preparação

Consolidando todas as entradas (primeiro registro) que ocorreram em um período de tempo, forma-se uma safra de clientes inadimplentes. Esse conceito permite organizar as duas etapas fundamentais da criação de uma base da dados analítica (*analytical base table* - ABT) necessária para qualquer processo de modelagem / classificação de dados:

- Criação de características
- Rotulagem dos dados

A figura 1 ilustra o processo de criação da ABT, do histórico são coletadas todas as características dos clientes e no período subsequente a safra o dado é rotulado. Especificamente nesse estudo o histórico é de no máximo 12 meses, cada safra corresponde a um mês (calendário) e o rótulo é atribuído após 30 dias da entrada em inadimplência.

Figura 1: Processo de criação da ABT

Dessa forma definimos X como vetor de características do cliente, sendo:

$$X = x_1, x_2, ..., x_{672} \tag{1}$$

E como vetor de rótulos y, sendo:

$$y = \begin{cases} 0 \\ 1 \end{cases} \tag{2}$$

Onde $y_i = 0$ quando o cliente não pagou em até 30 dias e $y_i = 1$ caso contrário.

3. Seleção de variáveis

Construir uma estratégia para seleção de variáveis é de fundamental importância para o aprendizado dos algoritmos, dado que o vetor de características deve carregar a máxima informação sobre os indivíduos e o evento que está sendo estudado [1]. Especificamente neste estudo, devido ao número de variáveis (672), a necessidade de redução de dimensão torna-se fundamental para o sucesso do aprendizado considerando o custo de processamento envolvido.

Na literatura são descritas muitas formas de seleção de atributos, dentre as quais está a seleção de subconjuntos (*subset selection*) que pode ser implementada usando algoritmo genético [2].

3.1. Algoritmo Genético (AG)

Este método descrito pela primeira vez por John Holland (1960) é baseado nos conceitos de evolução natural de populações anteriormente descritos por Charles Darwin, onde uma determinada população é reproduzida ao longo de gerações e avaliada segundo um critério. A seleção dos indivíduos de cada geração baseada nesse critério garante a evolução natural da população. O processo pode ser interrompido quando o critério atinge um valor desejado (convergência) ou um número fixo de gerações é reproduzida. As etapas principais podem ser descritas como:

- Inicialização da população
- Avaliação
- Operações genéticas (reprodução, cruzamento e mutação)

3.1.1. Parametrização

Em cada uma das etapas é necessário definir parâmetros de controle do algoritmo. No contexto de seleção de variáveis podemos definir a população inicial como um conjunto de n vetores de características definido em (1) e a função de avaliação sendo a medida F_1 Score. Calculada a partir da matriz de confusão (tabela 2) resultado de um classificador supervisionado baseado em árvore de decisão (*Random Forest*)(ver sessão 4), estratégia amplamente utilizada na avaliação de aprendizado em problemas de classificação.
Resumidamente os parâmetros são:

- Tamanho da populacão: n = 20
- Função de avaliação: $F_1 Score$
- Taxa de cruzamento = 0,8
- Taxa de mutação = 0,01
- Número de gerações = 100 (condição de término)

3.2. Execução

Levando em conta o custo de processamento envolvido, a execução do AG foi realizada sobre uma amostra aleatória de 20.000 casos divididos da seguinte forma:

- Treino = 50%
- Validação = 20%
- Teste = 30%

3.2.1. Inicialização da População

Considerando um vetor de mesma dimensão do definido em (1), cujos elementos são variáveis aleatórias

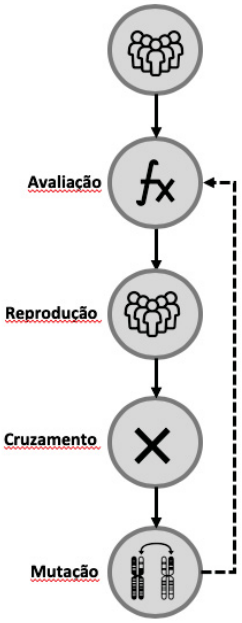

Figura 2: Etapas do Algoritmo Genético

indicadoras (0 e 1) e representam um subconjunto de características, temos a seguinte população de tamanho $n = 20$ (3.1.1) ilustrada na figura 3:

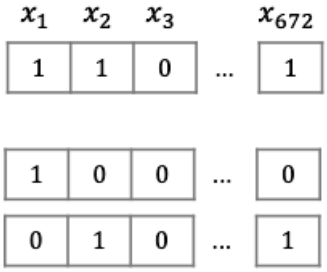

Figura 3: População inicial (AG)

3.2.2. Avaliação

Nesta etapa cada elemento da população deve ter sua função de avaliação (*fitness*) calculada, ou seja, para cada subconjunto de características um classificador (*Random Forest*) é aplicado na amostra de treino e o valor da *fitness* $F_1 Score$ (ver sessão 4) é obtido da amostra de validação.

3.2.3. Reprodução

Após o cálculo da *fitness* acontece a reprodução, onde uma nova população de mesmo tamanho ($n = 20$) deve ser gerada. A seleção dos novos indivíduos pode acontecer de duas maneiras [3]:

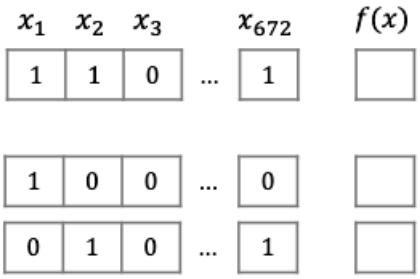

Figura 4: Função de avaliação (AG)

- Elitista: os melhores indivíduos são selecionados para a próxima geração.
- Não - Elitista: os inidivíduos são selecionados seguindo um critério, nesse caso existe a possibilidade do melhor indivíduo não pertencer a próxima geração.

O critério utilizado na seleção não - elitista também pode variar com a finalidade de manter a diversidade da população. Neste trabalho o critério escolhido foi o torneio. Uma seleção aleatória de 3 indivíduos é realizada e o que possui o maior valor de *fitness* é selecionado, repete-se até que $n = 20$ indivíduos sejam reproduzidos.

3.2.4. Cruzamento

Pares de indivíduos são escolhidos para trocar informações (alteração no subconjunto de variáveis), neste ponto uma característica (variável) que não foi utilizada pelo classificador (3.2.2) poderá ser incluída no novo indivíduo que será gerado do resultado deste cruzamento ou vice e versa. A troca acontece com probalidade de 0,8 (3.1.1) para cada indivíduo, ou seja, alguns serão apenas copiados para a próxima geração sem sofrer alteração.

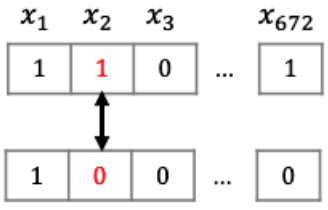

Figura 5: Cruzamento (AG)

A troca de informações pode acontecer para cada ponto (variável) ou de outras formas, neste caso a escolha foi a de um ponto.

3.2.5. Mutação

Os indivíduos resultantes do cruzamento podem sofrer mutação, com probabilidade de 0,01 (3.1.1), quando selecionados para esta etapa. Cada elemento do subconjunto características poderá ter seu valor alterado com probabilidade 0,05.

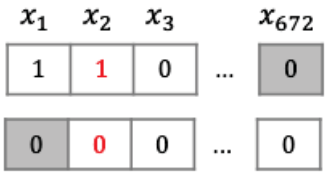

Figura 6: Mutação (AG)

Ao final desta etapa uma nova população foi gerada, temos novos indivíduos que representam diferentes subconjuntos de variáveis candidatas a seleção do classificador. O AG retorna à etapa de avaliação, calcula a *fitness* para cada indivíduo da nova população e executa as etapas seguintes até que 100 gerações (3.1.1) sejam reproduzidas.

3.2.6. Melhor indivíduo

Completada a evolução da população, selecionamos o indivíduo com o maior valor de *fitness* e este representa o melhor subconjunto de características para classificar o evento de interesse. Este indivíduo é fornecido ao classificador e aplicado na amostra de teste onde se obtém um novo valor da função de avaliação que pode ser comparado com os valores obtidos durante a evolução das gerações. As figuras 7 e 8 mostram os resultados.

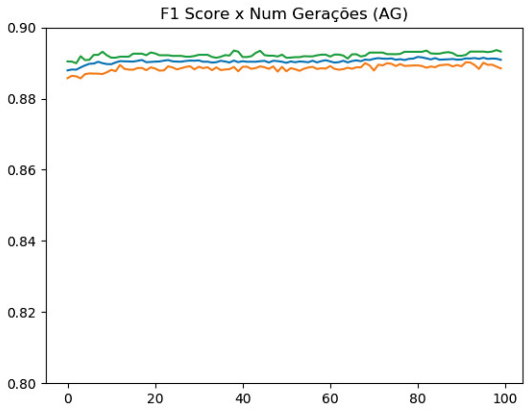

Figura 7: Evolução da *fitness* (AG)

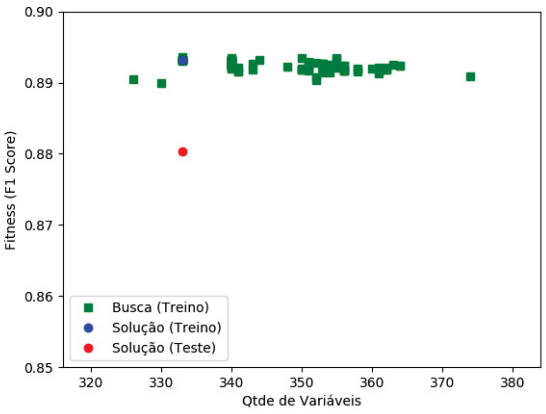

Figura 8: Evolução do número de variáveis selecionadas (AG)

A dimensão do vetor de características (1) foi reduzida de 672 para 333 e a escolha do classificador (*Random Forest*) proporciona a ordenação por importância dessas características (*feature importance*), dessa forma as 10 melhores, seguindo este critério, foram selecionadas. A escolha das variáveis além de ter impacto no sucesso da classificação também afeta o desempenho computacional dos algoritmos (custo de processamento), sendo assim a seleção de uma quantidade menor de atributos é fundamental.

4. Estratégia de aprendizado

Obter bons resultados de métodos de aprendizado passa pela construção de uma estratégia, algumas etapas dessa construção podem ser [5]:

- Pré-processamento
- Validação cruzada
- Reamostragem
- Escolha dos métodos (classificadores)
- Definição das métricas de avaliação

Não há obrigatoriedade em cumprir todas elas (ex: pré-processamento e reamostragem), cabe estudar o problema e definir a melhor estratégia. A seguir serão detalhadas as etapas da estratégia adotada para este trabalho.

4.1. Pré - processamento

Uma boa prática amplamente utilizada em problemas de clasificação é padronizar a escala das variáveis antes de fornecê-las aos métodos, isso contribui para a performance e resultados mais robustos podem ser encontrados [6]. A escala das variáveis selecionadas (3.2.6) foram padronizadas pela medida $Z - score$.

$$z = \frac{(x - \mu)}{\sigma} \tag{3}$$

Onde, μ é a média e σ o desvio - padrão do domínio de valores assumidos pela variável (x) na amostra de treino (4.2).

4.2. Validação cruzada

Técnica amplamente utilizada em problemas de aprendizado de máquina, consiste em particionar os dados de modo que o método seja treinado usando uma parte e avaliado no restante [7]. Isso permite observar o desempenho do método com dados novos do ponto de vista de qualidade das predições. Neste trabalho a partição dos dados (2) foi realizada da seguinte forma:

- Treino: 70%
- Teste: 30%

4.3. Reamostragem

Quando há desigualdade na distribuição de frequência do vetor de rótulos (2), os métodos tendem a classificar com melhor desempenho a classe em maior número, e consequentemente incorporar erro na distinção da classe minoritária [8]. Diante disso, optou-se pela reamostragem da classe $y_i = 0$ (Não pagou) na amostra de treino, proporcionando um equilíbrio entre as classes como mostra a tabela 1.

Rótulo	%	% (R)
$y_i = 0$ (Não pagou)	20%	50%
$y_i = 1$ (Pagou)	80%	50%

Tabela 1: Comparativo da distribuição do vetor de rótulos (pós reamostragem)

4.4. Classificadores

A escolha do método a ser utilizado tende a ser exaustiva, testar uma grande quantidade de classificadores, e decidir pelo melhor - segundo um critério de performance (4.5) - é a estratégia geralmente utilizada pela comunidade de *machine learning*. Especificamente neste trabalho a escolha foi limitada a três métodos de classificação supervisionados [4] cujos algoritmos estão implementados na biblioteca *scikit-learn* do software *Python*:

- Random Forest [9]
- kNN [10]
- Gradient Boosting [11]

4.5. Métricas de avaliação

Avaliar o desempenho dos métodos de classificação exige a escolha de uma ou mais medidas que possibilitem a correta leitura dos resultados e consequente escolha do melhor método. Em geral a medida amplamente utilizada é a acurácia, que apenas conta o número de acertos (classificações corretas) na amostra de teste. Porém, para dados desbalanceados (4.3), essa medida pode não fornecer o real desempenho do classificador [8].
Essa contagem pode ser resumida em uma tabela de contingência, denominada matriz de confusão (*confusion matrix*) [12].

	predição ($y_i = 0$)	predição ($y_i = 1$)
real ($y_i = 0$)	TN	FP
real ($y_i = 1$)	FN	TP

Tabela 2: Matriz de confusão

- TN: Taxa de verdadeiros negativos (Não Pagou)
- FP: Taxa de falsos positivos
- FN: Taxa de falsos negativos
- TP: Taxa de verdadeiros positivos (Pagou)

A partir da tabela 2 algumas medidas podem ser calculadas. Neste trabalho duas principais serão avaliadas com o objetivo de minimizar o custo do erro de classificação (FP), dado que a predição de um cliente pagador ($y_i = 1$) quando o verdadeiro rótulo é ($y_i = 0$) representa o maior risco para a instituição financeira (pagamento esperado não irá ocorrer). São elas:

$$Precision(y_i = 1) : P = \frac{TP}{TP + FP} \tag{4}$$

$$Recall(y_i = 0) : R = \frac{TN}{TN + FP} \tag{5}$$

Adicionalmente, a medida $F_1 Score$ combina as duas anteriores (4) (5) e fornece uma medida conjunta de desempenho.

$$F_1 = 2 * \frac{P * R}{P + R} \tag{6}$$

5. Resultados e discuções

Esse trabalho teve como objetivo explorar o uso de métodos de aprendizado de máquina (*machine learning*) na classificação de clientes inadimplentes. Porém, para chegar neste ponto a sessão 2 mostrou que a preparação dos dados é de fundamental importância e onde se despende grande parte do tempo total do trabalho. Transformar dados, presentes nos sistemas dedicados a suportar a operação de uma instituição financeira, em características comportamentais do cliente não é uma tarefa simples, ao mesmo tempo em que é primordial para os métodos de aprendizado.

A seleção das melhores características comportamentais (variáveis), descrita na sessão 3, embora limitada pela confidencialidade e anonimização, que não permitiu mostrar uma análise exploratória mais profunda dessas, obteve bons resultados na redução de dimensão com o uso do algoritmo genético. A obrigatoriedade da avaliação depender de uma métrica de desempenho ($F_1 Score$) resultado de um classificador na etapa anterior a de treinamento, mostrou resultados robustos nesta medida (tabelas 3 e 4) quando avaliada na amostra de teste.

O equilíbrio na distribuição do vetor de rótulos, descrito na sessão 4.3, teve grande impacto no treinamento dos classificadores [8], melhorando sensivelmente as taxas de acerto para a classe minoritária $y_i = 0$ (Não pagou) e consequente diminuição do erro de classificação de maior custo (FP). Conforme mostram as tabelas a seguir, contendo a matriz de confusão e as respectivas métricas (4.5) para cada classe resumindo o desempenho dos métodos na amostra de teste.

Sem Reamostragem						
		($y_i = 0$)	($y_i = 1$)	P	R	F_1
RF	($y_i = 0$)	**0,01**	**0,19**	0,41	**0,06**	0,10
	($y_i = 1$)	0,02	**0,78**	**0,80**	0,98	0,88
KNN	($y_i = 0$)	**0,04**	**0,17**	0,29	**0,17**	0,21
	($y_i = 1$)	0,08	**0,71**	**0,81**	0,89	0,85
GB	($y_i = 0$)	**0,01**	**0,19**	0,54	**0,06**	0,11
	($y_i = 1$)	0,01	**0,79**	**0,80**	0,99	0,79

Tabela 3: Resultados na amostra de teste sem reamostragem

Na matriz de confusão da tabela 3, nota-se claramente o efeito do desequilíbrio entre as classes, apesar de altas taxas de acerto para a classe $y_i = 1$ (Pagou) próximas a 0,80, reflete em FP perto de 0,20 e $Recall$

de no máximo 0,17 (KNN) para a classe $y_i = 0$ (Não pagou). Indicando que nenhum dos métodos produz classificações consistentes para a classe minoritária (Não pagou) e concentrando o erro em FP.

Com Reamostragem						
		$(y_i = 0)$	$(y_i = 1)$	P	R	F_1
	$(y_i = 0)$	**0,05**	**0,15**	0,36	**0,24**	0,29
	$(y_i = 1)$	0,09	**0,71**	**0,82**	0,89	0,85
	$(y_i = 0)$	**0,08**	**0,12**	0,25	**0,40**	0,31
	$(y_i = 1)$	0,24	**0,56**	**0,82**	0,70	0,75
	$(y_i = 0)$	**0,11**	**0,09**	0,33	**0,54**	0,41
	$(y_i = 1)$	0,23	**0,57**	**0,86**	0,71	0,78

Tabela 4: Resultados na amostra de teste com reamostragem

O comportamento das medidas se altera quando a amostra de treino é submetida ao processo de reamostragem, conforme mostra a tabela 4. As taxas de acerto para a classe $y_i = 0$ (Não pagou) sobem, chegando a 0,11 (GB) e com expressivo aumento de $Recall$ em todos os métodos. E ainda, não há perda de capacidade preditiva para a classe $y_i = 1$ (pagou), com $Precision$ acima de 0,80 em todos os classificadores.

Isso mostra que os métodos de aprendizagem de máquina podem ser considerados como alternativa viável na classificação de clientes inadimplentes, sobretudo quando há interesse em incluir o custo do erro de classificação na decisão final. Os modelos de mensuração de risco de inadimplência, citados na sessão 1, em geral fornecem probabilidades associadas a cada elemento do vetor de rótulos (2) considerando o mesmo custo para cada tipo de erro (FN e FP) presente na matriz de confusão.

Naturalmente cabe uma avaliação da robustez desses métodos considerando a dimensão temporal presente nos dados, fato que não foi objeto deste trabalho e poderá ser explorado futuramente. A divisão das amostras na validação cruzada pode ser realizada na unidade de tempo das safras definidas na sessão 2.2, possibilitando a avaliação da estabilidade das métricas no tempo, apresentando uma alternativa de variação na estratégia de aprendizado, e que torna salutar a discussão e avaliação de soluções para problemas de classificação utilizando aprendizado de máquina.

Agradecimentos

Agradeço à minha esposa Miriany, pelo apoio e compreensão durante o período do curso. Ao orientador Prof. Luiz Eduardo, pelas importantes contribuições e autonomia que me concedeu durante o trabalho. Aos meus colegas de turma e de trabalho Everson e Teo pelas discussões e troca de experiências. E por fim, aos professores coordenadores do curso Wagner e Walmes, por viabilizarem essa oportunidade de qualificação profissional.

Referências

[1] YANG, J. and HONAVAR, V, *Feature Subset Selection Using a Genetic Algorithm*. Computer Science Technical Reports, 1997. Disponível em: <https://lib.dr.iastate.edu/cs_techreports/156/> Acesso em: 08 jun.2019.

[2] SOUFAN, O, KLEFTOGIANNIS D, KALNIS P and BAJIC VB. *DWFS: a wrapper feature selection tool based on a parallel genetic algorithm*. PLoS One. 2015. Disponível em: <10.1371/journal.pone.0117988> Acesso em: 08 jun.2019.

[3] BALUJA S and CARUANA, R. *Removing the genetics from the standard genetic algorithm*. ICML 1995. Disponível em: <https://www.ri.cmu.edu/pub_files/pub2/baluja_shumeet_1995_1/baluja_shumeet_1995_1.pdf> Acesso em: 09 jun.2019.

[4] KOTSIANTIS, S. B., *Supervised Machine Learning: A review of classification techniques*. Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press, 2007. ISBN 978-1-58603-780-2.

[5] CHICCO, D., *Ten quick tips for machine learning in computational biology*. BioData Min., Dec/2017. Disponível em: <https://dx.doi.org/10.1186%2Fs13040-017-0155-3> Acesso em: 16 jun.2019.

[6] NIZRI, M. N., ATOMI, H. A. and REHMAN, M. Z., *The Effect of Data Pre-Processing on Optimized Training of Artifial Neural Networks*. The 4th Intl. Conference of Electrical Engineering and Informatics, Procedia Technology, p.p 32-39, Nov 2013. Disponível em: <https://doi.org/10.1016/j.protcy.2013.12.159> Acesso em: 16 jun.2019.

[7] BROWNE, M. W., *Cross-Validation Methods*. Journal of Mathematical Psycology, vol.44, p.p 108-132, Mar/2000. Disponível em: <https://doi.org/10.1006/jmps.1999.1279> Acesso em: 16 jun.2019.

[8] CASTRO, C. L. e BRAGA, A. P., *Aprendizado Supervisionado com conjutos de dados desbalanceados*. Sba Controle e Automação, Campinas, vol.22, n.5, Set/Out 2011. Disponível em: <http://ref.scielo.org/n86bkc> Acesso em: 25 mai.2019.

[9] HO, T. K., *Random Decision Forests*. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp. 278–282, Aug/1995. Disponível em: <http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf> Acesso em: 16 jun.2019.

[10] ALTMAN, N. S., *An introduction to kernel and nearest-neighbor nonparametric regression*. The American Statistician, Jun/1991. Disponível em: <https://hdl.handle.net/1813/31637> Acesso em: 16 jun.2019.

[11] FRIEDMAN, J. H., *Greedy Function Approximation: A Gradient Boosting Machine*. Reitz Lecture, Feb/1999. Disponível em: <https://statweb.stanford.edu/ jhf/ftp/trebst.pdf> Acesso em: 16 jun.2019.

[12] FAWCETT, T., *An introduction to ROC analysis*. Patter Recognition Letters, Dec/2005. Disponível em: <https://www.math.ucdavis.edu/ saito/data/roc/fawcett-roc.pdf> Acesso em: 16 jun.2019