RMP
Revista dos Mestrados Profissionais
UFPE / CCSA – MGP
https://periodicos.ufpe.br/revistas/rmp
ISSN – 2317 - 0115
V. 11 - n. 2 (2022)
Recife – PE

Tomada de decisões baseada no parâmetro *b* da Teoria da Resposta ao Item nas classificações de *ensemble*

Paulo Fernando Leite-Filho – Universidade Federal de Pernambuco
https://www.orcid.org/0000-0003-1345-8259 - pflf@cin.ufpe.br
Silvio Barros Melo – Universidade Federal de Pernambuco
https://www.orcid.org/0000-0001-8600-6427 - sbm@cin.ufpe.br

Resumo – A tomada de decisões baseada no auxílio de sistemas informatizados é uma tarefa complexa e alvo de muitos ajustes para evitar erros de classificações. Avaliar as classificações pelos acertos dos algoritmos e associar estes acertos com o parâmetro da dificuldade (*b)* da TRI em concordância com a estatística de Kappa proporciona uma maior segurança ao usuário de sistemas informatizados. Nesta avaliação entre os grupos de *undersample* e *oversample* foi possível selecionar grupos específicos de algoritmos de cluster baseados no intervalo de confiança de 5% e intensidade de concordância maiores que 0,9968 identificando concordância perfeita (0,48%) nos grupos de *SMOTE*.

Palavras-chave: Tomada de decisão, Teoria da Resposta ao Item, Classificação *Ensemble*

Decision making based on Item Response Theory b parameter in ensemble rankings

Abstract: Decision-making based on the aid of computerized systems is a complex task and the target of many adjustments to avoid classification errors. Evaluating the classifications by the hits of the algorithms and associating these hits with the Difficulty parameter (b) of the IRT in agreement with the Kappa statistics provides greater security to the user of computerized systems. In this evaluation between the undersample and oversample groups, it was possible to select specific groups of cluster algorithms based on the 5% confidence interval and agreement intensity greater than 0.9968, identifying Perfect agreement (0.48%) in the SMOTE groups.

Keywords: Decision-making, Item Response Theory, Ensemble, Classification

Data da Submissão: 13/08/2022 **Data de aceitação:** 11/12/2022

https://doi.org/10.51359/2317-0115.2022.256869

1. Introdução

Os usuários de sistemas de informação compartilham dados com grande velocidade, este compartilhamento ocorre em redes locais ou na internet. Um objetivo específico deve ser realizado por estes sistemas, independendo da localização dos dados. Esta tarefa muitas vezes torna-se complexa devido ao caráter decisório que os sistemas devam realizar. Os dados são obtidos para objetivos diferentes como agrupamento, classificações, previsão e outros. As classes que compõem estes dados muitas vezes estão desbalanceadas, o que torna complexos os objetivos, especificamente quando obtidos da *web* (WU; HU, 2012).

Os algoritmos estão mais eficientes em seus objetivos, obtendo bons resultados em diferentes tipos de bases de dados (HATHALIYA; TANWAR, 2020). Porém existe uma preocupação de como determinar o método mais eficiente de identificar quais algoritmos são ótimos em determinadas situações. A análise dos parâmetros do método da teoria de resposta ao item (TRI) tem sido considerada como uma solução para selecionar registros considerados difíceis de classificar. Possibilitando uma forma de avaliar complexidades dos dados a serem triados por algoritmos de *machine learning.*

O objetivo deste trabalho é identificar algoritmos ótimos em classificações com menor *mean squear error* – MSE em classes balanceadas por redução de registros e pela geração de dados artificiais nas respectivas classes, utilizando o parâmetro da TRI.

2. Trabalhos Relacionados

A inteligência artificial (IA) ganha importância em muitas áreas de aplicação e poderá em breve ser fator de grande impacto nas soluções de responsabilidades de tomada de decisão na gerencia destas áreas. Quando os principais fatores para a tomada de decisão de um determinado domínio são identificados, a tomada de decisão torna-se complexa para os seres humanos. Estes fatores são de difícil decisão, e a TRI tem grande contribuição para identificar estes pontos, auxiliando a IA. A TRI complementaria assim a tarefa de auxiliar sistemas baseados em IA nas tomadas de decisões (BURGGRÄF et al., 2020).

Para melhorar a eficiência de um algoritmo de m*achine learning* pode-se atribuir parâmetros associados aos classificadores que acerte corretamente instâncias difíceis de classificar. Para avaliar a dificuldade das amostras dos *datasets* e a capacidade dos classificadores simultaneamente deve-se observar se a atividade a ser desempenhada é supervisionada ou não. Dentre estes métodos o treinamento não-supervisionado exige uma maior atenção, pois a decisão está totalmente direcionada aos modelos obtidos sem a intervenção humana, e atribuir o poder de decisão a este método exige que o sistema seja o mais preciso possível. Os *datasets* desequilibrados inferem que a precisão não é um critério útil porque não reflete a precisão da classificação dos casos de classe minoritária. Entretanto para as classes majoritárias, classificadas corretamente, o valor da precisão é alto, mesmo que a classe minoritária seja classificada incorretamente (ZHENG; CHEON; KATZ, 2020).

O balanceamento de classes pode ocorrer pela redução de exemplos da classe majoritária, chamada de *undersample* (ZHOU et al., 2020), com perda de informação ou com a geração de dados artificiais na classe minoritária, chamada de *oversample*

(DOUZAS; RAUCH; BACAO, 2021), podendo alterar os centroides dos *clusters*. Além do desbalanceamento das classes, o fator da dimensionalidade dos dados também influencia nas classificações, impactados pela complexidade dos dados. Uma técnica muito comum na computação é ponderar a tomada de decisões por um conjunto de classificadores, esta técnica é chamada de *ensemble* que será também avaliada neste trabalho (MOHEDANO-MUNOZ et al., 2021).

Este trabalho abordará a correlação do parâmetro *b* da TRI com as classificações de ensemble de clusters em classes com *undersample* e *oversample*, assim como as alterações na métrica de *mean squear error (MSE)*

2.1 Mean squear error

As métricas de avaliação de classificações têm como objetivo identificar o comportamento de cada um dos algoritmos em *datasets*. A métrica a ser utilizada neste trabalho é a *mean squear error.* Esta métrica calcula os erros de classificação entre o predito pelo algoritmo e o valor original da classe. Para isto, deve-se observar que a seleção de cada um dos resultados calculados segue uma finalidade específica. Ao obter este resultado estaremos avaliando individualmente os acertos de cada classificador (TANG et al., 2019).

Esta métrica é sensível nas classificações de cluster quando há desbalanceamento de classes. O número de registros presentes nas diferentes classes deve ser equalizado para corrigir este desbalanceamento, Há diversas técnicas para resolver este desbalanceamento como as técnicas de *random under sampling – rus; random over sampling - ros; smote; chan and stolfo's method - chan* e *easy - ensemble and asymmetric bagging* (TAHIR; KITTLER; YAN, 2012). Neste trabalho foi realizada uma adaptação das técnicas de *Undersample* e *Oversample* com a dificuldade (*b*) da TRI para a obtenção de uma amostra ideal e balanceada para avaliar os *MSEs* dos algoritmos.

2.2 Undersample

É uma estratégia que preferencialmente realiza a *subamostragem* do conjunto de dados da classe majoritária como pode ser observada na figura 1.

Esta técnica utiliza os dados disponíveis no *dataset* e produz o mesmo número de amostras de dados que a classe minoritária, balanceando as classes (LIN et al., 2017).

2.3 Oversample

É uma técnica que ajustar a proporção de dados no conjunto de dados da classe minoritária com a geração de dados artificiais.

Igualando as classes e balanceando os grupos conforme pode ser observado na figura 2. Uma das técnicas mais utilizadas é o *synthetic minority oversampling technique - SMOTE* (WOTEKI; KINEMAN, 2003).

Figura 1 – *Undersample* da classe majoritária.

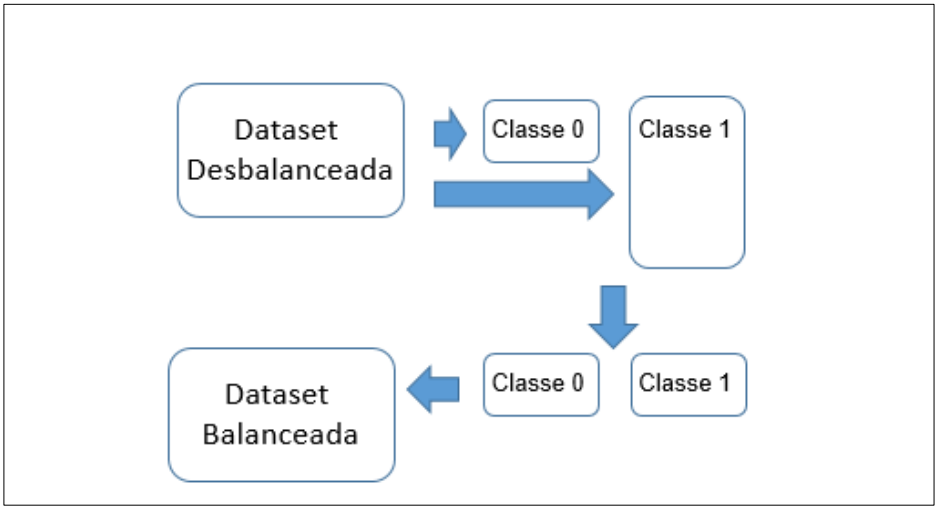
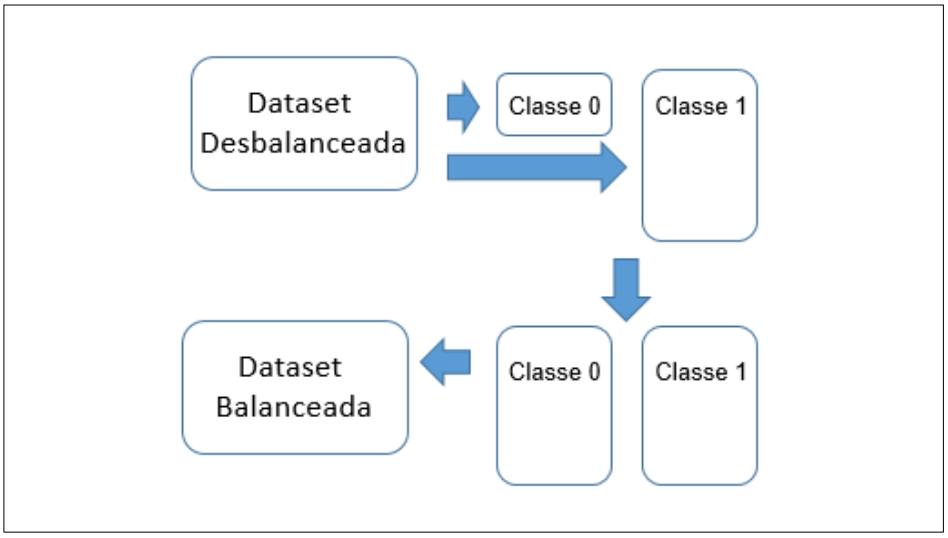


Figura 2 – *Oversample* da classe minoritária.



2.4 Synthetic Minority Oversampling Technique - SMOTE

É um método de *sobreamostragem* de registros da classe minoritária. Onde há a criação de novos registros artificiais com a interpolação com a classe minoritária. Esses registros são criados selecionando aleatoriamente um ou mais k-vizinhos mais próximos dos registros da classe minoritária do *dataset* (LIN et al., 2017).

Diferenciar os níveis de complexidade de amostras balanceadas ou não é uma tarefa que exige entendimento dos dados amostrados. Estes dados muitas vezes não estão balanceados, o que torna complexa a decisão dos usuários ao confiar nas classificações realizadas por sistemas que utilizem a IA. A TRI proporciona indicar quais registros são fáceis ou não em uma classificação.

2.5 Parâmetros da TRI

A teoria da resposta ao item é uma teoria de teste baseada em modelos probabilísticos para determinar a relação entre os itens de teste e a capacidade dos indivíduos, na qual define a probabilidade de resposta de um examinado a um item de teste como uma função da capacidade latente do examinado e das características do item. Para isto foi desenvolvida uma escala de pontuação que leva em conta as diferenças na dificuldade dos testes (CHANG; VATTIKUTI; CHOW, 2019).

$$\pi = c_{i+} \frac{1 - c_i}{1 + \exp(-a_i (\theta_j - b_j))} \qquad (1)$$

$$i = 1,2,3,\ldots,n$$

onde:

i - Item classificado
j - Classificador
π - Função dos parâmetros para os efeitos da habilidade do classificador e as propriedades do item
θ - Habilidade do classificador
a_i - Discriminante do item
b_i - Dificuldade do item
c_j - Probabilidade do classificador acertar ao acaso, também conhecida por *Guessing*

Os parâmetros da TRI estão presentes nos três modelos logísticos - ML. Estes modelos combinam os parâmetros da discriminação (a), da dificuldade (b) e da probabilidade de acerto ao acaso (c). É importante selecionar o ML especifico com maior importância de *b* para ser usado na seleção dos *clusters* do *ensemble*.

2.6 Modelos logísticos da TRI

Existem três modelos na TRI para dados dicotômicos a serem observados na tabela 1. Os parâmetros logísticos (PL) estão representados como se pode observar na equação 1 para cada um dos modelos. Cada modelo tem um papel importante na avaliação do item em relação ao conjunto a ser classificado pelos algoritmos (MARTÍNEZ-PLUMED et al., 2019)

Tabela 1**:** Relação dos parâmetros da TRI por modelo logístico.

Modelo PL da TRI	A (Discriminação)	B (Dificuldade)	C (*Guessing*)
1 PL	Constante igual a 1	Valor calculado entre [-4,+4]	Constante igual a 0
2 PL	Valor calculado entre [-4,+4]	Valor calculado entre [-4,+4]	Constante igual a 0
3 PL	Valor calculado entre [-4,+4]	Valor calculado entre [-4,+4]	Valor calculado entre [0,1]

Fonte: BAKER (2001).

Estes modelos seguem os três princípios da TRI, que são o foco na resposta, o reconhecimento da natureza randômica das respostas e a necessidade probabilística para explicar as distribuições e a presença de parâmetros separados para os efeitos das habilidades ou classificações dos algoritmos e as prioridades dos itens (MORAES et al., 2022).

3. Método

Como classificadores foram utilizados 13 algoritmos não supervisionados implementados no R com a biblioteca "clvalid", nominalmente: *fanny.fit.manhattan, fanny.fit.SqEuclidean, pam.fit.euclidean, pam.fit.manhattan, clara.fit.euclidean, clara.fit.manhattan, clara.fit.jaccard, sota.fit, k.means.fit.HartiganWong, k.means.fit.Lloyd, k.means.fit.Forgy e k.means.fit.MacQueen.*

Os algoritmos utilizados foram o *Fuzzy Analysis Clautering* (*Funny*); o *Partitioning Around Medoids* (*PAM*); o *Clustering large Application* (*Clara*); o *Self-Organization Tree Algorithm* (*SOTA*); e o *K-means*. Estes algoritmos também foram avaliados pela dissimilaridade na formação dos grupos.

Para os algoritmos de *Fanny, Pam, Clara* e *Sota* as medidas de dissimilaridades aplicadas foram a *Euclidean, Manhattan, Square Euclidean e Jaccard.* O algoritmo *K-means* foi avaliado em 4 métodos de atribuições de posicionamento de centroides, que são: *HartiganWong, Lloyd, Forgy e MacQueen*.

Para o cálculo do MSE em *datasets* de forma não-supervisionada deve-se realizar a classificação pelo algoritmo e posteriormente comparar com o vetor de classes do *dataset* original. Foram utilizados nove *datasets* distintos e originalmente desbalanceados conforme pode ser verificado na tabela 2. Os *datasets* D1 ao D4 foram obtidos de Martínez-Plumed *et al*. (2019) do *link* (https://github.com/nandomp/IRT4ML), com os dados tratados e sem dados faltantes. Os demais *datasets* utilizados neste trabalho foram obtidos do sítio https://archive.ics.uci.edu/ml/datasets.php, os quais não haviam dados faltantes. Os *datasets* de D6 à D9 foram normalizados e os demais já estavam normalizados com media igual a zero e variância igual a um.

Para cada um dos *datasets* foram calculados os três modelos logísticos da TRI e destes obtidos os respectivos valores de b de cada *dataset*. Para as avaliações nos modelos logisticos da TRI os dados foram transformados para dicotômicos pela técnica de *simulate response patterns* do pacote *mirt* do R (LA TORRE, DE; DENG, 2008).

Após obter os valores das respectivas classificações de cada algoritmo e os valores de b de cada modelo em cada *dataset*, foram avaliados estatisticamente os níveis de concordância dos algoritmos pelo método de Kappa de Fleiss com nível de confiança de $\alpha >= 0{,}95$. Esta avaliação indicaria o nível de concordância a ser ponderado na seleção dos algoritmos do *ensemble*, para uma decisão mais precisa de um sistema de tomada de decisão.

4. Resultados

Os algoritmos foram implementados na linguagem R versão 4.0.5, utilizando a biblioteca *mirt* na aplicação do TRI no código utilizado na análise. Os testes foram realizados em um microcomputador i3-8100 com *HD SSD* e 8*GB* de *RAM*. Os nove

datasets da tabela 1 foram classificados por 13 algoritmos não-supervisionados citados no descortinar do método. Os nove *datasets* escolhidos são compostos por diferentes de tipos de variáveis.

Esta composição de distintos tipos de variáveis pode impactar relevantemente na análise da TRI, mesmo que se faça a normalização dos dados. Não houve alteração do número de atributos, esta decisão foi tomada baseada na análise final ao comparar com os modelos para dados dicotômicos de 1PL, 2PL e 3PL da TRI. Na tabela 1 estão representadas percentualmente as quantidades de registros pelas respectivas classes desbalanceadas e balanceados pelo método de *SMOTE*.

Tabela 2: *Datasets* binárias utilizadas nas classificações pelos *clusters*.

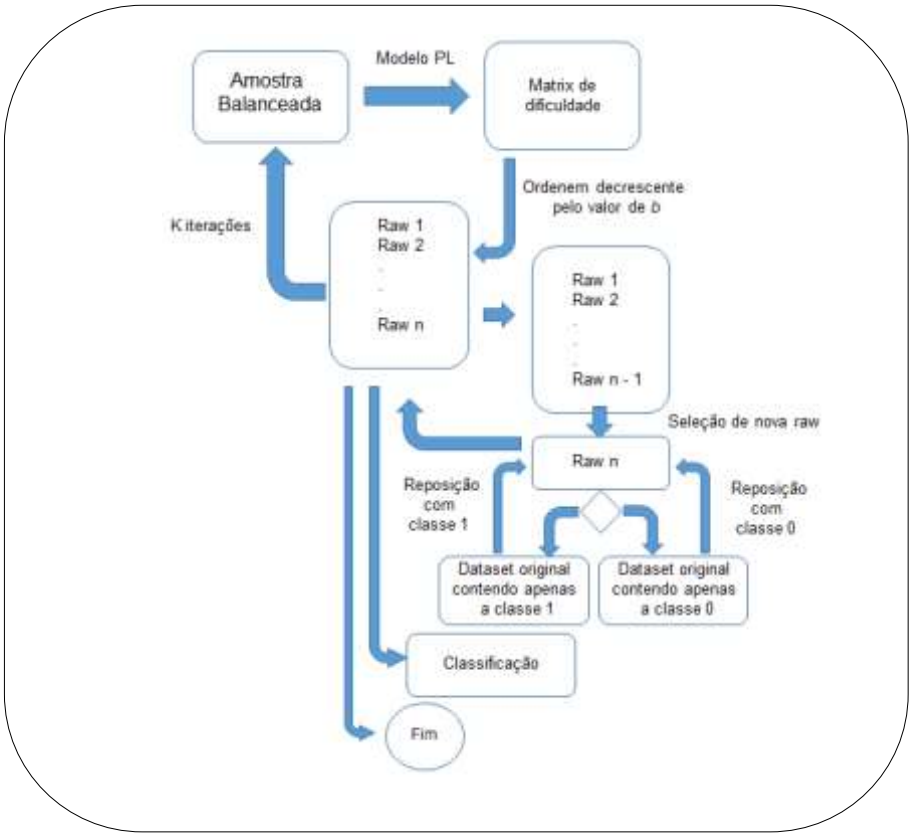
ID	Dataset	Características	Classes	Desbalanceada		Balanceada	
				Classe 0	Classe 1	Classe 0	Classe 1
D1	Echocardiogram	10	2	86 (77,47%)	25 (22,53%)	172 (49,57 %)	175 (50,43%)
D2	Heart-statlog	13	2	130 (56,52%)	100 (43,48%)	260 (56,52 %)	200 (43,48 %)
D3	Ionosphere	33	2	111 (34,68%)	209 (65,32%)	444 (51,51 %)	418 (48,49 %)
D4	Parkison	22	2	39 (13,18%)	145 (87,82%)	273 (48,49 %)	290 (51,51 %)
D5	Blood Transfusion	5	2	570 (76,20%)	178 (23,80%)	1140 (51,63 %)	1068 (48,37 %)
D6	Diabetes	8	2	268 (34,9%)	500 (65,1%)	1072 (51,74 %)	1000 (48,26 %)
D7	Madelon	500	2	1300 (50%)	1300 (50%)	1300 (50 %)	1300 (50 %)
D8	Vinnie	2	2	195 (48,68%)	185 (51,32%)	390 (51,32 %)	370 (48,68 %)
D9	Glass	10	2	138 (64,49%)	76 (35,51%)	276 (47,59 %)	304 (52,41 %)

Os tamanhos dos grupos para *undersample* foram ajustados para o grupo de menor tamanho (M) ou seja, 25 registros, no caso a classe 1 de D1, o tamanho de M foi mantido para todos os grupos de todos os *datasets* neste trabalho. Como a proposta de ajustar a amostra extraída dos *datasets* e permutar o elemento de menor dificuldade, a amostra padrão tem tamanho 24 (M-1), desta forma os grupos foram balanceados por subamostragem conforme experimentos realizados por (LIN et al., 2017). E para balancear as amostras com dados artificiais foi utilizado *SMOTE*, gerando mais registros do que os quantitativos originais, reduzindo o desbalanceamento. O D7 permaneceu com os mesmos quantitativos originais, pois a proposta é balancear, o D7 já é balanceado.

Após obter as amostras balanceadas por *undersample* ou *oversample,* como podemos observar na Figura 3*,* foram calculados os valores de *b* nos três ML da TRI, listados por *Rawx* onde *Raw* é o registro e *x* a linha correspondente, e armazenados em respectivas matrizes (MmLx), onde M é a matriz, mL (ML) e valor de *x* o ML. Em seguida ordenados decrescentemente pelos valores de *b*. Após esta ordenação, será selecionado o último elemento obtido de *b* na MmLx, como o valor de reposição do registro. Por conter o menor valor da Dificuldade. Para repor o registro mais fácil da amostra balanceada por outro valor proporcional a classe deste último registro é realizada uma amostragem aleatória na respectiva classe, sem que haja repetição de *Raw*, tornando a amostra sem

elementos repetidos, proporcionando uma seleção de amostras mais difíceis, visto que os elementos mais fáceis são substituídos gradativamente por elementos aleatórios dos grupos originais dos *datasets*.

Figura 3 – Diagrama da iteração da reposição de registros fáceis pelo valor de *b*.



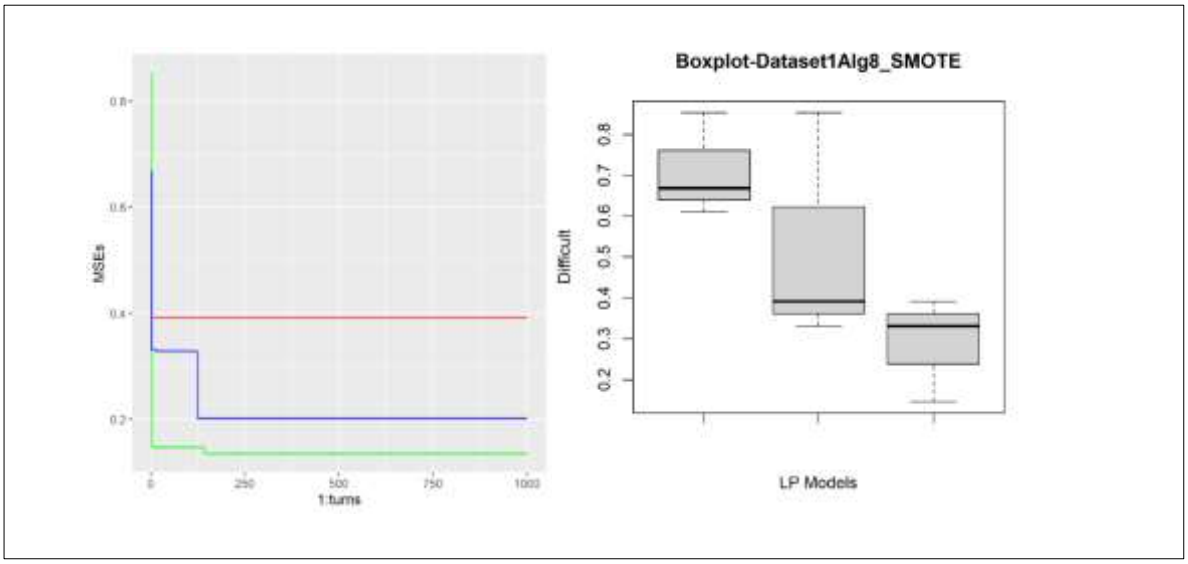
Na figura 3 também é possível observar que o ajuste ocorre nos valores da dificuldade após K interações, onde K é igual a 1000 (*turns*). A cada iteração são realizadas as respectivas classificações de cada um dos 13 algoritmos e destas classificações são calculados os *MSEs*

As amostras com a menor mediana de MSEs são conservadas utilizando a técnica de otimização de *machine learning* de *simulated annealing* (MOHAMMADI BIDHANDI et al., 2019), ajustando os valores das amostras quando o valor da nova reposição reduz a mediana do *MSE*, conforme a figura 3.

4.1. Avaliando as classificações dos algoritmos

Após as comparações entre os valores de *b* para obter os valores mais significativos. É possível avaliar entre as classificações dos algoritmos e os parâmetros dos ML do TRI para determinar quais parâmetros são mais relevantes dentre os três ML de 1PL, 2PL e 3PL da TRI, para realizar esta análise foram gerados 172 gráficos, onde o eixo Y representa os valores de MSEs, e o eixo X representa o número de iterações.

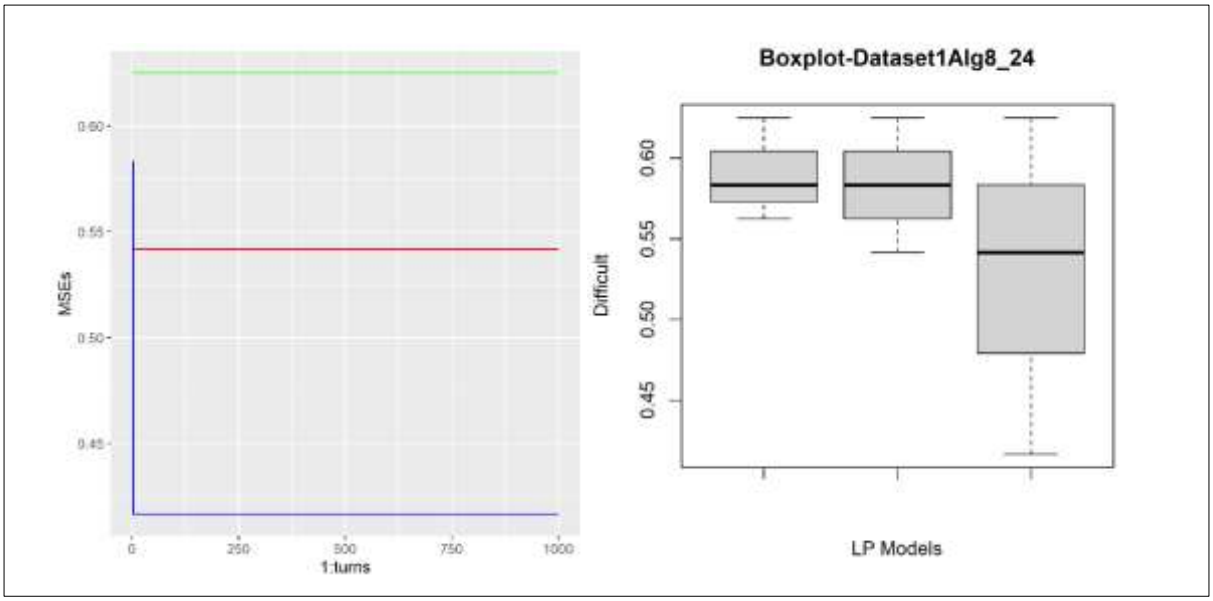
Figura 4 – Desempenho da classificação do algoritmo 8 ajustado pela dificuldade no *dataset* 1 – *SMOTE*.



1PL na cor vermelha, 2PL na cor verde e 3PL na cor Azul. E para comparar as medianas da amostra final das dificuldades foram gerados outros 172 gráficos de *boxplot*, representando como o 1PL o *boxplot* da esquerda, o 2PL o central e o 3PL o da direita. Os gráficos são proporcionais a cada uma das técnicas de undersample e oversample totalizando 354 gráficos.

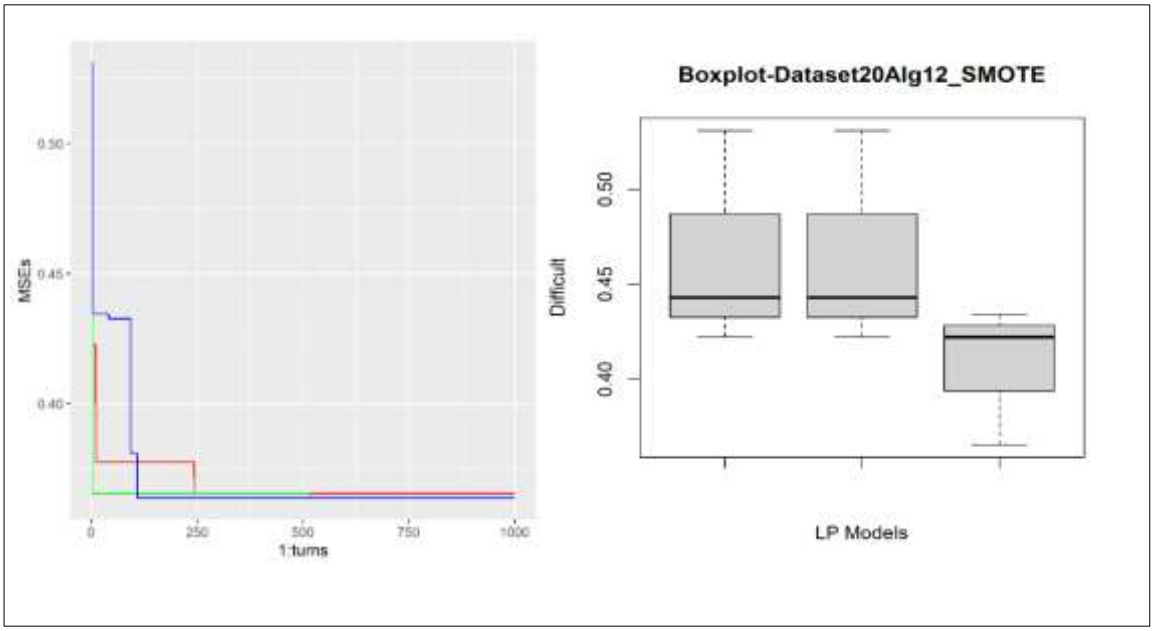
Após obter as amostras balanceadas por *undersample* ou *oversample,* como podemos observar na figura 3*,* foram calculados os valores de b nos três ML da TRI, listados por *rawx* onde *raw* é o registro e x a linha correspondente, e armazenados em respectivas matrizes (MmLx), onde M é a matrix, mL (ML) e x o ML. Em seguida ordenados decrescentemente pelos valores de *b*.

Figura 5 – Desempenho da classificação do algoritmo 8 ajustado pela dificuldade no *dataset* 1 – u*ndersample*.



Foram obtidos diferentes resultados dos valores de b nos três modelos PLs. Na Figura 4 à esquerda é possível observar que o 2PL, em verde, representa a tendência de menor MSE para o algoritmo 8 no dataset 1 na técnica de SMOTE, após o cálculo das dificuldades com o acréscimo de dados artificiais.

Figura 6 - Desempenho da classificação do algoritmo 12 ajustado pela dificuldade no *Dataset* 9 – *SMOTE*.



A amostra final da dificuldade nos três modelos PLs estão representados nos *boxplot* a direita da figura 7. Neste gráfico é possível observar que a distribuição de 1PL, à esquerda, tem a maior dificuldade, seguidos pelo 2PL, ao centro e no 3PL a direita. Observando que os gráficos de 2PL e 3PL estão com as medianas bem próximas, assim como nos traçados verde e azul dos MSEs.

Figura 7 - Desempenho da classificação do algoritmo 12 ajustado pela dificuldade no *dataset* 9 – *undersample*.

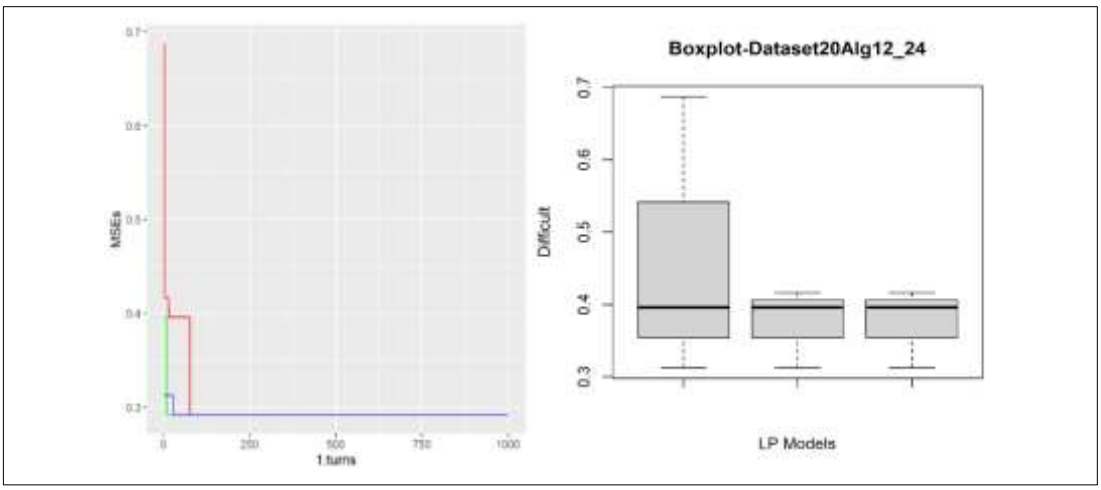
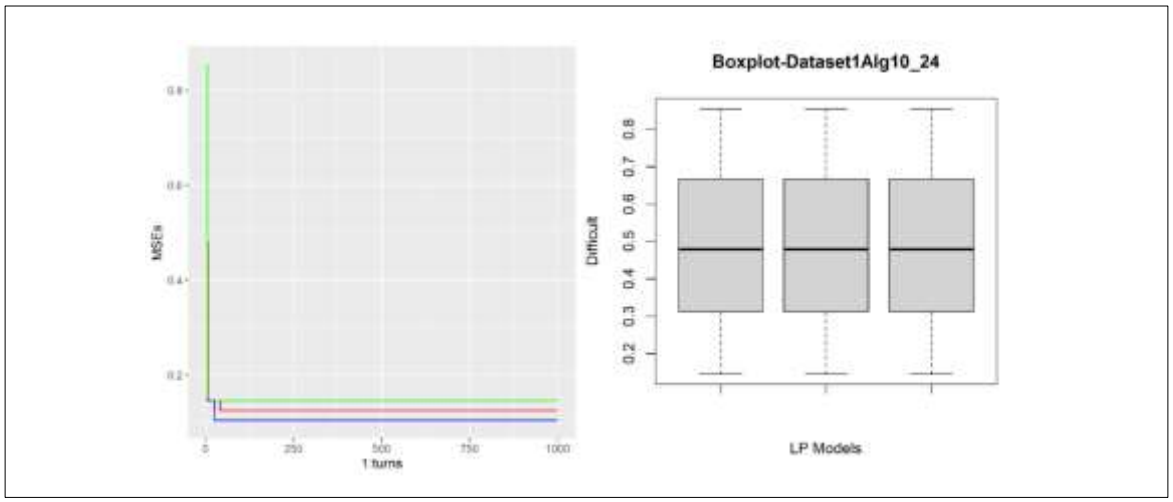
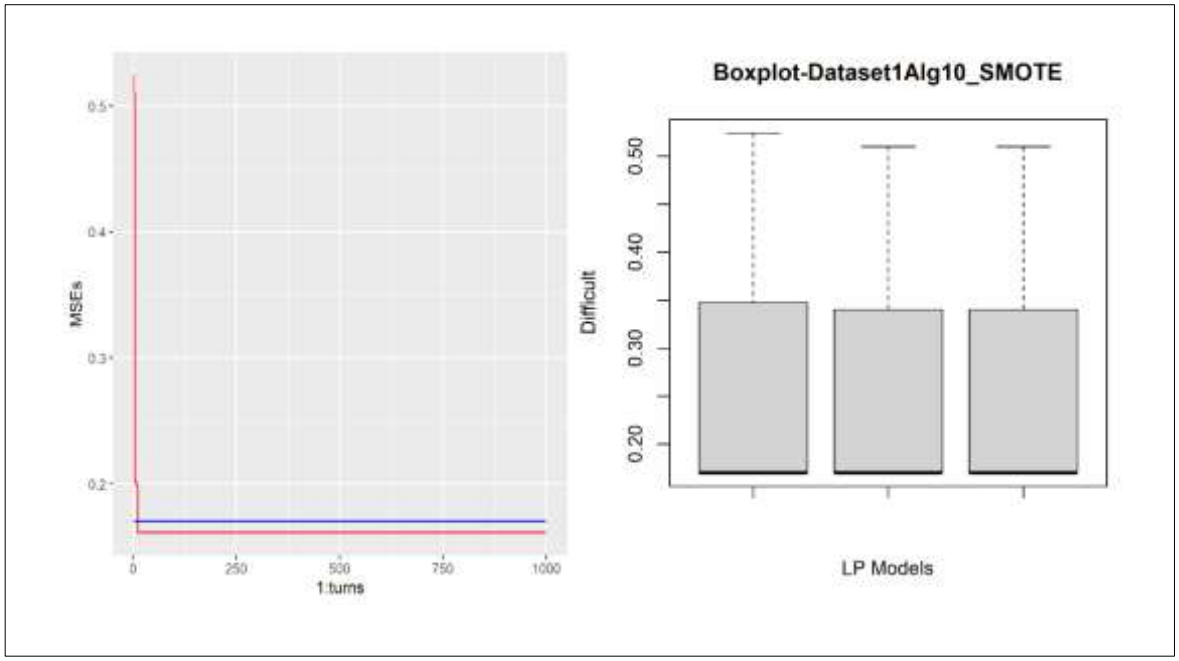


Figura 8 - Desempenho da classificação do algoritmo 10 ajustado pela dificuldade no *dataset* 1 – *undersample*.



Para os gráficos da figura 8 que representa as distribuições com *undersample* os modelos 1PL em vermelho e 2PL em verde estão com maiores *MSEs*. Assim como representados nos *boxplots* da esquerda e centro com medianas bem próximas, mas com a distribuição de 2PL maior do que 1PL. E observe que o 3PL em azul tem menor *MSE* e a distribuição de Dificuldade menor que as demais.

Figura 9 - Desempenho da classificação do Algoritmo 10 ajustado pela Dificuldade no *Dataset* 1 – *SMOTE*.



Na figura 8, se pode observar que a o algoritmo 12 no *dataset* 9 segue o método de *oversample* e que as medianas dos três modelos PLs estão bem próximas, assim como

nos traçados de *MSEs*. Observe-se que o menor *MSE* está representado no 3PL em azul, assim como o limite inferior de 3PL a direita.

Para o caso de *undersample* do algoritmo 12 no *dataset* 9 na figura 7 é possível observar que as medianas são muito similares para os *boxplots* dos três modelos. E os traçados dos *MSEs* são bem próximos na iteração final.

Nas figuras 8 e 9 os *boxplots* estão com as medianas da dificuldade semelhantes para cada ML Porem é possível observar que ao realizar a sobreamostragem na figura 9 há um deslocamento da dificuldade para valores menores. Agrupando os traçados dos *MSEs* para mais próximos entre os modelos PLs, como se pode observar na figura 9.

Como se depreende, é muito complexo analisar os comportamentos das dificuldades com os *MSEs* de cada algoritmo nos diferentes *datasets*. Para uma melhor avaliação foram calculados os *MSEs* de cada algoritmo no *Turn* igual a 1.000 e submetidos à estatística de Kappa de Fleiss para avaliar o nível de concordância dos algoritmos.

2.5 Nível de concordância das classificações pelo Kappa de Fleiss.

A estatística de Kappa de Fleiss tem como objetivo identificar o grau de concordância e confiança das classificações dos algoritmos nos diferentes modelo PL em cada *dataset* (MUNOZ; BANGDIWALA, 1997). Os valores da estatística de Kappa de Fleiss foram calculados para os grupos de 24 amostras e *SMOTE*, com $\alpha = 0.05$ e H_0 como concordância e H_1 como discordância. Esta avaliação estatística é muito importante para avaliar o nível de concordância e confiança dos algoritmos de classificação na formação dos *ensembles* descritos no método, proporcionando assim maior segurança nas tomadas de decisões das classificações.

Foram avaliados os resultados de todas as combinações das classificações dos algoritmos nos respectivos *datasets* em um total de 144 combinações (excluindo os grupos de apenas 1 elemento), multiplicado por 3 ML e totalizando 1.872 registros para cada um dos dois grupos de *undersample* e *oversample*. Os *p-values* menores que α proporcionaram selecionar no grupo de 24 amostras (*undersample*) os algoritmos para comporem o *ensemble* ideal, porém o valor da força da classificação da Kappa foi igual a 0,1665, considerada como muito fraca, equivalente a 18 no total de1.872 ocorrências. Onde: [<0] pobre, [0 – 0,2] muito fraca, [0,21 – 0,4] fraca, [0,41 – 0,6] moderada, [0,61 – 0,8] substancial e [0,81 – 1] perfeita (MUNOZ; BANGDIWALA, 1997), indicando como ideias para compor o *ensemble* os algoritmos *fanny.fit.SqEuclidean, pam.fit.euclidean e pam.fit.manhattan.*

Para o grupo de *SMOTE* (*oversample*) todos foram classificados como perfeitos, variando de [0.9968 a 1] equivalente ao total de 1.872 ocorrências. Ao selecionar apenas o valor igual a 1 equivalente a 9 de 1.872 ocorrências, foram selecionados como ideias para compor o *ensemble* os algoritmos *fanny.fit.manhattan, fanny.fit.SqEuclidean, pam.fit.euclidean, pam.fit.manhattan, clara.fit.euclidean, clara.fit.manhattan, clara.fit.jaccard, sota.fit e k.means.fit.HartiganWong.*

Ao realizar esta seleção foi possível observar que os melhores ML para *b* foram o de 1PL para o grupo de 24 amostras e para o de *SMOTE* o de 2PL. A partir desta especificação foram realizadas as avaliações pelos *MSEs* nos dois grupos e foram comparados os quantitativos de acertos pelos níveis de *b* nos respectivos modelos PLs.

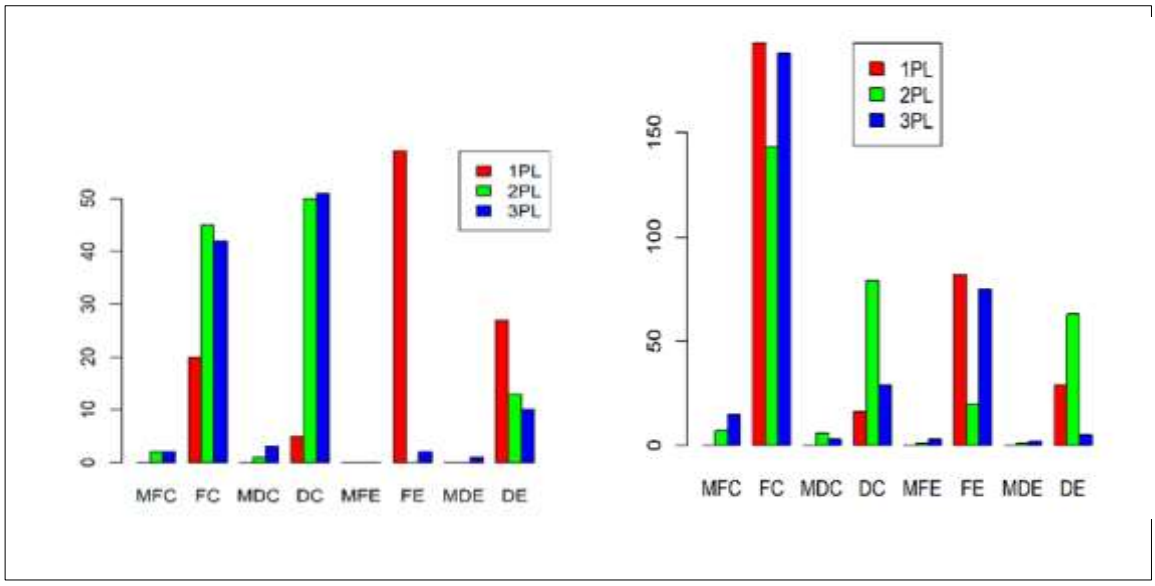
4.3 Proporcionalidade de questões difíceis por ML

Para uma melhor compreensão dos níveis de *b*, os quais foram divididos da seguinte maneira: MFC (Muito fácil e classificado corretamente) [-4,-2], FC (Fácil e classificado corretamente) [-2,0], MDC (Muito difícil e classificado corretamente) [0,+2], DC (Difícil e classificado corretamente) [+2,+4], MFE (Muito fácil e classificado incorretamente) [-4,-2], FE (Fácil e classificado incorretamente) [-2,0], MDE (Muito Difícil e classificado incorretamente) [0,+2] e DE (Difícil e classificado incorretamente) [+2,+4] (ANDRADE; VALLE, 2000).

Na figura 10 estão representados os gráficos dos somatórios de acertos e erros de classificação em todos os *datasets* da seguinte maneira: o gráfico da esquerda representa o grupo de 24 amostras (*undersample*) e no da direita o de *SMOTE* (*oversample*) nos 3 modelos PLs.

O eixo Y dos gráficos informa o somatório de itens classificados em todos os *datasets*, e no eixo X as categorias de acerto ou erro informadas no parágrafo anterior. É possível observar que ao realizar a sobreamostragem pelo *SMOTE* o quantitativo de erros cai e o nível de dificuldade também decai em relação aos três modelos PLs. Uma particularidade nos algoritmos avaliados neste trabalho é o comportamento de inclusão nos *ensembles*, particularmente falando o caso do *k-means*.

Figura 10 – Gráficos das comparações das classificações nos dois grupos avaliados.



O algoritmo baseado em *k-means* obteve grande quantidade de acertos em DC no grupo de SMOTE. Porém na subamostragem o *k-means* sofre penalização devido o desbalanceamento. No grupo de *undersample* o *k-means* tem rendimento muito afetado. Desta forma o *k-means* é bastante avaliado em estudos envolvendo desbalanceamentos de classes e (PURI; GUPTA, 2022).

5. Conclusão

O processo de tomada de decisões em *datasets* não-supervisionadas é uma tarefa complexa e está associada à capacidade de classificação pelos algoritmos. Identificar o nível de acerto na classificação de determinado registro no *dataset* não explicita que um algoritmo é ótimo, só informa que ele acertou ou não. Inferir graus de dificuldades dos itens e associar estas dificuldades aos acertos de classificação indicam um nível mais seguro de tomada de decisão. Avaliar estatisticamente se os algoritmos que compõe um *ensemble* concordam nas classificações e identificar a intensidade desta concordância também proporciona mais um nível de segurança na tomada de decisão em sistemas baseado em inteligência artificial.

Referências

BAKER, F. B.; KIM, S.-H. The Basics of Item Response Theory. S**tatistics for Social and Behavioral Sciences**, Springer International Publishing. Cham, Switzerland. p. 188. 2017.

BURGGRÄF, P.; WAGNER, J.;KOKE, B.; BAMBERG, M. Performance assessment methodology for AI-supported decision-making in production management. **Procedia CIRP**, v. 93, p. 891–896. 2020.

DOUZAS, G.; RAUCH, R.; BACAO, F. G-SOMO: An oversampling approach based on self-organized maps and geometric SMOTE. **Expert Systems with Applications**, v. 183, n. March, p. 115230. 2021.

HATHALIYA, J. J.; TANWAR, S. An exhaustive survey on security and privacy issues in Healthcare 4.0. **Computer Communications**, v. 153, n. September 2019, p. 311–335. 2020

LA TORRE, J. DE; DENG, W. Improving person-fit assessment by correcting the ability estimate and its reference distribution. **Journal of Educational Measurement**, v. 45, n. 2, p. 159–177. 2008.

LIN, W. C.; TSAI, C. F.; HU, Y. H.; JHANG, J. S..Clustering-based undersampling in class-imbalanced data. **Information Sciences**, v. 409–410, p. 17–26. 2017.

MOHAMMADI BIDHANDI, H. . Capacity planning for a network of community health services. **European Journal of Operational Research**, v. 275, n. 1, p. 266–279. 2019.

MOHEDANO-MUNOZ, M. A.; ALIQUE-GARCÍA, S.; RUBIO-SÁNCHEZ, M. RAYA, L. SANCHEZ, A. Interactive visual clustering and classification based on dimensionality reduction mappings: A case study for analyzing patients with dermatologic conditions. **Expert Systems with Applications**, v. 171, n. January, p. 114605. 2021.

MORAES, J. V. C.; REINALDO, J. T.S.; FERREIRA-JUNIOR, M.; FILHO, T. S.;PRUDÊNCIO, R. B.C. Evaluating regression algorithms at the instance level using item response theory. **Knowledge-Based Systems**, v. 240, p. 108076. 2022.

MUNOZ, S. R.; BANGDIWALA, S. I. Interpretation of Kappa and B statistics measures of agreement. **Journal of Applied Statistics**, 1 fev. v. 24, n. 1, p. 105–112. 1997.

PURI, A.; GUPTA, M. K. Improved Hybrid Bag-Boost Ensemble with K-Means-

SMOTE-ENN Technique for Handling Noisy Class Imbalanced Data. **Computer Journal**,v. 65, n. 1, p. 124–138. 2022.

TAHIR, M. A.; KITTLER, J.; YAN, F. Inverse random under sampling for class imbalance problem and its application to multi-label classification. **Pattern Recognition**, v. 45, n. 10, p. 3738–3750. 2012.

TANG, W.; YANG, Y.; ZENG, L.; ZHAN, Y.. Optimizing MSE for clustering with balanced size constraints. **Symmetry**, v. 11, n. 3. 2019.

WOTEKI, C. E.; KINEMAN, B. D. C Hallenges and a Pproaches To R Educing. **Review Literature And Arts Of The Americas**, 2003. n. June, p. 82–89. 2003.

WU, X.; LI, P.; HU, X. Learning from concept drifting data streams with unlabeled data. **Neurocomputing**,v. 92, p. 145–155. 2012.

ZHENG, Y.; CHEON, H.; KATZ, C. M. Using Machine Learning Methods to Develop a Short Tree-Based Adaptive Classification Test: Case Study With a High-Dimensional Item Pool and Imbalanced Data. **Applied Psychological Measurement**, v. 44, n. 7–8, p. 499–514. 2020.

ZHOU, Q.; SUN, B.; SONG, Y.; LI, S.. K-means Clustering Based Undersampling for Lower Back Pain Data. **ACM International Conference Proceeding Series**, p. 53–57. 2020.