LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

# Subamostragem Apoiada por Sistemas Imunológicos Artificiais para Aprimorar o Balanceamento de Dados em Contextos de Saúde

**Reinaldo Junio Dias de Abreu**
Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade de São Paulo (USP)
Av. Trab. São Carlense, 400 - Centro, São Carlos - SP, 13566-590
reinaldodias@usp.br

**Graciely Duarte Dias, Luciana Balieiro Cosme, Laércio Ives Santos**
Instituto Federal de Educação, Ciência e Tecnologia do Norte de Minas Gerais (IFNMG)
Rua Dois, 300 - Village do Lago I - Montes Claros/MG - CEP: 39404-058

**Marcos Flávio Silveira Vasconcelos D'Angelo**
Departamento de Ciência da Computação - UNIMONTES
Campus Universitário Professor Darcy Ribeiro – Avenida Rui Braga, S/Nº – Vila Mauricéia

## RESUMO

O desbalanceamento de dados apresenta um desafio significativo na análise e classificação de conjuntos de dados, especialmente quando se trata de técnicas de subamostragem. O *One-Sided Selection (OSS)* é uma abordagem comum, porém sua eficácia pode ser limitada devido à remoção ineficiente de exemplos redundantes da classe majoritária. Este artigo propõe uma solução alternativa, denominada *Subsampling Supported by Artificial Immune Systems* (SSAIS), que incorpora princípios de detecção de padrões das redes imunes artificiais para uma seleção no *OSS* com mais informação. Lidar com dados desbalanceados permanece como um desafio contínuo no campo da saúde; por essa razão, este artigo utiliza um conjunto de dados relacionado à gestão da diabete como exemplo para ilustrar a eficácia da metodologia proposta.

**PALAVRAS CHAVE. Subamostragem, Sistemas Imunes Artificiais, Dados Desbalanceados.**

## ABSTRACT

Data imbalance poses a significant challenge in the analysis and classification of datasets, especially when it comes to subsampling techniques. One-Sided Selection (OSS) is a common approach, but its effectiveness may be limited due to inefficient removal of redundant examples from the majority class. This paper proposes an alternative solution, named Subsampling Supported by Artificial Immune Systems (SSAIS), which incorporates principles of pattern detection from artificial immune networks for a more informed selection in OSS. Addressing imbalanced data remains an ongoing challenge in the field of healthcare, which is why this paper utilizes a dataset related to diabetes management as an example to illustrate the efficacy of the proposed methodology.

**KEYWORDS. Subsampling, Artificial Immune Systems, Unbalanced Data.**

## 1. Introdução

A presença de dados desbalanceados é uma questão comum em muitos problemas de aprendizado de máquina e mineração de dados [Sarker, 2021]. Refere-se a situações em que as classes de interesse em um conjunto de dados são representadas por um número significativamente menor de exemplos em comparação com outras classes. Esse desequilíbrio pode levar a modelos enviesados e sub-ótimos, pois os algoritmos tendem a favorecer a classe majoritária em detrimento das classes minoritárias. Várias abordagens são propostas para lidar com dados desbalanceados em uma variedade de contextos. Uma das técnicas mais comuns é a reamostragem [Dablain et al., 2022], que envolve a modificação da distribuição das classes no conjunto de dados. Isso pode ser feito aumentando a representação das classes minoritárias (sobreamostragem) ou reduzindo a representação das classes majoritárias (subamostragem). Técnicas populares incluem SMOTE (*Synthetic Minority Over-sampling Technique*) [Chawla et al., 2002], ADASYN (*Adaptive Synthetic Sampling*) [He et al., 2008], *Tomek Links* [Tomek, 1976] e *OSS* (*One-Sided Selection*) [Kubat and Matwin, 1997].

Na aplicação de técnicas de aprendizado na área da saúde, o desbalanceamento de dados é uma preocupação significativa e pode comprometer a eficácia dos modelos de aprendizado de máquina e análise de dados. Além disso, a qualidade dos dados na saúde pode variar consideravelmente, com ruídos, *outliers*, e inconsistências sendo comuns. Esses desafios adicionais podem dificultar a criação de modelos robustos e confiáveis [Dakka et al., 2021]. É comum encontrar desbalanceamento entre as classes de interesse nos conjuntos de dados da saúde, especialmente em tarefas de diagnóstico médico, nas quais classes de doenças raras podem ter representatividade significativamente menor do que as classes de doenças comuns [Li et al., 2017].

O *One-Sided Selection (OSS)* ou Seleção Unilateral, proposto por Kubat and Matwin [1997], é uma técnica de subamostragem que seleciona aleatoriamente elementos da classe majoritária para classificação, remove redundâncias e preserva a integridade da classe minoritária. O trabalho de Batista et al. [2000] enfatiza a importância da remoção de casos da classe majoritária pelo *OSS*, ressaltando sua eficácia. Contudo, é evidente a demanda por novas heurísticas que aprimorem tal seleção para reduzir a seleção de exemplos redundantes.

Neste contexto, a contribuição deste trabalho se destaca ao introduzir e validar a hipótese de empregar Sistemas Imunes Artificiais (SIA), por meio de Redes Imunes Artificiais, para identificar padrões na distribuição da classe majoritária. Essa abordagem, denominada Subamostragem Apoiada por Sistemas Imunes Artificiais, em inglês *(Subsampling Supported by Artificial Immune Systems - SSAIS)*, propõe utilizar o SIA como heurística na seleção de exemplos majoritários no *OSS*, como ilustrado na Figura 1. O objetivo é alcançar uma seleção mais criteriosa desses exemplos, maximizando a eliminação de redundâncias na base de dados. Para avaliar sua eficácia, foram realizados experimentos comparativos com o método *OSS* tradicional em base de dados artificial e, depois, aplicados ao problema de classificação de indivíduos com diabetes.

O restante desse trabalho está organizado da seguinte maneira: a seção 2 apresenta o método proposto e o delineamento do experimento, na seção 3 os resultados obtidos são apresentados e discutidos e, por fim, as conclusões do trabalho são apresentadas na seção 4.

## 2. Abordagem Proposta

Os métodos de classificação têm dificuldade em tratar dados desequilibrados por tenderem a classificar todas as instâncias como pertencentes à classe majoritária em detrimento da classe minoritária que, em geral, se caracteriza como o evento de interesse [Li et al., 2017]. Dessa forma, é necessário suavizar este desequilíbrio ou utilizando uma subamostragem da classe majoritária, ou reamostrando a classe minoritária. As duas abordagens têm o objetivo de igualar a distribuição entre as classes [He and Garcia, 2009].

A técnica de subamostragem *One-Sided Selection (OSS)*, proposta por Kubat and Matwin [1997], consiste na seleção aleatória de um elemento na classe majoritária, utilizado para classificar

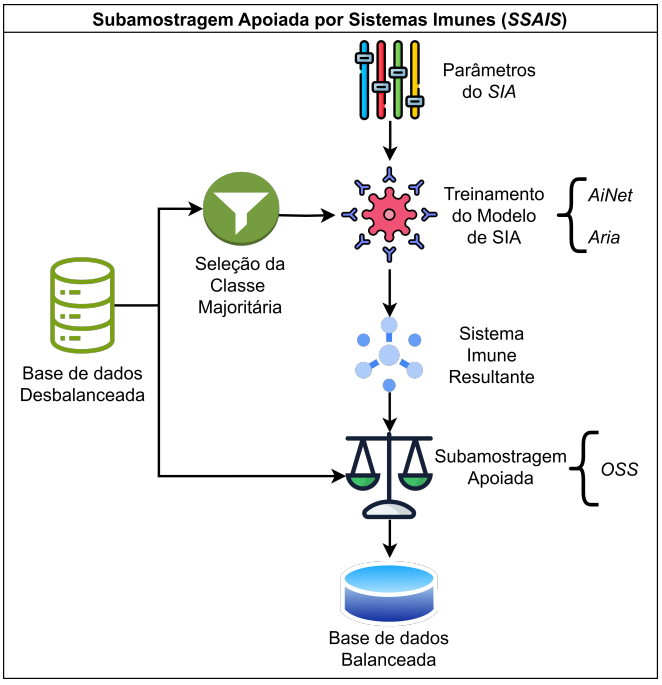



Figura 1: Subamostragem Apoiada por Sistemas Imunes Artificiais (*SSAIS*). Fonte: Dos Autores.

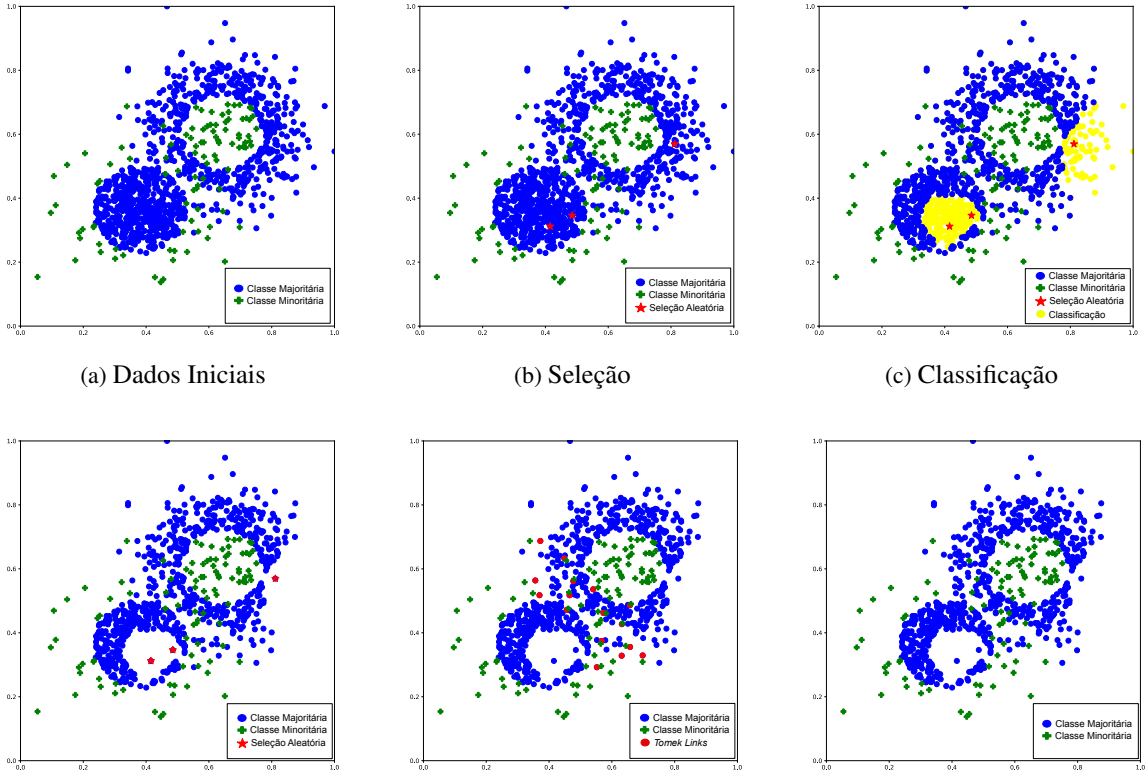
e eliminar os elementos redundantes dessa classe (elementos que não participam da fronteira de decisão entre as classes). Posteriormente, usando o conceito de *Tomek Links* [Tomek, 1976], remove os exemplos limítrofes e ruidosos, subamostrando a classe majoritária e mantendo a classe minoritária intacta. Essa estratégia está descrita no Pseudocódigo 1. Em implementações disponíveis, o *OSS* pode selecionar mais de um indivíduo aleatório da classe majoritária, com a quantidade definida pelo usuário, como na implementação disponível no pacote *Imbalanced-learn* [Lemaitre et al., 2017] da linguagem *Python*.

**Pseudocódigo 1** *One-Sided Selection (OSS)*

1: Dado um conjunto de dados desequilibrado inicialmente, chamado de $S$, com todos os exemplos minoritários e majoritários.
2: Cria-se um novo conjunto chamado de $C$, que inclui todos os exemplos da classe minoritária e **N exemplos aleatórios da classe majoritária**.
3: Usando o conjunto $C$ como base de treinamento, classifica-se $S$ usando **o vizinho mais próximo (1-NN)** e move-se todos os exemplos classificados incorretamente para $C$, que se torna consistente com $S$, mas que agora, não mantêm exemplos redundantes.
4: O próximo passo é remover em $C$ todos os exemplos majoritários que participam de *Tomek Links* com a classe minoritária, assim eliminando todos os exemplos ruidosos e limítrofes e mantendo todos da classe minoritária. O resultado é um novo conjunto de dados balanceado ($T$).

A seleção aleatória de exemplos pode não representar de forma eficiente a distribuição da classe majoritária, o que pode prejudicar a remoção de suas redundâncias. Este comportamento é ilustrado na Figura 2. Na Figura 2-A, o método *OSS* seleciona aleatoriamente $N$ exemplos majoritários (neste caso, 3 exemplos, destacados em vermelho, sendo essa quantidade definida pelo

usuário), conforme ilustrado na Figura 2-B. Esses exemplos selecionados são usados para classificar a classe majoritária, logo, os exemplos destacados em amarelo, na Figura 2-C, são os exemplos mais próximos (classificados corretamente) e por isso são descartados (Figura 2-D). Após essa etapa, os exemplos majoritários pertencentes aos *Tomek Links* (destacados em vermelho na Figura 2-E) são removidos. O resultado pode ser observado na Figura 2-F, onde vemos que ainda existem exemplos em outras regiões dos dados que poderiam ser eliminados, pois não fazem parte da fronteira de decisão com a classe minoritária. Isso ocorre devido à falta de seleção de exemplos majoritários nessas áreas da distribuição.




(a) Dados Iniciais (b) Seleção (c) Classificação

(d) Remoção (e) *Tomek Links* (f) Dados Finais

Figura 2: Exemplo de execução passo a passo do *OSS*. Fonte: Dos Autores.

Portanto, propomos neste trabalho a utilização dos Sistemas Imunes Artificiais (SIA) para obter uma amostragem mais informada do *OSS*. Os SIAs podem ser definidos como sistemas computacionais abstratos ou metafóricos desenvolvidos a partir das ideias, teorias e componentes existentes nos sistemas imunes naturais. Esses sistemas podem ser aplicados tanto a problemas de otimização de funções quanto ao reconhecimento de padrões. Quando aplicados no reconhecimento de padrões, a metáfora comumente utilizada é definir a base de dados que se deseja reconhecer como antígenos não próprios (como células invasoras), que serão reconhecidos pelos anticorpos artificiais criados [de Castro and Timmis, 2002].

Duas principais abordagens de SIAs baseadas na teoria das Redes Imunes Artificiais são utilizadas neste trabalho. Essa teoria sugere que o sistema é composto por uma rede regulada de células que se reconhecem, mesmo na ausência de antígenos ou estímulos externos. A abstração utilizada se baseia em que as moléculas receptoras contidas na superfície das células imunes apresentam marcadores, denominados idiótipos, que podem ser reconhecidos por receptores em outras

células imunes [de Castro and Timmis, 2002].

Como resultado dos reconhecimentos efetuados pela rede, sugere-se que quando um receptor de um anticorpo é reconhecido por outro anticorpo, isso resultará em supressão da rede, realizando assim a eliminação da redundância e a redução da população final de células de memória que reconhecem e seguem a distribuição espacial dos antígenos [Jeyse Freire Pinheiro, 2006]. Enquanto o reconhecimento de um antígeno por um anticorpo resulta na ativação de rede e proliferação celular.

Dada a teoria da rede imune, diversos modelos do mais simples ao mais complexo foram propostos, e dentre esses modelos podemos destacar os algoritmos *AiNet (Artificial Immune Network)* [De Castro and Von Zuben, 2002] e *Aria (Adaptive Radius Immune Algorithm)* [Bezerra et al., 2005] que foram utilizados para a realização dos experimentos. No método *AiNet* os anticorpos possuem a habilidade de reconhecer uns aos outros, o que possibilita a implementação de etapas de supressão destinadas a eliminar redundâncias. A rede é adaptativa quanto à quantidade de anticorpos, com a limitação estabelecida pelo raio de supressão empregado; consequentemente, os anticorpos finais mantêm uma distância mínima de acordo com o raio de supressão da rede. Já no método *Aria*, da mesma maneira que o *AiNet*, os anticorpos têm a capacidade de se reconhecerem. A diferença está na definição de um raio de supressão adaptativo, inversamente proporcional à densidade de antígenos alcançados, calculado a partir de um raio inicialmente aleatório. Durante as iterações, os raios são ajustados com base na densidade e a rede é suprimida, priorizando a manutenção dos anticorpos com raios menores.

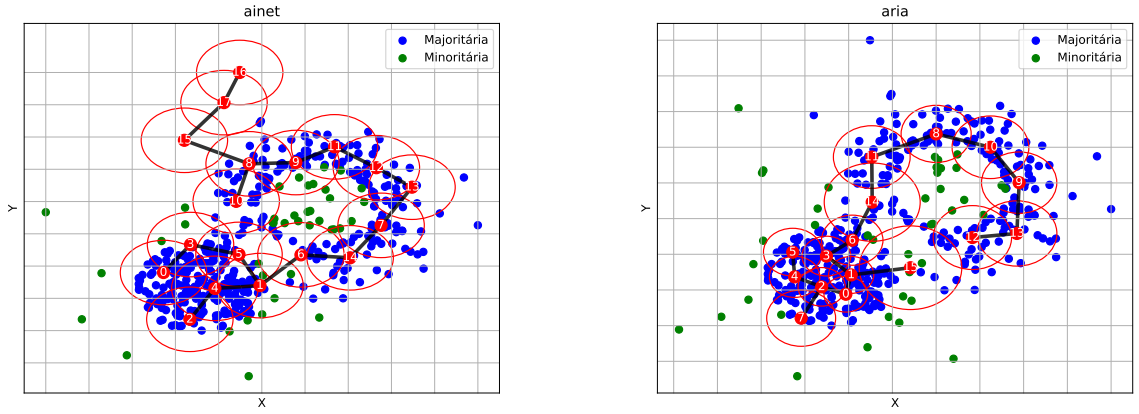
Ambos os algoritmos produzem uma rede de anticorpos artificiais (*Ab*) adaptados ao padrão presente na base de dados fornecida. Com um número específico de iterações e utilizando o conjunto de dados como antígenos (*Ag*), é possível treinar uma rede *Ab* visando detectar o referido padrão em $Ag$. A estratégia do algoritmo *AiNet* está representada no Pseudocódigo 2, no qual: $N$, representa o tamanho inicial da população de anticorpos; $\alpha_1$, é o raio de apoptose; e $\alpha_2$, é o raio de supressão clonal e de rede.

**Pseudocódigo 2** *AiNet para reconhecimento de padrões*

1: Dado um conjunto de dados inicial (*Ag*).
2: Gere aleatoriamente uma população de $N$ anticorpos iniciais *Ab* e divida em dois subconjuntos (memória -$Abm$ e Restantes - $Abr$)
3: **for** $\forall Ag_i \in Ag$ **do**
4: $Ab_n \leftarrow n$ mais similares a $Ag_i$
5: Cria C sendo uma população de clones de $Ab_n$ com tamanho proporcional a similaridade com $Ag_i$
6: Cria $C^m$ sendo a C após sofrer mutação de forma inversamente proporcional a $Ag_i$
7: Elimina todos os clones de $C^m$ com afinidade $< \alpha_1$ em relação ao $Ag_i$ (Apoptose)
8: Elimina todos os clones de $C^m$ com $afinidade > \alpha_2$ em relação a outro clone de $C^m$ (Supressão clonal)
9: Insere os clones restantes em $Abm$
10: **end for**
11: Elimina todos os $Abm$ com $afinidade > \alpha_2$ em relação a outro $Abm$ (Supressão em Rede)
12: $Ab \leftarrow Abm$ U $Abr$

Para ilustrar o funcionamento desses algoritmos, a Figura 3 apresenta a identificação do padrão antígeno criado por cada algoritmo (*AiNet* e *Aria*) sobre a classe majoritária de uma base de dados gerada artificialmente. Os dados utilizados foram propositalmente gerados de forma desequi-

librada, com 400 indivíduos na classe majoritária e 40 na classe minoritária. As redes imunes foram treinadas com um número fixo de 100 iterações. Na Figura 3, a classe majoritária é representada pela cor azul, e a minoritária, pela cor verde. Em vermelho e enumerados, estão os anticorpos artificiais, cada um com o raio de abrangência da supressão ilustrado por círculos também em vermelho.




(a) Rede AiNet (b) Rede Aria

Figura 3: Comparação da Detecção do Padrão Antígeno. Fonte: Dos Autores.

Podemos visualizar na Figura 3a a detecção do padrão antígeno criado pela rede *AiNet* para os dados majoritários, note que o valor definido para os raios é fixo, o que limitou o número de anticorpos produzidos. Na Figura 3b, o padrão detectado pela rede *Aria* possui raios diferentes por serem adaptativos; note que para a região com maior densidade de exemplos majoritários há um maior número de anticorpos devido a essa característica da rede.

É importante observar, nas redes produzidas no exemplo da Figura 3, que a rede de anticorpos conseguiu reproduzir o padrão que representa a disposição dos exemplos majoritários. Podemos supor que os anticorpos representem e formem o padrão característico da dispersão dessa classe.

### 2.1. Método Proposto

Diante da necessidade de uma seleção mais bem informada para a classificação de redundâncias no *OSS*, propõe-se o uso do reconhecimento de padrões em redes imunes artificiais para aprimorar esse processo, contribuindo, assim, para uma escolha mais precisa dos exemplos mais relevantes para a seleção no *OSS*.

O método proposto consiste em treinar uma rede imune artificial utilizando a classe majoritária e, em seguida, utilizar os anticorpos gerados para identificar os exemplos prioritários dessa classe. Esses exemplos serão então utilizados para a classificação de redundâncias, com o objetivo de alcançar um balanceamento de maior qualidade, que elimine mais exemplos majoritários e reduza o risco de perda de dados importantes para a classificação. Essa modificação para o *OSS* está descrita em detalhes no Pseudocódigo 3.

### 2.2. Experimentos Realizados

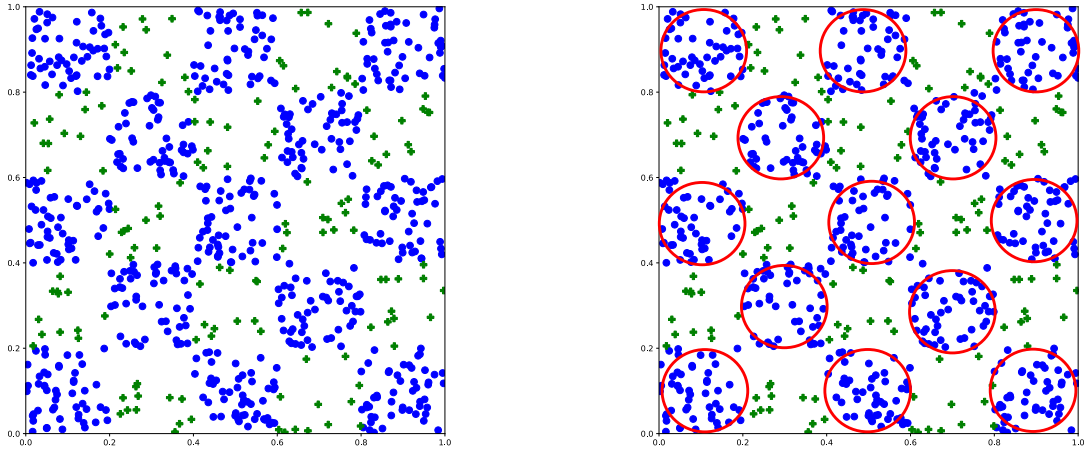
A aplicação do método proposto (*SSAIS* com *OSS* apoiado pelas redes *AiNet* ou *Aria*), em comparação ao *OSS* tradicional, será avaliada por meio do uso de uma base de dados artificial e de uma base de dados contendo pacientes com diabete.

No primeiro caso, a base de dados artificial foi gerada com o formato de tabuleiro de xadrez, inspirado no padrão utilizado por Fernández et al. [2018] para apresentar os efeitos do problema de aprendizagem desbalanceada. A base de dados apresentada na Figura 4a possui um total

**Pseudocódigo 3** *SSAIS para OSS*

1: Dado um conjunto de dados desequilibrado inicial, chamado de *S*.
2: **Treine um *SIA* (como *AiNet* ou *Aria*) sobre os exemplos da classe majoritária e retorne o conjunto de anticorpos resultante (*Ab*) que represente o padrão dos dados.**
3: Cria-se um subconjunto de *S* chamado *C*.
4: Adicione em *C*, todos os exemplos da classe minoritária de *S*.
5: **for** $\forall Ab_i \in Ab$ **do**
6: **Adicione em *C*, o exemplo da classe majoritária de *S*, mais próximo de** $Ab_i$**.**
7: **end for**
8: Usando o conjunto *C* como base de treinamento, classifique *S* usando o vizinho mais próximo (1-NN) e mova todos os exemplos classificados incorretamente para *C*, que se torna consistente com *S* mas que agora não mantêm exemplos redundantes.
9: Remova em *C* todos os exemplos majoritários que participam de *Tomek Links* com a classe minoritária, assim eliminando todos os exemplos ruidosos e limítrofes e mantendo todos da classe minoritária. O resultado é um novo conjunto de dados balanceado (*T*).

de 650 exemplos de dados para classe majoritária (em azul) e 144 exemplos da classe minoritária (em verde), dispostos em 25 grupos/casas com mesma quantidade de exemplos (13 grupos de 50 exemplos para classe majoritária, conforme destaque na Figura 4b, e 12 grupos com 12 exemplos para classe minoritária), gerados aleatoriamente no intervalo real uniforme entre 0 e 1 para ambas dimensões.



(a) Distribuição das Classes (b) Destaque dos grupos majoritários

Figura 4: Base de Dados Artificial com Padrão Xadrez. Fonte: Dos Autores.

A segunda base de dados utilizada para testar o método proposto é a base de dados indiana PIMA[1]. Ela é composta por pacientes do sexo feminino de até 21 anos; possui 768 pacientes, dos quais 268 (35%) são diabéticos e 500 (65%) não são; e 8 atributos, quais sejam: IMC, nível de insulina, idade, pressão arterial, espessura da pele, glicose, número de gestações e o desfecho (classe do problema).

Na base de dados PIMA, os experimentos foram elaborados utilizando 100 ensaios execu-

[1]Disponível em: https://data.world/data-society/pima-indians-diabetes-database

tados pelo *Framework Optuna* [Akiba et al., 2019] para ajustar os hiperparâmetros dos modelos de forma dinâmica e realizar a validação cruzada *5-fold*, na qual uma pasta foi utilizada para validação e as demais para treinamento do modelo. Os resultados médios das 5 pastas, em termos de AUC-ROC e da quantidade de elementos selecionados da classe majoritária por cada técnica de balanceamento, foram medidos em cada ensaio e apresentados. Para testar a eficiência da classificação, utilizamos os métodos Árvore de Decisão e *Perceptron* disponíveis na biblioteca *scikit-learn* da linguagem *Python*.

As duas técnicas de classificação *Perceptron* e Árvore de Decisão foram escolhidas por serem técnicas mais simples. O objetivo deste experimento foi avaliar a influência do balanceamento realizado pelo método proposto na classificação, em comparação ao *OSS* tradicional durante a execução dos ensaios. É importante mencionar que os hiperparâmetros dos modelos treinados das técnicas de classificação não foram ajustados pelo *Optuna*; foram utilizados os valores padrão configurados na biblioteca *scikit-learn*.

A implementação do método *SSAIS* para o *OSS* foi realizada em linguagem Python. A implementação do *OSS* é disponibilizada pelo pacote *Imbalanced-learn* para Python [Lemaitre et al., 2017]. Os resultados obtidos do balanceamento pelo método proposto foram tabelados e a base de dados balanceada é exibida para comparação com o resultado pelo *OSS* tradicional.

## 3. Resultados

O resultado da aplicação do método proposto na base de dados artificial é apresentado na Tabela 1 e ilustrado nas Figuras 5, 6 e 7. Na Tabela 1, esses resultados foram apresentados começando pelo método com maior redução da classe majoritária até o com menor redução. Em que #MAJ representa o número de exemplos majoritários após subamostragem; #RED, Número de exemplos majoritários reduzidos; e #AB, o tamanho da rede de anticorpos criada. Foram utilizados parâmetros de supressão nas redes imunes artificiais para que se obtivesse um número de anticorpos aproximado ao número de grupos majoritários da base de dados. Para o *OSS*, o número de elementos selecionados ($N$) foi definido como a quantidade exata desses grupos (13 elementos).

Tabela 1: Resultado do balanceamento na base de dados Artificial

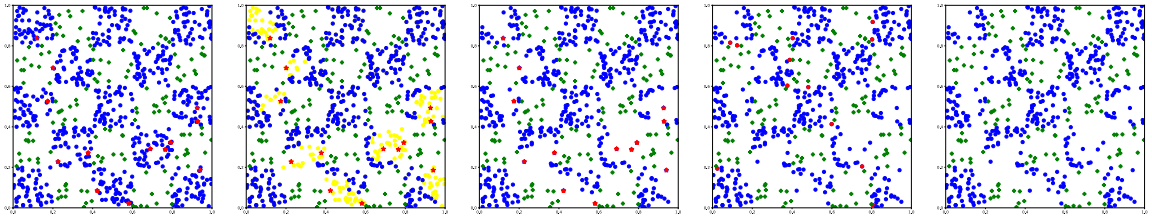
| **Balanceador** | **#MAJ** | **#RED** | **#AB** |
|---|---|---|---|
| *OSS+ARIA* | 369 | 281 | 13 |
| *OSS+AINET* | 381 | 269 | 15 |
| *OSS* | 475 | 175 | – |

Pode-se constatar, pelos resultados obtidos na Tabela 1, que as estratégias de subamostragem por *SSAIS* com as redes imunes, apresentaram uma melhor remoção de exemplos majoritários para o padrão da base de dados utilizada, e eliminaram dos grupos somente os exemplos mais centrais, que não fazem parte da fronteira de decisão.

As Figuras 5, 6 e 7, apresentam o passo a passo do processo de subamostragem para o *OSS* e do *SSAIS* com *AiNet* e *Aria*, respectivamente, executados na base de dados artificial. As Figuras A apresentam em vermelho os exemplos majoritários selecionados, as Figuras B, apresentam a classificação realizada (em amarelo estão os exemplos classificados corretamente), as Figuras C, exibem a eliminação dos exemplos classificados corretamente, as Figuras D, destacam em vermelho os exemplos que fazem parte dos *Tomek Links*, e por fim, as Figuras E, apresentam a base de dados final balanceada pelo método.

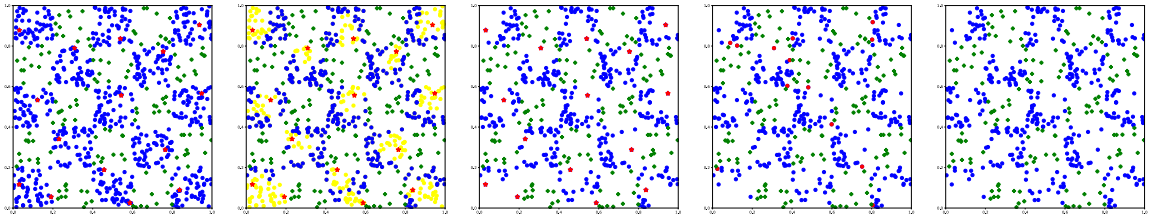
Alguns achados importantes podem ser derivados da análise das Figuras 5, 6 e 7. Primeiramente, o método *SSAIS* com *Aria* (Figura 7 - A) selecionou exatamente um indivíduo central de cada grupo, o que lhe permitiu uma melhor classificação e eliminação de redundâncias. O *SSAIS*

com *AiNet* apesar de selecionar mais de um indivíduo em dois dos grupos (Figura 6 - A), não deixou de ter selecionado algum representante de cada grupo majoritário, o que lhe conferiu uma boa redução da classe majoritária. O *OSS* tradicional, selecionando aleatoriamente 13 exemplos (Figura 5 - A), não obteve ao menos um exemplo majoritário selecionado em cinco dos treze grupos, o que conferiu uma menor redução de exemplos, ao deixar esses grupos intactos (Figura 5 - E).



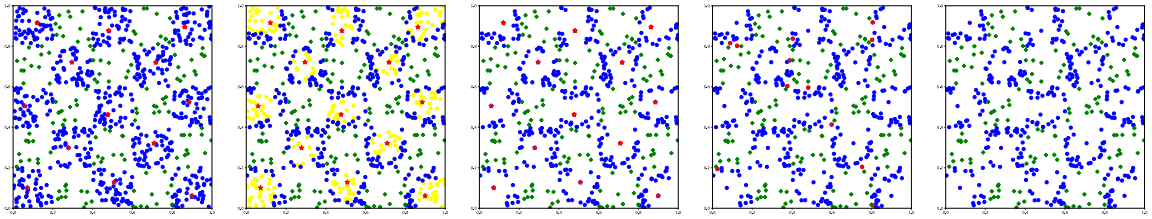
(a) (b) (c) (d) (e)

Figura 5: Execução passo a passo do *OSS*. Fonte: Dos Autores.



(a) (b) (c) (d) (e)

Figura 6: Execução passo a passo do *OSS+AINET*. Fonte: Dos Autores.



(a) (b) (c) (d) (e)

Figura 7: Execução passo a passo do *OSS+ARIA*. Fonte: Dos Autores.

É importante ressaltar que os resultados obtidos pelo método proposto, não só eliminaram melhor os exemplos redundantes da classe majoritária, como mantiveram os exemplos pertencentes a fronteira de decisão entre as classes. A seleção feita pela detecção de padrões das redes imunes artificiais foi importante para direcionar a seleção de um número relevante de exemplos majoritários, sem que esse número seja arbitrado pelo usuário diretamente, como também, a boa capacidade de percorrer o espaço de busca, que permitiu que grupos tivessem ao menos um representante majoritário selecionado. Em especial para este padrão de dados, o *Aria* obteve vantagem devido à característica adaptativa dos raios de supressão da rede, pois se aproveitou da densidade idêntica de exemplos majoritários de cada grupo.

Os resultados da aplicação do método proposto na base de dados PIMA são apresentados

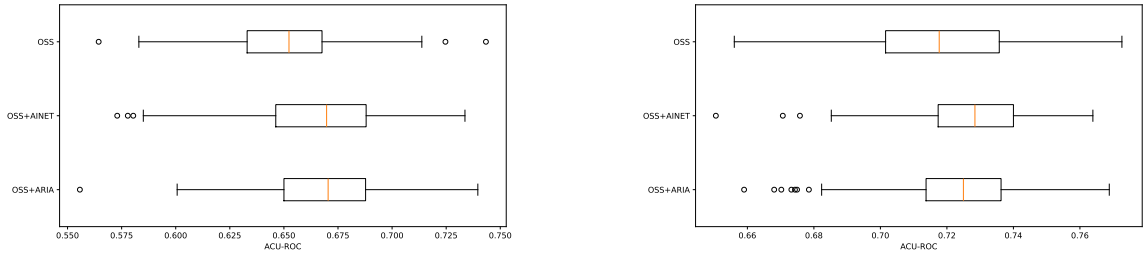
na Tabela 2 e na Figura 8. A Tabela 2 exibe os resultados médios dos 100 ensaios realizados com o *Optuna*, destacando as métricas AUC-ROC e a quantidade de indivíduos selecionados durante o processo de amostragem pelas abordagens OSS+Aria, OSS+AiNet e OSS tradicional, utilizando os classificadores *Perceptron* e Árvore de Decisão. A Figura 8 ilustra a variação dos resultados para cada um dos ensaios, com o *Perceptron* (Figura 8-A) e com a Árvore de Decisão (Figura 8-B), com o eixo $x$ representando os valores de AUC-ROC.

Na Tabela 2, é notável uma ligeira melhora em termos de AUC-ROC dos métodos apoiados pelo SIA (OSS + ARIA e OSS+AINET) em relação ao *OSS* tradicional, principalmente, quando o classificador utilizado é o *Perceptron*. Embora o método *OSS* puro tenha obtido resultados semelhantes de AUC-ROC, ele necessitou selecionar aleatoriamente uma quantidade média superior de exemplos majoritários. Isso representa que, em média, o *OSS* tradicional teve que utilizar 24% (no *Perceptron*) e 68.4% (na Árvore de Decisão) de exemplos da classe majoritária para atingir resultados comparáveis aos do método proposto, isso se deve ao otimizador utilizado, que sugere esses valores no decorrer dos ensaios para maximizar a métrica, neste caso foi sugerida uma maior seleção ao *OSS* puro para provavelmente não arriscar eliminar aleatoriamente exemplos relevantes. Já com o método proposto, a seleção guiada por SIA não enfrentou este problema, em que no máximo foram usados, em média, de 9.4% a 13.2% exemplos da classe majoritária.

Tabela 2: Resultados da Aplicação do método proposto na base de dados PIMA utilizando os classificadores *Perceptron* e Árvore de Decisão

| **Balanceador** | **AUC-ROC (*Perceptron*)** | **Selecionados (*Perceptron*)** | **AUC-ROC (Árvore)** | **Selecionados (Árvore)** |
|---|---|---|---|---|
| *OSS+ARIA* | 0.67 | 47 | 0.72 | 59 |
| *OSS+AINET* | 0.67 | 52 | 0.73 | 66 |
| *OSS tradicional* | 0.65 | 120 | 0.72 | 342 |

Já na Figura 8, observa-se a variação dos resultados de cada classificador e seus respectivos métodos de balanceamento. Essa visualização evidencia a superioridade do método proposto em comparação ao tradicional, ao exibir a distribuição da AUC-ROC nos 100 ensaios do otimizador. Nota-se que, ao utilizar apenas o *OSS* tradicional, há uma maior concentração de resultados com valores mais baixos de AUC-ROC, em comparação com os métodos suportados pelo SIA. Este comportamento ocorre principalmente nos ensaios realizados utilizando o classificador *Perceptron*.




(a) *Perceptron* (b) Árvore de Decisão

Figura 8: Resultados de AUC-ROC na execução dos 100 ensaios com Optuna. Fonte: Dos Autores.

A abordagem proposta representa um avanço no processo de seleção e classificação de redundâncias do *OSS*. Ao empregar redes imunes artificiais, como AiNet e Aria, o *SSAIS* demonstrou uma capacidade aprimorada de selecionar exemplos majoritários de forma mais informada,

resultando em uma melhor eliminação de redundâncias. Os resultados obtidos com a base de dados artificial desbalanceada e na base de dados PIMA evidenciam a eficácia do método proposto. Além de aprimorar a métrica AUC-ROC em relação ao *OSS* tradicional nos experimentos, o método também reduz significativamente a quantidade de exemplos selecionados necessários.

## 4. Conclusão

Este trabalho apresentou uma abordagem denominada Subamostragem Apoiada por Sistemas Imunes Artificiais. Essa abordagem utiliza os Sistemas Imunológicos Artificiais (SIA) para prover uma seleção mais criteriosa ao método Seleção Unilateral (*OSS*). Nos experimentos realizados em dois conjuntos de dados, a abordagem mostrou ter uma boa capacidade de selecionar exemplos relevantes da classe majoritária durante o processo de subamostragem, por meio da geração de anticorpos pelos SIA. As vantagens da abordagem incluem a capacidade de seleção baseada no padrão de distribuição dos dados, a eliminação de redundâncias de forma mais eficiente e a manutenção dos principais exemplos para classificação, sem a necessidade de conhecimento prévio do número exato de exemplos majoritários. No entanto, é importante reconhecer que o treinamento dos modelos de rede imune pode aumentar a complexidade do tempo de execução, uma vez que o processo deixa de ser uma simples seleção aleatória, exigindo também uma atenção maior na escolha dos hiperparâmetros ideais. Sugere-se a utilização de otimizadores de hiperparâmetros sobre a base de dados para contornar esse desafio. Para trabalhos futuros, recomenda-se a análise da abordagem em conjuntos de dados diversos e reais, a fim de destacar seus benefícios em problemas de classificação com bases desequilibradas e em diferentes contextos. Além disso, explorar estratégias adicionais, como a utilização de múltiplos SIA ou outras técnicas de subamostragem, pode ampliar ainda mais sua eficácia e aplicabilidade. Em última análise, essa abordagem representa uma contribuição promissora para o campo da seleção de exemplos majoritários e remoção de redundâncias em conjuntos de dados desbalanceados, com potencial para impulsionar o desenvolvimento de soluções mais eficientes e precisas para problemas de classificação.

## 5. Agradecimentos

Os autores agradecem a FAPEMIG, CAPES e CNPq.

## Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 2623–2631.

Batista, G. E. A. P. A., Carvalho, A. C. P. L. F., and Monard, M. C. (2000). Applying one-sided selection to unbalanced datasets. *MICAI 2000: Advances in Artificial Intelligence*, 1793:315–325.

Bezerra, G. B., Barra, T. V., de Castro, L. N., and Von Zuben, F. J. (2005). Adaptive Radius Immune Algorithm for Data Clustering. In Jacob, C., Pilat, M. L., Bentley, P. J., and Timmis, J. I., editors, *Artificial Immune Systems*, Lecture Notes in Computer Science, p. 290–303, Berlin, Heidelberg. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Dablain, D., Krawczyk, B., and Chawla, N. V. (2022). Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34:6390–6494.

Dakka, M., Nguyen, T., Hall, J., Diakiw, S., VerMilyea, M., Linke, R., Perugini, M., and Perugini, D. (2021). Automated detection of poor-quality data: case studies in healthcare. *Scientific Reports*, 11(1):18005.

de Castro, L. N. and Timmis, J. (2002). Artificial immune systems: A novel approach to pattern recognition. In Corchado, J. M., Alonso, L., and Fyfe, C., editors, *Artificial Neural Networks in Pattern Recognition*, p. 67–84. University of Paisley. URL https://kar.kent.ac.uk/13832/.

De Castro, L. N. and Von Zuben, F. J. (2002). aiNet: an artificial immune network for data analysis. In *Data mining: a heuristic approach*, p. 231–260. IGI Global.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer Cham, 1 edition.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *IEEE Transactions on Neural Networks*, 22(10):1322–1331.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284. ISSN 1558-2191. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Jeyse Freire Pinheiro, E. (2006). Investigação da abordagem de sistemas imunológicos artificiais para reconhecimento de padrões. URL https://repositorio.ufpe.br/handle/123456789/2647. Publisher: Universidade Federal de Pernambuco.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, p. 179.

Lemaitre, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5. URL http://jmlr.org/papers/v18/16-365.html.

Li, J., Liu, L.-s., Fong, S., Wong, R. K., Mohammed, S., Fiaidhi, J., Sung, Y., and Wong, K. K. (2017). Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PloS one*, 12(7):e0180830.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3):160. ISSN 2661-8907. URL https://doi.org/10.1007/s42979-021-00592-x.

Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:769–772.