Aplicando Lógica *Fuzzy* na Interpretação de Grandes Volumes de Dados Cromatográficos no Controle de Dopagem

Guy M. B. Junior[1], Giseli Rabello Lopes[1], Sergio Manuel Serra da Cruz[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro
Rio de Janeiro - RJ – Brasil

guyjunior@iq.ufrj.br, giseli@ic.ufrj.br, serra@ppgi.ufrj.br

Resumo. *Este trabalho apresenta uma proposta computacional para auxiliar na interpretação de grandes volumes de dados baseada na lógica fuzzy e uso do coeficiente de determinação R² aplicados a dados oriundos de cromatógrafos líquidos de alta eficiência acoplados a espectrômetros de massas* (HPLC-MS). *A abordagem apoia analistas na detecção semiautomatizada de substância alvo nas amostras de urina de atletas submetidos ao controle de dopagem. Os primeiros resultados indicam que o método não apenas acelera a detecção, mas também permite a identificação simultânea de múltiplas substâncias alvo.*

1. Introdução

Novas substâncias dopantes e novos métodos de dopagem em atletas são problemas que requerem constante atenção por parte de governos e organizações desportivas. Somente em 2012, os gastos com controle de dopagem aproximaram-se de US$ 500 milhões [Maennig 2014]. Logo, é urgente desenvolver novas estratégias que apoiem as atividades de detecção de *doping,* pois a lista de substâncias e métodos proibidos é atualizada anualmente e a sofisticação da dopagem acompanha, par e passo, a evolução da Farmacologia e da Medicina Desportiva [Aquino Neto 2001].

Embora os avanços científicos propiciem a melhora da detecção de *doping*, há um crescente esforço para desenvolver novas estratégias para evitar a dopagem e manipulação dos resultados esportivos [Maennig 2014]. Nesse cenário, organizações antidopagem, como a Agência Mundial Antidoping (WADA), têm um papel fundamental na luta contra a dopagem [Pereira 2022]. As agências buscam proporcionar maior integridade, garantir condições justas nas competições e preservar a saúde dos atletas.

Mundialmente, são coletadas milhares de amostras de urina para fins de controle de dopagem. O processamento analítico gera *big data* com terabytes de dados brutos a cada nova competição. Esses dados são altamente sensíveis e protegidos, possuindo uma grande variedade de formatos e riqueza de informações. No entanto, seu potencial ainda não é plenamente explorado para uso em pesquisas na Farmacologia, Medicina Desportiva ou mesmo na E-Ciência [Gleaves et al. 2021].

Atualmente, o Laboratório Brasileiro de Controle de Dopagem (LBCD) é o único laboratório acreditado pela WADA na América do Sul. Ele é responsável pela análise de aproximadamente 7.000 amostras/ano, um número que cresce anualmente. Os analistas do LBCD avaliam cada amostra individualmente. Eles inspecionam os dados manualmente através de cromatogramas (sob a forma de relatório ilustrado na Figura 1) gerados por equipamentos científicos do tipo cromatógrafos líquidos de alta eficiência acoplados a espectrômetros de massas (HPLC-MS). Esse procedimento, apesar de robusto e padronizado, é caro e pode ser propenso a erros humanos devido ao grande

volume de informações e ao curto tempo de análises, entre outros [Mogollon et al. 2014].

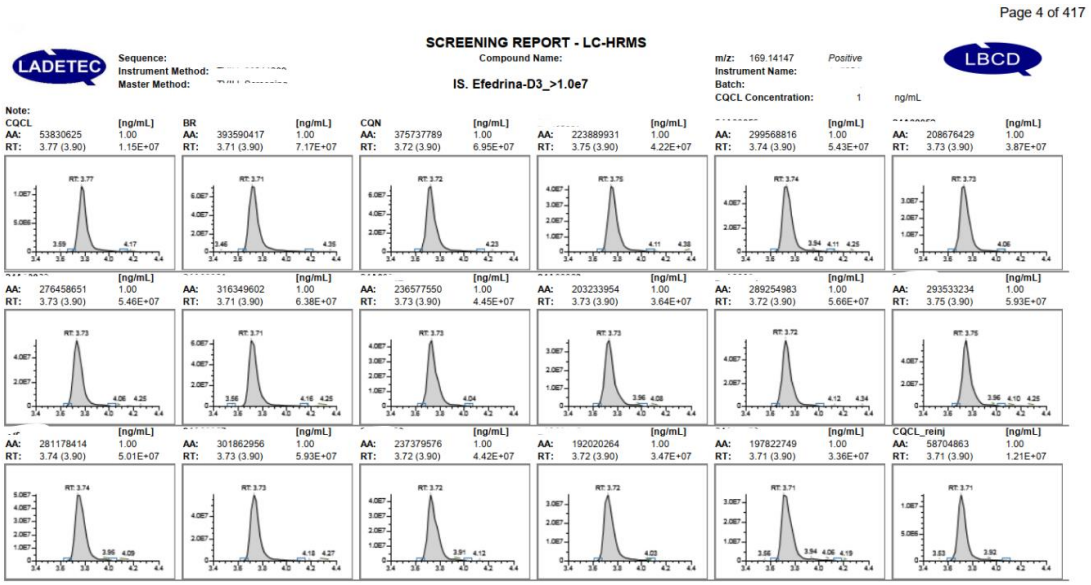

Figura 1. Exemplo de relatório com as gaussianas de cromatogramas (picos cromatográficos) para a substância efedrina-D3 obtidos através de HPLC-MS.

Adicionalmente, para assegurar sua acreditação internacional, anualmente o LBCD passa por rigorosos testes de controle de qualidade. A WADA envia amostras fortificadas com substâncias alvo para que através da cromatografia os analistas e o laboratório as identifiquem e cumpram os requisitos de acreditação. Após a conclusão, o laboratório consolida os resultados reportando-os à WADA, assegurando que todas as análises estejam em conformidade com os padrões internacionais. No entanto, esses testes sobrecarregam ainda mais os analistas.

Logo, com o objetivo de agilizar esses processos, defende-se a criação de um novo método computacional para apoiar os analistas a realizarem análises semiautomatizadas dos cromatogramas. O método proposto é baseado em *pipelines* que usam lógica *fuzzy* [Marro et al. 2011] e ciência de dados através do coeficiente de determinação R^2 [Chiode 2021] para apoiar a interpretação de dados dos cromatogramas. A abordagem visa auxiliar os profissionais no processo de tomada de decisões e reduzir o tempo e as possibilidades de erros humanos.

2. Trabalhos Relacionados

Essa seção aborda trabalhos relacionados à utilização de lógica *fuzzy* e coeficiente de determinação R^2 em análises cromatográficas. No contexto de avaliação não linear, Ukić et al. (2022) demonstram a eficácia da lógica *fuzzy* no apoio à classificação de substância do tipo açúcares usando modelos quantitativos de relação estrutura-retenção (QSRR). Essa abordagem permite a definição de critérios flexíveis para a detecção das substâncias nos cromatogramas, melhorando a precisão da análise quando comparada a métodos booleanos.

O coeficiente de determinação R^2 é uma medida estatística que indica a proporção da variância dos dados de um modelo. Abed e Rasheed (2020) aplicaram o R^2 para avaliar a qualidade do ajuste de modelos de regressão aos dados cromatográficos. Os autores

mostraram que o uso do R^2 é crucial para validar a precisão dos modelos empregados, garantindo resultados confiáveis e compatíveis com os realizados por humanos.

Conforme destacado por Gowd et al. (2018), a integração de lógica *fuzzy* com o coeficiente de determinação R^2 tem se mostrado promissora na identificação de padrões em dados. Essa pesquisa demonstrou que a combinação permite não só uma avaliação robusta, mas também a detecção de *outliers* e outros padrões em longas séries de dados.

Recentemente, Ryoo et al. (2024) têm explorado os algoritmos de aprendizado de máquina na área da antidopagem, visando automatizar e otimizar as análises das amostras. Porém, diferente da nossa proposta, o trabalho utiliza amostras de sangue de atletas do gênero feminino em uma única modalidade esportiva. De qualquer modo, essas inovações afirmam reduzir o tempo de análise e aumentar a precisão dos resultados, representando um avanço significativo na área, no entanto, ainda carecem de estudos mais conclusivos.

A avaliação dos trabalhos relacionados indicou que a combinação de lógica *fuzzy* e coeficiente de determinação R^2 indica ser um método com elevado potencial na avaliação de cromatogramas de HPLC-MS. As pesquisas destacam não apenas a viabilidade dessa combinação, indicam perspectivas para o desenvolvimento de técnicas mais sofisticadas para detecção de substâncias alvo do controle de dopagem ao se incorporar os conceitos de *pipelines* capazes de analisar grandes massas de dados.

3. Materiais e métodos

O LBCD utiliza cromatógrafos líquidos de alta eficiência (HPLC) acoplados a equipamentos de espectrometria de massas de alta resolução (MS) da *Thermo Scientific* para realizar as corridas cromatográficas das amostras de urina fornecidas pelos atletas. Esses equipamentos são essenciais para a obtenção de dados científicos precisos sobre as substâncias presentes nas amostras a serem analisadas. Eles produzem os *datasets* que são utilizados nessa pesquisa, estes consistem de centenas de arquivos com dados textuais estruturados contendo as sequências de injeções das amostras (etapas analíticas validadas pelo LBCD e pelas normas ISO17.025 e ISL2021 da WADA), garantindo maior padronização, auditabilidade, reprodutibilidade e confiabilidade dos resultados.

A sequência de injeção das amostras nos HPLC-MS para a geração dos *datasets* possuem a seguinte lógica: 1) **CQCL** - referente à injeção da amostra que contém as substâncias alvo a serem avaliadas; 2) **CQN** – referente à injeção de solventes que não contém as substâncias alvo; 3) **Amostras de urina**: referente à injeção da amostras de urina de atletas coletadas para controle de dopagem; 4) **CQCL_REINJ**: referente à reinjeção do CQCL, ao final de cada corrida, para comparar, verificar e confirmar (ou não) as ocorrências de substâncias alvo nas injeções das amostras.

As substâncias presentes nas amostras injetadas são separadas através do processo de cromatografia líquida e sua detecção é feita através da espectrometria de massas produzindo os dados brutos (arquivos .RAW). Esses arquivos contêm um número enorme de dados para cada amostra, parâmetros das análises, substâncias alvo, tempos de retenção (RT) e outras variáveis importantes para a interpretação dos resultados cromatográficos.

No LBCD, cada analista é responsável pela interpretação qualitativa dos cromatogramas. Trata-se de um processo lento, caro e sujeito a falhas de interpretação. O tempo de avaliação das amostras é um desafio, especialmente considerando a crescente quantidade de amostras que o laboratório processa anualmente. Por exemplo, cada relatório contém aproximadamente 400 páginas, sendo composto por até 20 amostras

envolvendo mais de 300 diferentes substâncias alvo. Cada arquivo .RAW contém cerca de 15.800 registros e o tamanho de cada arquivo é de aproximadamente 2.0 GB.

Para assegurar a veracidade das análises, a interpretação de cada cromatograma é realizada de forma isolada, distinta por dois analistas distintos. A interpretação das amostras é baseada no CQCL onde o analista compara as amostras de urina dos atletas com o CQCL para identificar sua similaridade e emitir os laudos. A necessidade de uma interpretação rápida, inequívoca e precisa é essencial para manter a acreditação e eficiência do laboratório na interpretação dos resultados e emissão dos laudos.

A interpretação feita pelo analista é manual, pode ter duas conclusões possíveis: (i) *negativo*, a amostra não possui a presença da substância alvo ou caso contenha baixo valor de concentração da substância, ou seja, a concentração da substância presente na amostra pode ser considerada muito baixa; (ii) *presumível*, quando há a presença da substância alvo e cabe ao analista concluir se a concentração da substância na amostra necessita de um novo procedimento de avaliação mais profundo e detalhado para caracterizar como *doping*.

Após a interpretação por pares de analistas, em caso de indicação de negativo por ambos, a amostra segue para a liberação de resultado conclusiva como negativo. Caso ocorra ao menos um presumível, a amostra passa por novo estágio de avaliação, onde a substância alvo será evidenciada e destacada em novos métodos analíticos de maior precisão.

Propomos uma nova metodologia baseada na teoria dos *workflows* científicos para experimentos em larga escala [Mattoso et al, 2010], implementados sob a forma de *pipelines* para o processamento de grandes massas de dados oriundos das amostras (Figura 2). O *pipeline* fornece uma solução computacional de apoio aos analistas e auxiliando-os no processo de tomada de decisão possivelmente reduzindo o tempo de avaliação e falhas humanas.

Figura 2. Etapas do *pipeline* para processamento das amostras.

As etapas do *pipeline* conceitual: (1) Injeção das amostras e configuração dos equipamentos HPLC-MS da *Thermo Scientific*, (2) Geração dos arquivos .RAW, extração dos mesmos parâmetros utilizados pelos analistas para avaliar o tempo de retenção e concentração das substâncias presentes nas amostras, (3) Tratamento (engenharia) dos dados de espectrometria de massas com arquivos .RAW, envolvendo ações de estruturação, limpeza e tabulação dos dados dos cromatogramas. A seguir, (4) calcular os coeficientes de determinação R^2 sobre os *datasets* com o objetivo de efetuar um comparativo assemelhado ao realizado pelo analista humano. O método computacional proposto calculará o R^2 do CQCL e o R^2 da amostra do atleta. Após essa extração, serão calculadas as métricas de indicação de presença ou ausência de uma determinada substância.

Por fim, definição do filtro de detecção (5), utiliza-se a lógica *fuzzy* para a interpretação qualitativa dos resultados do R^2; as definições não booleanas serão fundamentais para equiparar o nível de interpretação automatizada com análises manuais

realizadas pelos analistas. O uso da lógica *fuzzy* e suas funções de pertinência apoiam o analista na aceleração da interpretação qualitativa dos resultados.

4. Implementação do *pipeline*

O *pipeline*, descrito na seção 3, foi materializado utilizando a linguagem de programação *Python* e suas bibliotecas, ele é capaz de realizar todos as computações relativas aos procedimentos analíticos. A primeira tarefa do *pipeline* é a carga dos *datasets* (dados do HPLC-MS com os arquivo .RAW). Em seguida, efetua a verificação e tratamento dos dados através da biblioteca *pymsfilereader*, desenvolvida por François [2019], onde a função *GetChroData* acessa os dados .RAW e extrai os parâmetros de corrida cromatográfica (*StartTime, EndTime, MassRange, ScanFilter, SmoothingType* e *SmothingValue)*, conforme a Tabela 1.

Tabela 1. Parâmetros parâmetros de corrida cromatográfica utilizados na biblioteca *pymsfilereader*

GetChroData	Função
start_time *end_time*	Os parâmetros definem o intervalo de tempo para a extração dos dados cromatográficos, permitem focar em uma janela de tempo específica durante a análise, facilitando a identificação de picos relevantes.
mass_range	O parâmetro especifica o intervalo de massa considerado durante a extração dos dados. Filtrar as massas de interesse, ignorando as massas fora do intervalo definido.
scan_filter	Filtro de varredura usado para selecionar tipos de escaneamento durante a extração dos dados. Pode incluir filtros de modos de ionização, tipos de fragmentação ou outras características específicas do experimento.
smoothingType e *smoothingValue*	Os parâmetros são utilizados para aplicar uma suavização aos dados cromatográficos. O *smoothingType* define o método de suavização, o *smoothingValue* determina o grau de suavização aplicado. A suavização reduz o ruído nos dados, facilita a identificação dos picos mais significativos.

Os parâmetros são cruciais para a análise dos dados que utiliza o tempo de retenção (RT) na coluna cromatográfica através do *StartTime* e *EndTime*, a detecção da substância alvo será calculada através do *MassRange* e *ScanFilter*. Para, a seguir, organizar as gaussianas (picos dos cromatogramas) e agilizar a interpretação visual pelos analistas. Os parâmetros de *Smoothing* são responsáveis por ajustar os cromatogramas para visualização. Os dados são estruturados em *DataFrames* (biblioteca *pandas*) que representam os tempos de retenção e as intensidades de íons de cada massa molecular das substâncias em análise.

Para comparar os picos dos cromatogramas, utilizamos a função *r2_score* (da biblioteca *scikit-learn*) para calcular o R^2 que avalia a qualidade do modelo de regressão. Para determinar o R^2, utilizamos um conjunto de valores reais (*y_true*) extraídos dos arquivos de CQCL e um conjunto de valores preditos pelo modelo (*y_pred*), que são valores extraídos das amostras de urina dos atletas. A lógica *fuzzy* trata as incertezas e imprecisões presentes nos dados, ela é implementada através do módulo *skfuzzy* (da biblioteca *scikit-fuzzy*).

O *skfuzzy* oferece funcionalidades para criar sistemas de inferência *fuzzy*, que podem ser aplicados em diversas áreas, desde processos de tomada de decisão até o reconhecimento de padrões. O *pipeline* também utiliza funções de pertinência para o coeficiente, classificação, regras e variáveis de entrada e saída, com o objetivo de classificar cada amostra como sendo: *Negativo, Presumível Muito Baixo, Presumível Baixo, Presumível Médio, Presumível Alto* ou *Presumível Muito Alto*. Utilizamos a biblioteca *matplotlib.pyplot* para a visualização gráfica dos resultados, gerando gráficos complementares que auxiliam os analistas na interpretação dos cromatogramas e dos resultados das análises *fuzzy*.

Os experimentos envolveram a utilização integrada das bibliotecas *pymsfilereader*, *pandas*, *matplotlib.pyplot*, *sklearn.metrics*, *numpy* e *skfuzzy*, proporcionando uma abordagem capaz de analisar grandes quantitativos de amostras e compará-las com os controles positivos, culminando na classificação das amostras com base na lógica *fuzzy*. O código fonte dos *pipelines* está disponível em https://github.com/guyjunior/dopinho.

5. Resultados e discussões

Os experimentos computacionais buscaram 125 substâncias alvo presentes em 300 amostras cegas, 300 amostras de rotina e 20 amostras de CQCL. Foram realizadas várias rodadas de testes e os indicadores de acurácia, precisão e F1-score estão representados na Tabela 2. A 1ª rodada utiliza os dados com substâncias alvo, comparando a similaridade do CQCL com amostras fortificadas com todos as substâncias alvo. A 2ª rodada utiliza os dados cegos das amostras de rotina, comparando a similaridade do CQCL com amostras de rotina que os analistas interpretam. A 3ª rodada contém os dados com substâncias alvo e amostras cegas, comparando a similaridade do CQCL com amostras fortificadas e negativas, misturadas de modo proposital para verificação pelo *pipeline*.

Tabela 2. Indicadores das rodadas dos experimentos computacionais

	Acurácia	**Precisão**	**F1-Score**
1ª Rodada	98%	98%	99%
2ª Rodada	98%	98%	99%
3ª Rodada	86%	86%	93%

Os experimentos avaliaram mais de 600.000 dados das amostras (cada uma com 125 substâncias alvo com 8 parâmetros) relativos aos conjuntos de amostras de urina do controle de dopagem submetidas ao HPLC-MS. As Figuras 3A e 3B, ilustram a presença de quatro diferentes substâncias alvo presentes nas amostras #45 e #31; o aumento da intensidade das substâncias alvo nas amostras (curva em vermelho) elevam o grau de similaridade com o CQCL (curva em azul). Verifica-se que o *pipeline* classificou corretamente a presença das substâncias de acordo com a similaridade dos cromatogramas, utilizando o valor do R^2 e as inferências da lógica *fuzzy*. Ambas as detecções classificaram as amostras como baixo e médio.

As Figuras 3A, 3B e 3C ilustram graus de intensidade distintos independentemente das substâncias; o *pipeline* foi capaz de interpretar a intensidade para diferentes substâncias. A Figura 3D ilustra com clareza a similaridade; o *pipeline* classificou-a como um presumível "muito alto".

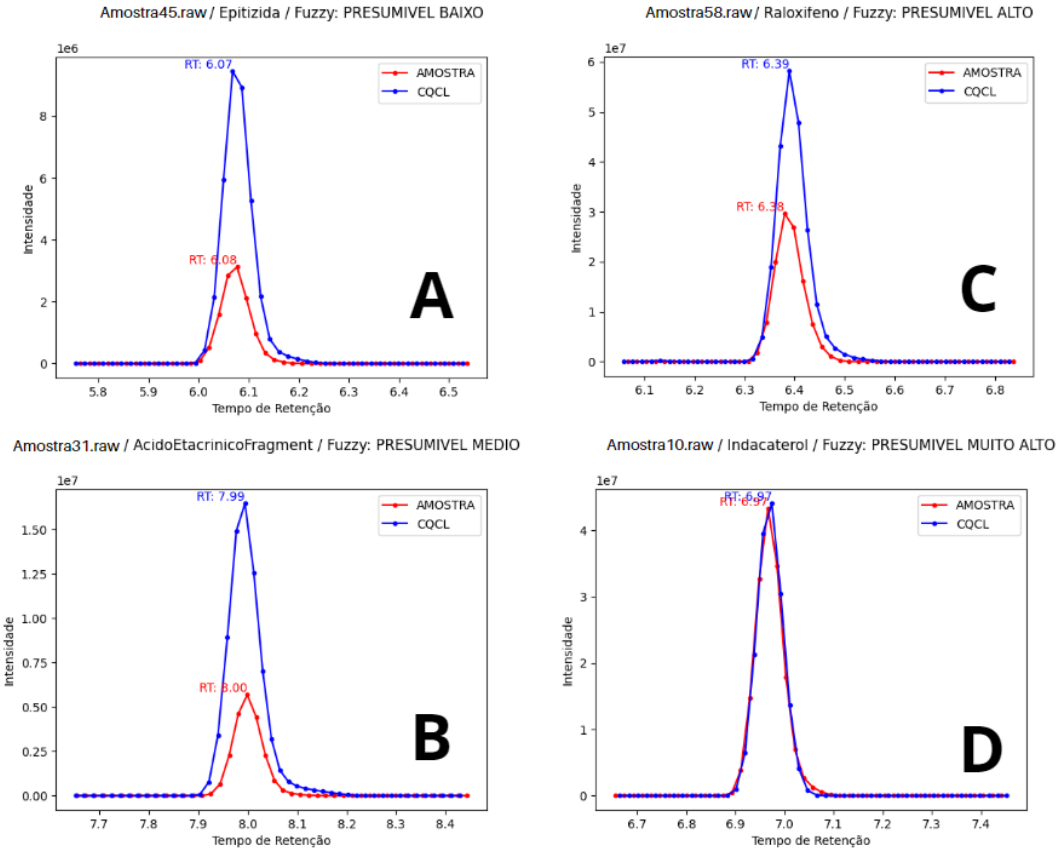

Figura 3. Amostras classificadas como: A = "presumível baixo", B = "presumível médio", C = "presumível alto" e D = "presumível muito alto" para substâncias diferentes.

Os experimentos computacionais evidenciaram uma nova forma comparativa de classificação e visualização entre as amostras com substâncias e os controles positivos ou fortificados (CQCL). Essa nova forma facilita a percepção visual das intensidades das substâncias em cada amostra pelos analistas.

O *pipeline* foi capaz de analisar as 125 substâncias presentes nas amostras, onde para cada substância identificada como *presumível*, independentemente da intensidade (muito baixa, baixa, média, alta ou muito alta), a lógica *fuzzy* sinaliza ao analista uma possível presença da substância alvo.

6. Conclusão

O controle de *doping* em atletas é um problema em aberto e de natureza interdisciplinar. Nossa contribuição, experimentos e seus resultados ainda são preliminares, no entanto, indicam que o método proposto tem indicativos de eficiência e precisão; atualmente ele está sendo avaliado de forma experimental na rotina dos analistas do LBCD.

A implementação computacional do método permitiu uma interpretação mais rápida dos dados cromatográficos quando comparada aos métodos tradicionais, evidenciando uma significativa redução no tempo de interpretação dos resultados sem perda de correção e um possível aumento na confiabilidade dos resultados.

A capacidade do *pipeline* em processar grandes volumes de amostras em um curto período aliada à precisão na identificação de substâncias alvo está se mostrando plausível para o ambiente do controle de dopagem. A visualização facilitada de dados e a utilização

da lógica *fuzzy* para a interpretação dos resultados poderá apoiar os analistas no processo de tomada de decisões mais assertivas e seguras. A flexibilidade e a eficácia do método proposto indicam seu potencial para ser utilizado em outras áreas que requerem interpretação de dados complexos provenientes do HPLC-MS.

Como trabalhos futuros, pretende-se ampliar o número de amostras, agregar novos métodos de aprendizado de máquina o *pipeline* para classificação das amostras e oferecer interfaces gráficas mais intuitivas para os analistas. Além disso, pretende-se oferecer a solução sob a licença de código livre para outros laboratórios de controle de dopagem ou setores que necessitem desse tipo de suporte computacional.

Referências

Abed, S. S., & Rasheed, A. S. (2020). Estimação simultânea de clonazepam e metronidazol em comprimidos farmacêuticos pelo modo de cromatografia líquida de alta eficiência de fase reversa com detecção uv. *Periódico Tchê Química*, *17*(36).

Aquino Neto, F. R. D. (2001). O papel do atleta na sociedade e o controle de dopagem no esporte. *Revista Brasileira de Medicina do Esporte*, *7*, 138-148.

Chiode, A. D. S. (2021). *Avaliação de propostas de coeficientes de determinação do tipo R^2 em modelos de regressão logística com resposta nominal* (Doctoral dissertation, Instituto de Matemática e Estatística, Universidade de São Paulo).

François, A. (2019). *pymsfilereader: Thermo MSFileReader Python bindings*. GitHub. https://github.com/frallain/pymsfilereader

Gleaves, J., et al. (2021). Doping prevalence in competitive sport: evidence synthesis with "best practice" recommendations and reporting guidelines from the WADA Working Group on Doping Prevalence. *Sports Medicine*, *51*(9), 1909-1934.

Gowd, B. P., Jayasree, K., & Hegde, M. N. (2018). Comparison of artificial neural networks and fuzzy logic approaches for crack detection in a beam like structure. *Int. J. Artif. Intell. Appl*, *9*(1), 35-51.

Maennig, W. (2014). Inefficiency of the anti-doping system: Cost reduction proposals. *Substance use & misuse*, *49*(9), 1201-1205.

Mattoso et al. (2010). Towards supporting the life cycle of large scale scientific experiments. *Int. J. of Business Process Intehration and Management*. 5(1), 79-92.

Marro, A. A., et al. (2010). Lógica fuzzy: conceitos e aplicações. *Natal: Universidade Federal do Rio Grande do Norte (UFRN)*, 2.

Mogollon, N. G., et al. (2014). State of the art two-dimensional liquid chromatography: fundamental concepts, instrumentation, and applications. *Química Nova*, *37*, 1680-1691.

Pereira, S. L. R. (2022). *Desenvolvimento e validação de um método de detecção e quantificação de THC-COOH, em urina, por cromatografia gasosa acoplada a espectrometria de massa* (Desenvolvimento e validação de um método de detecção e quantificação de THC-COOH, em urina, por cromatografia gasosa acoplada a espectrometria de massa, Universidade de São Paulo).

Ryoo, H., et al. (2024). Identification of doping suspicions through artificial intelligence-powered analysis on athlete's performance passport in female weightlifting. *Frontiers in Physiology*, *15*, 1344340

Ukić, Š., et al. (2015). Development of gradient retention model in ion chromatography. Part III: Fuzzy logic QSRR approach. *Chromatographia*, *78*, 889-898.Dias, M. D. S., Visintin, L., & Reiser, R. Estudo Introdutório da Lógica Fuzzy Intuicionista Intervalar.