Identificação de metabólitos no plasma e urina com potencial uso como biomarcadores para o monitoramento dos efeitos adversos dos AINEs em gatos via aprendizado de máquina

Cecilia Aparecida Santos Silva
Departamento de automática
Universidade Federal de Lavras
Lavras, Brasil
cecilia.silva1@estudante.ufla.br

Danton Diego Ferreira
Departamento de automática
Universidade Federal de Lavras
Lavras, Brasil
danton@ufla.br

Marcos Ferrante
Departamento de medicina veterinária
Universidade Federal de Lavras
Lavras, Brasil
marcos.ferrante@ufla.br

Nicolas Villarino
College of Veterinary Medicine
Washington State University
Washington, Estados Unidos
nicolas.villarino@wsu.edu

***Resumo*—Os anti-inflamatórios não esteróides (AINEs), como o meloxicam, podem ser usados para melhorar a dor crônica e, ao mesmo tempo, ajudar a reduzir o uso de opióides. No entanto, a administração crônica desses medicamentos pode causar danos gastrointestinais e renais em alguns pacientes, limitando as opções terapêuticas desses pacientes que necessitam de controle da dor e da inflamação. A fim de discriminar animais que receberam AINEs de animais saudáveis (controle) e determinar em qual estágio estão os efetivos dos AINEs no organismo do paciente, foram desenvolvidos 2 modelos, um k-nearest neighbors algorithm (KNN) e um Gaussian Naive Bayes algorithm (NB). A base de dados usada é composta por um grupo de 6 gatos saudáveis e um grupo de 6 gatos que foram tratados com meloxicam. Foram coletadas informações metabolômica plasmática e urinária 5 vezes durante um período de até 31 dias em ambos os grupos. A partir disso, separou-se os dados em três bases: (i) base de substâncias hidrossolúveis, composta por 114 substâncias e 6 amostras, (ii) base de substâncias lipossolúveis com 195 substâncias e 5 amostras, e (iii) base composta por todas as substâncias analisadas com 5 amostras. Além disso, cada base de dados possui 5 classes, uma para cada vez em que as substâncias foram coletadas e medidas. Foi utilizado o Fisher 's Discriminant Ratio (FDR), com a finalidade de ranquear as substâncias que mais se alteraram de uma classe para a outra. Apesar da baixa quantidade de amostras e da necessidade de se usar a menor quantidade de substâncias possíveis, resultados satisfatórios foram alcançados.**

Palavras-chave—Meloxicam, Fisher's Discriminant Ratio, Machine Learning, K-nearest neighbors, Gaussian Naive Bayes.

I. INTRODUÇÃO

Os anti-inflamatórios não esteróides (AINEs), como o meloxicam, podem ser usados para melhorar a dor crônica e, ao mesmo tempo, ajudar a reduzir o uso de opióides. No entanto, a administração constante desses medicamentos pode causar danos gastrointestinais e renais em alguns pacientes, limitando as opções terapêuticas daqueles que necessitam de controle da dor e da inflamação.

Expandir o entendimento sobre os efeitos dos AINEs no corpo pode levar a revelar os mecanismos associados com intolerância a esse tipo de medicamento, o que permitirá descobrir novos alvos de drogas e/ou estratégias terapêuticas para o controle ideal da dor e da inflamação [1]. Além disso, a identificação de alterações nos lipídios induzidas por medicamentos pode resultar na descoberta de novas substâncias que podem se tornar marcadores candidatos para monitorar o uso de AINEs em pacientes. Isso é interessante pois marcadores metabólicos podem auxiliar, após validação adequada, a categorização dos pacientes como tendo um risco alto ou baixo de sofrer efeitos adversos [2], o que ajudaria a estabelecer novas estratégias de tratamento direcionadas, melhorando assim o atendimento ao paciente.

Os AINEs têm um efeito pleiotrópico no corpo, influenciando a função celular e o metabolismo na maioria dos tecidos [3-6]. Os AINEs podem alterar o metabolismo lipídico; de fato, um de seus principais mecanismos de ação está relacionado às vias lipídicas intracelulares [4]. Vários estudos relataram, inclusive, que os AINEs induzem a dessaturação de ácidos graxos celulares, alteram os componentes fosfolipídicos celulares e elevam o nível de ácidos graxos livres [4].

Considerando todas essas evidências, hipotetiza-se que a administração repetida do AINE meloxicam altera o conteúdo lipídico plasmático e urinário em gatos. Porque os lipídios são moléculas diversas e abundantes [7,8], essa hipótese foi abordada traçando um perfil do repertório total de metabólitos no plasma e na urina usando uma abordagem metabolômica não direcionada em gatos tratados repetidamente com meloxicam. O estudo de lipídios no plasma e na urina fornece um poderoso meio não invasivo para avaliar o status dos processos metabólicos celulares que ocorrem nos tecidos do corpo. A aplicação da lipidômica em várias doenças incluindo diabetes [9,10], doenças cardiovasculares [11] e outros processos inflamatórios [12] já forneceu assinaturas lipídicas características e insights mecânicos sobre os processos de doenças [13,14]. O perfil lipídico resultante é análogo a uma impressão digital química deixada para trás por alterações induzidas por AINEs nos processos celulares. Portanto, necessita-se identificar metabólitos no plasma e na urina que possam ser avaliados prospectivamente como biomarcadores para monitorar o efeito de AINEs em gatos.

Os modelos de *machine learning*, por sua vez, vêm sendo muito usados no auxílio à tomada de decisão em inúmeros momentos da atenção à saúde, especialmente no diagnóstico, intervenção e acompanhamento de problemas de saúde, tanto humana quanto animal, como foi feito no trabalho de desenvolvimento de algoritmos de aprendizado de máquina para estimar limites máximos de resíduos de medicamentos veterinários [15]. Isso deve-se a capacidade de identificar indícios que determinam o ocasionamento de determinadas observações de maneira muito mais eficiente que um humano.

II. Técnicas Empregadas

No presente trabalho, dois algoritmos de *machine learning* foram utilizados para classificar os estágios da doença renal gerada pelo tratamento com meloxicam. Optou-se pelo k-nearest neighbors (KNN) e Gaussian Naive Bayes (NB) por serem bons para problemas com poucas amostras, situação esta encontrada neste trabalho.

A. K-nearest Neighbors

A classificação KNN é um dos métodos de classificação mais fundamentais e simples. Quando há pouco ou nenhum conhecimento prévio sobre a distribuição dos dados, o método KNN deve ser uma das primeiras escolhas para classificação. É um poderoso não-paramétrico sistema de classificação que contorna completamente o problema das densidades de probabilidade, quando estimativas paramétricas confiáveis de densidades de probabilidade são desconhecidas ou difíceis de serem determinar. Ele realiza a classificação de um dado desconhecido por meio das suas "*k*" distâncias relativas mais próximas a um conjunto de treino, baseando-se na ideia de que seus padrões mais próximos de um padrão alvo x, fornecem informações úteis do rótulo [16]. Para isso, temos que ser capazes de definir uma medida de similaridade no espaço de dados. É comumente baseado na distância euclidiana entre uma amostra de teste e as amostras de treinamento especificadas, mas também pode-se utilizar a distância de Manhattan ou a distância de Minkowski.

B. Gaussian Naive Bayes

Já o NB, conhecido por sua simplicidade e eficácia, apresentando benefícios significativos, uma vez que requer poucos parâmetros para estimativa em comparação com seus concorrentes para classificação, aprende as probabilidades de um objeto com determinadas características pertencentes a uma determinada classe ou grupo, ou seja, é um modelo preditivo probabilístico. Esse tipo de modelo recebe o nome de "ingênuo" pois desconsidera a correlação entre as *features*, pressupondo que elas são independentes, e verifica se uma amostra analisada pertence ou não a uma determinada classe ao encontrar a probabilidade a Posteriori, através do teorema de Bayes [17]. Para isso ele calcula a probabilidade de uma dada amostra pertencer a cada uma das classes e a maior probabilidade determinará a classe da amostra. Foi usada uma adaptação para dados numéricos chamada algoritmo Gaussiano de Naive Bayes, que assume que a variável analisada possui uma distribuição Gaussiana ou normal. Com a distribuição Gaussiana basta calcular a média e o desvio padrão de cada variável numérica para cada classe [18].

C. Fisher's discriminant ratio

Coletar substâncias para análise é diretamente proporcional ao custo para isso, uma vez que a eficiência dos equipamentos necessários também precisa aumentar. Dessa forma, do ponto de vista prático de se automatizar uma análise de amostra, é importante diminuir a quantidade de substâncias usadas para formar as amostras que compõem a base de dados, e obter ganhos relativos à praticidade e custo. Por isso, empregou-se a técnica *Fisher's discriminant ratio* (FDR), uma das mais eficientes abordagens para redução de dimensionalidade, para selecionar aquelas variáveis (substâncias) que mais contribuem para determinar o estágio da doença renal [19] baseado no cálculo da discrepância entre os valores coletados de uma mesma substância em estágios diferentes. A FDR calcula a divisão do quadrado da diferença entre as médias de dois tipos de pontos de amostra, uma para cada classe, pela soma de suas dispersões dentro da sua respectiva classe, a fim de encontrar a linha na qual a projeção dos dados maximiza a separação entre as classes [20].

A equação 1 mostra como se deu o cálculo das discrepâncias, onde J representa o vetor de discrepâncias, m1 e m2 são os vetores de media e $d1^2$ e $d2^2$ são os vetores de variância das amostras dos estágios analisados no momento.

$$J = (m1 - m2)^2 * \frac{1}{d1^2 + d2^2} \quad (1)$$

III. Base de dados

A base de dados usada para esse trabalho veio do estudo de metabolômica plasmática e urinária em gatos com administração repetida de meloxicam realizado pelo grupo de pesquisa coordenado pelo Dr. Nicolas F. Villarino da Washington State University [21]. Importante ressaltar que todos os métodos foram realizados de acordo com as diretrizes e regulamentos relevantes. Resumidamente, gatos (n = 6) foram tratados por via subcutânea com solução salina ou com 0,3 mg/kg de peso corporal de meloxicam diariamente por até 31 dias. Os metabolitos no plasma e na urina foram determinados por *Liquid chromatography–mass spectrometry* (LC-MS) antes do primeiro tratamento e aos 4, 9 e 13 e 17 dias após a primeira administração de meloxicam. Importante ressaltar que esses períodos serão tratados como Dn, onde n indica a sequência do período. Por exemplo, D0 representa a primeira coleta de metabolitos. A partir disso, separou-se os dados em uma base de substâncias hidrossolúveis composta por 114 substâncias e 6 amostras, outra para substâncias lipossolúveis com 195 substâncias e 5 amostras, isso por que um dos gatos não apresentou mudanças nos valores das substâncias analisadas, e uma última base de dados composta por todas as substâncias analisadas com 5 amostras. Além disso, cada uma das bases tinha 5 classes, uma para cada vez que as substâncias foram coletadas e medidas. Como consequência do custo para coletar cada substância, aplicou-se nos dados a FDR, com a finalidade de ranquear as substâncias que mais se alteraram de uma classe para a outra. Ademais, as mesmas substâncias, de cada uma das três bases de dados criadas foram coletadas em 6 gatos saudáveis e que não tiveram contato com meloxicam, chamados de gatos de controle, durante os mesmos 31 dias com o propósito de coletar informações que possibilitam analisar como essas substâncias se comportam naturalmente.

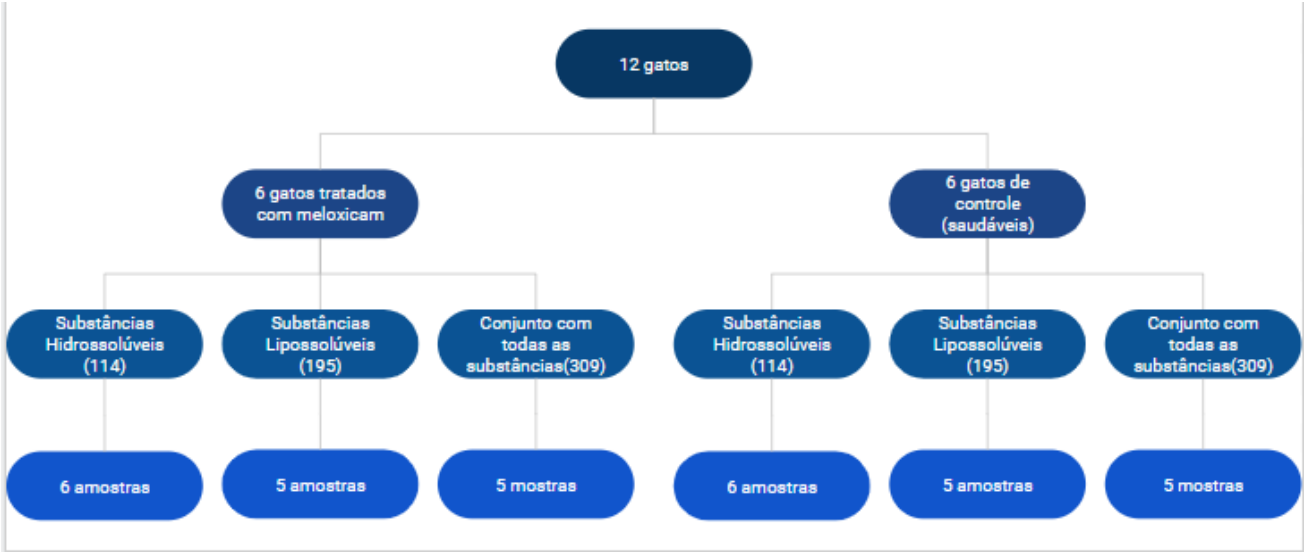

Fig. 1 Diagrama da divisão dos dados coletados em bases de dados.

Na Fig. 1 há um diagrama que mostra como é a divisão dos dados analisados. Na Fig. 2 há um gráfico que exemplifica as discrepâncias para a base de dados das substâncias hidrossolúveis do estágio D0 e D1 dos gatos doentes. Nota-se o tamanho do vetor J, indicado no eixo X, é igual ao número de substâncias hidrossolúveis. O eixo Y, por sua vez, indica a discrepância.

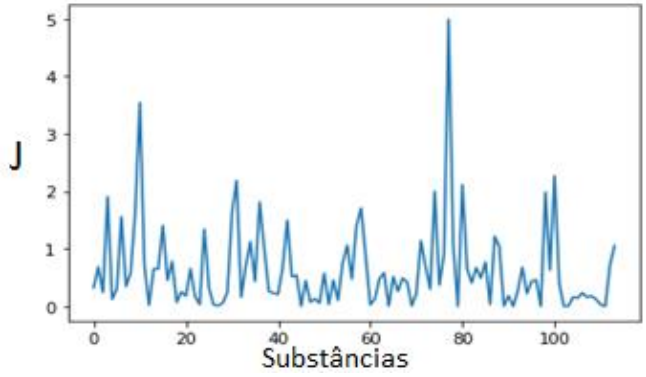

Fig. 2 Exemplo de gráfico das discrepâncias das substâncias hidrossolúveis.

IV. Método Proposto

A Fig. 3 mostra como se deu o método proposto. O primeiro passo, após a organização dos dados, foi aplicar o FDR nas bases de dados. Esse processo foi realizado para calcular a discrepância de cada uma das substâncias que compõem as amostras de dois a dois estágios. Assim, ao todo, obteve-se 10 resultados do FDR para cada base de dados. A Fig. 4 ilustra como se deu essa combinação de estágios para aplicar o FDR.

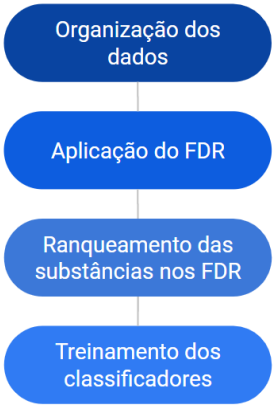

Fig. 3 Diagrama em blocos do método proposto.

Fig. 4 Diagrama de combinação de estágios para aplicar o FDR.

A fim de selecionar quais substâncias serão levadas em consideração na análise, ordenou-se decrescentemente os FDRs e escolheram-se as 5 substâncias que mais alteraram de um estágio para o outro, totalizando 50. Nas três bases, foram feitas análises com o vetor de amostras composto por todas elas, excluindo repetições, com o vetor feito apenas com as substâncias que se repetem e com o vetor feito apenas por aquelas que não se repetiram. Somado a isso, também reduziu-se a quantidade de substâncias nesses vetores para os menores números possíveis que mantivessem uma acurácia satisfatória dos classificadores.

Dessa maneira, chegou-se a 9 substâncias do conjunto das hidrossolúveis, 5 para as lipossolúveis e 6 para o conjunto que uniu as substâncias hidrossolúveis com as lipossolúveis, o

que levou a uma redução de dimensão considerável de 114 para 9.

Como só era possível criar no máximo 6 amostras para cada estágio em cada uma das três bases de dados, algumas táticas foram usadas para avaliar se um número maior de amostras faria uma diferença significativa nos resultados. A primeira delas foi unir os estágios 1 e 2 e os estágios 3 e 4, além de usar as amostras do estágio 0 dos gatos do grupo de controle, uma vez que foram medidos junto com os gatos que foram aplicados à meloxicam.

Já a segunda tática foi usar a ferramenta de sobreamostragem conhecida como *Synthetic Minority Oversampling* (SMOTE). Ela usa amostras originais para criar amostras sintéticas da classe minoritária com o objetivo de balancear a base de dados [22] e faz isso de maneira semelhante ao funcionamento do KNN, ou seja, a cada iteração, uma nova amostra da classe com menos amostras é selecionada aleatoriamente e os seus K vizinhos mais próximos são encontrados. Um desses é escolhido formando uma nova amostra que combina os dois vetores de substâncias. Para evitar o problema de *over-fitting* enquanto expande regiões de classe minoritária, o SMOTE gera novas instâncias operando dentro do espaço de recursos existente. Os valores de nova instância são derivados de interpolação em vez de extrapolação, portanto, eles ainda carregam relevância para o conjunto de dados subjacente.

Para cada dado minoritário, uma nova instância de dados sintéticos (I) é gerada tomando a diferença entre o vetor de características de I e seu vizinho mais próximo (J) pertencente à mesma classe, multiplicando-o por um número aleatório entre 0 e 1 e, em seguida, adicionando-o a I. Isso cria um segmento de linha aleatório entre cada par de recursos existentes das instâncias I e J, resultando na criação de uma nova instância dentro do conjunto de dados [23]. Esse processo se repete para os outros vizinhos k-1 da instância minoritária I. Como resultado, o SMOTE gera regiões mais gerais da classe minoritária.

O modelo de KNN usado para a base lipossolúvel, foi configurado para encontrar os 2 vizinhos mais próximos e p=2, sendo p o parâmetro indica a medida de distância utilizada, que nesse caso é a distância euclidiana. Já a base hidrossolúvel e por aquela composta por essas duas apresentou melhor resultado com o NB.

V. RESULTADOS E DISCUSSÃO

Ao todo, três KNNs e três NBs foram analisados, um para cada grupo de substâncias analisadas (Hidrossolúveis, Lipossolúveis e Todas).

A. *Grupo Hidrossolúvel*

Para o grupo Hidrossolúvel reduziu-se o número de substâncias selecionadas, sendo elas: phenol, 3-(4-hydroxyphenyl) propionic acid, quinic acid, uridine, isoribose, oxoproline, galactonic acid, Xylose, tyrosine. Importante ressaltar que essas substâncias são aquelas que se repetem em mais de um FDR. O SMOTE foi aplicado a partir da base de dados com substâncias selecionadas via FDR. Para criar o balanceamento necessário para o funcionamento da análise, também foram usadas as amostras do estágio 0, que corresponde os valores coletados antes do primeiro contato com meloxicam, dos gatos não afetados pelo meloxicam para aumentar a classe referente ao estágio 0 dos gatos afetados, já que nesse estágio nenhum dos 12 gatos teve contato com o AINE. A Tabela I apresenta a matriz de confusão do algoritmo gaussiano de Naive Bayes que apresentou uma acurácia de 93% na identificação dos estágios, sendo o melhor resultado. Percebe-se que o modelo classificou uma amostra do estágio 2 como estagio 3, o que é justificável, uma vez que são estágios vizinhos e há semelhança entre os dados de suas substâncias. Para essa mesma situação, o KNN, apresentou acurácia de 72,22%. Pode-se ver a matriz de confusão desse resultado na Tabela II.

TABELA I. Matriz de confusão do melhor resultado encontrado para a base hidrossolúvel.

		Estágio predito				
		Estágio 0	Estágio 1	Estágio 2	Estágio 3	Estágio 4
Estágio verdadeiro	Estágio 0	5				
	Estágio 1		2			
	Estágio 2			2	1	
	Estágio 3				1	
	Estágio 4					4

TABELA II. Matriz de confusão do KNN para a base hidrossolúvel.

		Estágio predito				
		Estágio 0	Estágio 1	Estágio 2	Estágio 3	Estágio 4
Estágio verdadeiro	Estágio 0	2				1
	Estágio 1		2			
	Estágio 2		1	4		
	Estágio 3			2	2	
	Estágio 4		1			3

B. *Grupo Lipossolúvel*

As amostras da base de dados Lipossolúvel foram compostas pelas substâncias LPC (17:1), LPC (20:0), CE (16:1), Gal-Gal-Cer(d18:1/16:0) or Lactosylceramide (d18:1/16:0) e PC (35:4), totalizando 6, selecionadas dentre aquelas que repetiram nos FDRs. O melhor resultado encontrado usou as amostras do estágio D0 dos gatos não afetados pelo meloxicam para aumentar a classe referente ao estágio D0 dos gatos afetados apenas na etapa de teste. O algoritmo KNN apresentou uma acurácia de 64%, e sua matriz de confusão está presente na Tabela III Para essa situação e as demais a maior acurácia encontrada pelo NB foi 60%.

TABELA III. Matriz de confusão do melhor resultado encontrado para a base lipossolúvel.

		Estágio predito				
		Estágio 0	Estágio 1	Estágio 2	Estágio 3	Estágio 4
Estágio verdadeiro	Estágio 0	5		2	1	
	Estágio 1		1			
	Estágio 2					
	Estágio 3				1	
	Estágio 4		1			

C. *Grupo Hidrossolúveis e Lipossolúveis*

O conjunto com substâncias Hidrossolúveis e Lipossolúveis, foi composto pela LPC (17:1), adenosine , PC (p-36:3) or PC (o-36:4), LPC (22:6), LPC (20:1) e tyrosine, selecionadas dentre as primeiras substâncias que mais alteraram de um FDR para o outro. O algoritmo gaussiano de Naive Bayes apresentou uma acurácia de 80%. Na Tabela IV é possível ver

a matriz de confusão desse resultado. O único resultado maior que 50% encontrado pelo KNN foi de 60%.

TABELA IV. Matriz de confusão do melhor resultado encontrado para a base que une as substâncias hidro e lipossolúveis.

		Estágio predito				
		Estágio 0	Estágio 1	Estágio 2	Estágio 3	Estágio 4
Estágio verdadeiro	Estágio 0				1	
	Estágio 1		1			
	Estágio 2			1		
	Estágio 3				1	
	Estágio 4					1

Numa tentativa de aumentar a quantidade de amostras por classe, agrupou-se a classe 1 com a classe 2 e a classe 3 com a classe 4. Para promover o balanceamento, a classe 0 dos dados de controle foi acrescentada aos dados da classe 0 dos gatos doentes. Além disso, as substâncias usadas foram as mesmas dos testes anteriores. Essa análise com 3 classes obteve acurácias para as substâncias hidrossolúveis de 63,63% com o KNN e 72,72% com o NB. A matriz de confusão do melhor resultado pode ser vista na Tabela V.Para as substâncias lipossolúveis o NB encontrou uma acurácia de 100%. Por último, o NB obteve uma acurácia de 80% para a base de dados composta por todas as substâncias. A Tabela VI e a Tabela VII mostram, respectivamente, as matrizes de confusão dos dois últimos resultados apresentados.

TABELA V. Matriz de confusão do algoritmo NB para a base hidrossolúvel.

		Estágio predito		
		Estágio 0	Estágios 1 e 2	Estágios 3 e 4
Estágio verdadeiro	Estágio 0	4		
	Estágios 1 e 2		2	2
	Estágios 3 e 4		1	2

TABELA VI. Matriz de confusão do melhor resultado encontrado par a base de substâncias lipossolúveis com NB.

		Estágio predito		
		Estágio 0	Estágios 1 e 2	Estágios 3 e 4
Estágio verdadeiro	Estágio 0	4		
	Estágios 1 e 2		1	
	Estágios 3 e 4			2

TABELA VII. Matriz de confusão do melhor resultado encontrado para a base que une as substâncias hidro e lipossolúveis.

		Estágio predito		
		Estágio 0	Estágios 1 e 2	Estágios 3 e 4
Estágio verdadeiro	Estágio 0	2	1	1
	Estágios 1 e 2		4	
	Estágios 3 e 4			2

VI. Conclusão

A base de dados, composta por 5 classes para gatos doentes e 5 classes para os gatos sem contato com o meloxicam, tendo, cada uma, 6 amostras para as substâncias hidrossolúveis e 5 amostras pra lipossolúveis e para o grupo composto por todas as substâncias, foi dividida em 3 bases de dados, cada um para um tipo de substância. Essas novas bases de dados tiveram suas substâncias ranqueadas e selecionadas através do FDR, e foram submetidas aos NB e KNN. Obtiveram-se então inúmeros resultados, sendo o melhor deles 100% com o NB para as substâncias lipossolúveis e com a análise de 3 classes. Já para a análise de 5 classes, o melhor resultado foi 93% com o NB para as substâncias hidrossolúveis. Com isso, conclui-se que o uso do FDR foi muito útil para selecionar o menor número possível, visto que nas duas situações foram usadas menos de 10 substâncias. Além disso, apesar da baixa quantidade de amostras, ambas acurácias foram maiores que 90%, o que indica um bom desempenho do NB. Com isso, o modelo de NB com a acurácia de 100% é o mais indicado por apresentar boa acurácia e utilizar apenas 6 substâncias. Vale ressaltar que o método ainda é limitado ao uso de poucas amostras no treinamento, o que pode impactar sua capacidade de generalização. Espera-se em trabalhos futuros melhorar o tamanho amostral para o desenvolvimento do modelo.

AGRADECIMENTOS

Os autores gostariam de agradecer à agência brasileira de fomento CNPq e também a UFLA, pelo suporte dado à pesquisa, o qual possibilitou a realização da mesma.

REFERÊNCIAS

[1] Rivera-Velez, S.M., Broughton-Neiswanger, L.E., Suarez, M. et al. Repeated administration of the NSAID meloxicam alters the plasma and urine lipidome. Sci Rep 9, 4303 (2019).

[2] Clayton, T. A. et al. Pharmaco-metabonomic phenotyping and personalized drug treatment. Nature. 440, 1073–1077 (2006).

[3] Takeuchi, K. et al. Metabolic profiling to identify potential serum biomarkers for gastric ulceration induced by nonsteroid antiinflammatory drugs. J. Proteome. Res. 12, 1399–1407 (2013).

[4] Peng, H., Wu, X., Zhao, L. & Feng, Y. Dynamic analysis of phospholipid metabolism of mouse macrophages treated with common non-steroidal anti-inflammatory drugs. Mol. cell. biochem. 411, 161–171 (2016).

[5] Basivireddy, J., Vasudevan, A., Jacob, M. & Balasubramanian, K. A. Indomethacin-induced mitochondrial dysfunction and oxidative stress in villus enterocytes. Biochem. Pharmacol. 64, 339–349 (2002).

[6] Galunska, B. et al. Effects of paracetamol and propacetamol on gastric mucosal damage and gastric lipid peroxidation caused by acetylsalicylic acid (ASA) in rats. Pharmacol. Res 46, 141–147 (2002).

[7] Fahy, E. et al. Update of the LIPID MAPS comprehensive classification system for lipids. J. lipid Res. 50, S9–S14 (2009)

[8] Afshinnia, F. et al. Lipidomic signature of progression of chronic kidney disease in the chronic renal insufficiency cohort. Kidney Int.Rep. 1, 256–268 (2016).

[9] Rhee, E. P. et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. J. Clinical Investig 121, 1402–1411 (2011).

[10] Dutta, T. et al. Concordance of changes in metabolic pathways based on plasma metabolomics and skeletal muscle transcriptomics in type 1 diabetes. Diabetes. 61, 1004–1016 (2012).

[11] Hinterwirth, H., Stegemann, C. & Mayr, M. Lipidomics: quest for molecular lipid biomarkers in cardiovascular disease. Cir. Genom. Precis Med 7, 941–954 (2014).

[12] Zhou, X. et al. Identification of plasma lipid biomarkers for prostate cancer by lipidomics and bioinformatics. PLoS ONE 7, e48889 (2012).

[13] Sas, K. M., Karnovsky, A., Michailidis, G. & Pennathur, S. Metabolomics and diabetes: analytical and computational approaches. Diabetes. 64, 718–732 (2015).

[14] Su, H. et al. Lipid deposition in kidney diseases: interplay among redox, lipid mediators, and renal impairment. Antioxid. Redox signal. 28, 1027–1043 (2018).

[15] Zad, N. et al. Development of machine learning algorithms to estimate maximum residue limits for veterinary medicines. Food and Chemical Toxicology179, 113920, (2023).

[16] Bedê, F. D. T. M., Dutra, L. V., & Sandri, S. Uma proposta de versão contextual para o classificador de vizinhos mais próximos. XI Workshop de Computação Aplicada, (2011).

[17] Pivetta, S. P. Classificação de documentos do exército brasileiro utilizando o classificador Naive Bayes e técnicas de seleção de sentenças. (2013).

[18] de Abreu Batista, R., Bagatini, D. D., and Frozza, R. Classificação Automática de Códigos NCM Utilizando o Algoritmo Naïve Bayes. iSys-Revista Brasileira de Sistemas de Informação, 11, 2, 4-29, (2018).

[19] Wang, S., Li, D., Song, X., Wei, Y., & Li, H. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Applications, 38, 7, 8696-8702, (2011).

[20] WEBB, A. R. Statistical pattern recognition. John Wiley & Sons, (2003).

[21] Rivera-Velez, S.M., Broughton-Neiswanger, L.E., Suarez, M. et al. Repeated administration of the NSAID meloxicam alters the plasma and urine lipidome. Sci Rep 9, 4303 (2019).

[22] Chawla, N. V., Bowyer, K. W., Hall, L. O., e Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357, (2002).

[23] Bifet, A., Holmes,G., Kirkby, R. and Pfahringer, B. "MOA: Massive Online Analysis." In Journal of Machine Learning Research, 11, 1601-1604, (2010).