SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA - PPGQUI

THIFANNY TIBURCIO PEREIRA DA SILVA

USO DA TÉCNICA LC-HRMS ASSOCIADA A MÉTODOS QUIMIOMÉTRICOS (PLS-DA e SVM) PARA DETECTAR CÂNCER DE PRÓSTATA ATRAVÉS DA URINA

Uberlândia
2023

THIFANNY TIBURCIO PEREIRA DA SILVA

USO DA TÉCNICA LC-HRMS ASSOCIADA A MÉTODOS QUIMIOMÉTRICOS (PLS-DA e SVM) PARA DETECTAR CÂNCER DE PRÓSTATA ATRAVÉS DA URINA

Dissertação apresentada ao Programa de Pós-graduação em Química do Instituto de Química da Universidade Federal de Uberlândia, como requisito para obtenção de Título de Mestre em Química.

Área de Concentração: Química Analítica

Orientador: Prof. Dr. Waldomiro Borges Neto

Uberlândia
2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU com dados informados pelo(a) próprio(a) autor(a).

S586 2023

Silva, Thifanny Tiburcio Pereira da, 1999-
Uso da técnica LC-HRMS associada a métodos quimiométricos (PLS-DA e SVM) para detectar câncer de próstata através da urina [recurso eletrônico] / Thifanny Tiburcio Pereira da Silva. - 2023.

Orientador: Waldomiro Borges Neto.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Química.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2023.478
Inclui bibliografia.

1. Química. I. Borges Neto, Waldomiro ,1970-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Química. III. Título.

CDU: 54

Bibliotecários responsáveis pela estrutura de acordo com o AACR2: Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Coordenação do Programa de Pós-Graduação em Química
Av. João Naves de Ávila, 2121, Bloco 5I - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902 Telefone: (34) 3239-4385 -
www.cpgquimica.iq.ufu.br - cpgquimica@ufu.br

ATA

Programa de Pós-Graduação em:	Química				
Defesa de:	Dissertação de Mestrado Acadêmico, 376, PPGQUI				
Data:	Vinte e sete de julho de dois mil e vinte e três	Hora de início:	9:00	Hora de encerramento:	11:50
Matrícula do Discente:	12122QMI008				
Nome do Discente:	Thifanny Tiburcio Pereira da Silva				
Título do Trabalho:	"Uso da técnica LC-HRMS associada a métodos quimiométricos (PLS-DA e SVM) para detectar câncer de próstata através da urina"				
Área de concentração:	Química				
Linha de pesquisa:	Instrumentação Analítica e Preparo de Amostras				
Projeto de Pesquisa de vinculação:	Aplicação de Métodos Quimiométricos				

Reuniu-se, por webconferência, links https://meet.google.com/tbm-hsmx-hb; https://meet.google.com/cqb-ewtd-ios; https://meet.google.com/rqh-yxxf-axa, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Química, assim composta: Professores Doutores: Welington de Oliveira Cruz, do IQUFU - Universidade Federal de Uberlândia; José Augusto da Col, da Universidade do Estado de Santa Catarina, e Waldomiro Borges Neto, orientador da candidata.

Iniciando os trabalhos o presidente da mesa, Dr. Waldomiro Borges Neto, apresentou a Comissão Examinadora e a candidata, agradeceu a presença do público, e concedeu à Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos(às) examinadores(as), que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o(a) candidato(a):

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do titulo de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Documento assinado eletronicamente por **Waldomiro Borges Neto**, **Professor(a) do Magistério Superior**, em 27/07/2023, às 11:42, conforme horário oficial de Brasília, com fundamento no art. 6º,

§ 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **José Augusto da Col**, **Usuário Externo**, em 27/07/2023, às 11:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de

8 de outubro de 2015.

Documento assinado eletronicamente por **Welington de Oliveira Cruz**, **Professor(a) do Magistério Superior**, em 27/07/2023, às 11:44, conforme horário oficial de Brasília, com fundamento no art. 6º,

§ 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site

https://www.sei.ufu.br/sei/controlador_externo.php?

acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4566008** e o código CRC **3F69B062**.

Referência: Processo nº 23117.040884/2023-28 SEI nº 4566008

Dedico,

Aos meus pais, Marizane e Gildo, aos meus irmãos Kauã e Gildo Filho, pela confiança, amor e incentivo.

AGRADECIMENTOS

Agradeço a Deus por sempre estar ao meu lado, dando forças para superar as dificuldades e não deixar desistir daquilo que sempre lutei. Foi preciso trilhar caminhos diferentes do planejado, para crescimento pessoal e profissional no qual me proporcionaram momentos únicos.

À minha família, que sem dúvidas nada seria possível sem seu apoio. Mencionar nome por nome seria uma tarefa difícil, portanto, minha eterna gratidão por todos que de uma forma direta ou indiretamente participou desse processo.

Ao meu orientador, Prof. Dr. Waldomiro Borges Neto, pelos ensinamentos, dedicação, paciência e, principalmente, pela compreensão e confiança depositada em mim.

Aos meus colegas do Laboratório de Quimiometria do Triângulo: Anisia Durans, Grazielly Amaral, Rodrigo, Carlos Alface, Assumane, Ademar Máquina, Maria Teresa, Edvando Teles, Lucas Gustavo, por todo conhecimento e ajuda compartilhada.

À Universidade Federal de Uberlândia e o Instituto de Química, por todo suporte e estrutura necessário para desenvolvimento desse projeto.

Aos professores e pesquisadores parceiros, Prof. Dr. Frederico Garcia Pinto, da Universidade Federal de Viçosa Campus Rio Paranaíba, Prof. Dr. Timothy J. Garrett do departamento de Patologia e Imunologia do Laboratório de Medicina da Universidade da Flórida, pelo fornecimento dos dados obtidos que viabilizou esse projeto.

Aos meus colegas de graduação, Celyvânia, Suzanna, Iorran e Léticia pelo apoio moral e pela troca de experiência. Aos mestrandos que passaram na minha trajetória na pós-graduação, gratidão pelo apoio durante os estudos.

Por fim, aos que estavam ao meu lado, o meu muito obrigada pelo incentivo, pela compreensão, pelo companheirismo.

RESUMO

O câncer de próstata (CP) é o tipo de câncer mais frequente entre a população masculina no Brasil, compondo 30% dos diagnósticos da doença no país. O diagnóstico prévio é uma forma de reduzir o número de mortalidade e, assim, obter maiores chances de cura. Os testes clínicos mais aplicados são o Antígeno Prostático Específico (PSA) e o toque retal. A biopsia é indicada quando é preciso visualizar a lesão identificada mais de perto. Embora seja uma doença comum entre a população masculina, é um problema de saúde grave, pois ao ser descoberto na fase inicial a chance de cura é superior a 90%, sendo assim, é necessário o desenvolvimento de novos testes clínicos eficientes para diagnóstico da doença, utilizando uma abordagem não invasiva, rápida e reprodutível do câncer de próstata. A técnica de cromatografia em fase líquida acoplada a espectrometria de massas de alta resolução (LC-HRMS) tornou-se uma ferramenta analítica importante em metabolômica, pois é uma técnica que apresenta alta sensibilidade e seletividade, contudo, apresenta alta complexidade da matriz, portanto, torna-se essencial o desenvolvimento de métodos analíticos para superar os desafios propostos nesse projeto em combinação com métodos quimiométricos, visando correlacionar as informações extraídas de espectros com as propriedades de interno caso diagnosticar pacientes com câncer de próstata pela análise da urina. O uso dos métodos quimiométricos de análise discriminante por quadrados mínimos parciais (PLS-DA) e máquina de vetores de suporte (SVM) para extrair as informações contidas nos espectros de massas permitiu o desenvolvimento de modelos estáveis, robustos e de fácil interpretação, capazes de discriminar amostras de pacientes com câncer de próstata e pacientes sadios com eficiência de 100%. Para o modelo PLS-DA os resultados obtiveram valores de RMSEC = 0,54; RMSECV = 0,64; RMSEP = 0,57; Sensibilidade = 1 e Especificidade = 1. Devido fornecer através do gráfico dos pesos as variáveis mais significativas, o método quimiométrico PLS-DA se torna o melhor a ser aplicado na busca de novos biomarcadores e métodos alternativos na detecção precoce da doença.

Palavras-chave: Câncer de próstata. LC-HRMS. Quimiometria. Biomarcadores.

ABSTRACT

Prostate cancer (PC) is the most frequent type of cancer among the male population in Brazil, comprising 30% of the diagnoses of the disease in the country. Prior diagnosis is a way to reduce the number of deaths and thus obtain greater chances of cure. The most applied clinical tests are Prostate Specific Antigen (PSA) and digital rectal examination. Biopsy is indicated when it is necessary to visualize the identified lesion more closely. Although it is a common disease among the male population, it is a serious health problem, because when it is discovered in the initial phase, the chance of cure is greater than 90%, therefore, it is necessary to develop new efficient clinical tests for the diagnosis of the disease. , using a non-invasive, rapid and reproducible approach to prostate cancer. The liquid chromatography technique coupled with high resolution mass spectrometry (LC-HRMS) has become an important analytical tool in metabolomics, as it is a technique that presents high sensitivity and selectivity, however, it presents high complexity of the matrix, therefore , it is essential to develop analytical methods to overcome the challenges proposed in this project in combination with chemometric methods, aiming to correlate the information extracted from spectra with the internal properties of diagnosing patients with prostate cancer through urine analysis. The use of chemometric methods of discriminant analysis by partial least squares (PLS-DA) and support vector machine (SVM) to extract the information contained in the mass spectra allowed the development of stable, robust and easy to interpret models, capable of discriminate samples from patients with prostate cancer and healthy patients with 100% efficiency. For the PLS-DA model, the results obtained values of RMSEC = 0.54; RMSECV = 0.64; RMSEP = 0.57; Sensitivity = 1 and Specificity = 1. As it provides the most significant variables through the graph of weights, the PLS-DA chemometric method becomes the best to be applied in the search for new biomarkers and alternative methods for early detection of the disease.

Keywords: Prostate cancer. LC-HRMS. Chemometrics. Biomarkers.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CP Câncer de Próstata

INCA Instituto Nacional de Câncer

PSA Antígeno Prostático Específico (do inglês, *Prostate Specific Antigen*)

LC-HRMS Cromatografia Líquida Acoplada a Espectrometria de Massas de Alta Resolução (do inglês, *Liquid Chromatography Coupled – Mass Spectrometry*)

PLS-DA Análise Discriminante por Quadrados Mínimos Parciais (do inglês, *Partial Least Squares - Discriminant Analysis*)

SVM Máquina de Vetores de Suporte (do inglês, *Support Vector Machine*)

VL Variáveis Latentes

RMSEC Erro Quadrático Médio de Calibração (do inglês, *Root Mean Squared Error of Calibration*)

RMSECV Erro Quadrático Médio de Validação Cruzada (do inglês, *Root Mean Squared Error of Cross Validation*)

RMSEP Erro Quadrático Médio de Previsão (do inglês, *Root Mean Squared Error of Prediction*)

FP Taxa de Falso Positivo

VP Taxa de Verdadeiro Positivo

FN Taxa de Falso Negativo

VN Taxa de Verdadeiro Negativo

SUMÁRIO

1. INTRODUÇÃO

A próstata é um órgão presente nos homens, localizada na frente do reto, abaixo da bexiga, englobando parte superior da uretra, canal onde a urina armazenada na bexiga é eliminada. A próstata não é responsável pela ereção nem pelo orgasmo, sua função é produzir e armazenar parte do líquido seminal em função de nutrir e proteger os espermatozoides. O câncer de próstata (CP) é o tipo de câncer mais frequente entre a população masculina no Brasil, compondo 30% dos diagnósticos da doença no país. Dados do Instituto Nacional de Câncer (INCA, 2022) apontam para 71.730 novos casos a cada ano para o triênio 2023 a 2025.

O diagnóstico prévio é uma forma de reduzir o número de mortalidade e assim, obter maiores chances de cura. No caso do câncer de próstata, a detecção é realizada através de exames clínicos e laboratoriais, sendo os mais comuns o exame de toque retal, onde o médico insere um dedo com luva e lubrificante no reto do paciente para determinar qualquer inchaço ou áreas endurecidas na próstata que possam eventualmente ser um câncer, e o exame de sangue do antígeno prostático específico (PSA), que trata-se de moléculas produzidas pela próstata, inclusive quando ela está saudável, o que muda, na verdade, é a dosagem de PSA em circulação. Ambos os exames não são específicos para CP, apresentando uma alta taxa de falsos positivos ou negativos, porém são incluídos como parte do rastreamento.

Biopsia ou ultrassom trans retal é indicado quando é preciso visualizar mais de perto a lesão identificada em outros exames. Na biópsia pequenas amostras da próstata são removidas e enviadas para análise em um laboratório de patologia.

Embora seja uma doença comum entre a população masculina, devido à falta de conhecimento, medo, constrangimento, métodos de detecção precoce invasivos, é um problema de saúde grave, pois ao ser descoberto na fase inicial a chance de cura é superior a 90%. Dessa forma, é necessário o desenvolvimento de novos testes clínicos eficientes para diagnóstico da doença, utilizando uma abordagem não invasiva, rápida e reprodutível do câncer de próstata.

As aplicações da metabolômica nos últimos anos tem sido empregada em diversas áreas, como por exemplo, para identificação de marcadores bioquímicos de tumores, estudo da diversidade biológicas de interesse medicinal e alimentar (MADSEN, et al., 2010), entre outros.

A técnica de cromatografia em fase líquida acoplada a espectrometria de massas de alta resolução (LC-HRMS) pode ser empregada na busca de biomarcadores para câncer, devido fornecer uma imagem dos constituintes químicos com ampla caracterização de polaridade e massa molecular (BUSTILLOS, 2020). Esta imagem pode detectar um câncer a partir de dados

químicos e não apenas pela morfologia das células como é feito atualmente. Um dos principais desafios desta técnica inclui a alta complexidade da matriz, pois na cromatografia irá ocorrer a separação de compostos (moléculas) pertencentes a uma mistura em duas fases, estacionária e móvel, gerando assim, uma banda no espectrômetro para cada composto. Na espectrometria de massas irá ocorrer a fragmentação desses compostos previamente ionizados, dessa forma, os íons gerados são separados em razão massa carga, gerando espectros desses compostos presentes na amostra.

As informações contidas nos espectros obtidos são complexas e de difícil interpretação por análise visual, sendo necessário o uso de métodos quimiométricos para extrair e correlacionar com a propriedade de interesse, no caso, a detecção do câncer de próstata. Para tal finalidade os métodos de classificação apresentam eficiência e resultados em uma dimensão menor, tornando possível a visualização dos possíveis agrupamentos em gráficos de 2 ou 3 dimensões. Nos métodos propostos no seguinte trabalho, além de informações sobre a distribuição das amostras é possível através da análise de pesos avaliar quais foram as variáveis (picos do espectro) mais significativas para distinguir as amostras dos pacientes com câncer das amostras de pacientes sadios. Ao identificar as moléculas que deram origem aos picos mais significativos, pode-se ter candidatos a possíveis marcadores biológicos para o câncer de próstata.

O método de Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA do inglês, *Partial Least Squares Discriminant Analysis*) é um método multivariado utilizado para classificação de amostras onde a redução de variáveis e a variabilidade entre grupos seja maior que adentro do grupo. O bloco **Y** em um modelo de PLS1-DA indica a classe à qual uma amostra pertence. Neste caso existe uma variável dependente **y** que normalmente assumem os valores 0 ou 1. Para a separação das classes é necessário o cálculo de um valor limite, denominado de *threshold* representado por uma linha tracejada no gráfico. Quando a amostra apresentar um valor previsto acima do *threshold*, ela é considerada como pertencente à classe 1 e as amostras abaixo deste valor limite são pertencentes à classe 0. Sendo assim, o PLS-DA (um método supervisionado) permite construir modelos de classificação capazes de prever novas amostras (BARKER; RAYENS, 2003).

As Máquinas de Vetores de Suporte (SVM do inglês, *Support Vector Machine*), desenvolvido inicialmente pelo pesquisador russo Vladmir Vapnik (1995), é um modelo estatístico, supervisionado, com capacidade de resolver problemas de classificação e análise de regressão. As propostas das SVMs é a construção de um hiperplano de separação, tendo-se dados de treinamento rotulados, o algoritmo gera um hiperplano ideal que classifica novos

dados. Em um espaço bidimensional, esse hiperplano é uma linha que separa em duas classes, onde cada classe fica em cada lado. Dessa forma, o presente trabalho tem como objetivo desenvolver método analítico capaz de detectar o câncer de próstata pela análise da urina de pacientes usando LC-HRMS e quimiometria.

1.1 OBJETIVOS

O trabalho tem como objetivo geral, desenvolver métodos analíticos capazes de detectar o câncer de próstata pela análise da urina de pacientes usando a técnica de cromatografia em fase líquida acoplada a espectrometria de massas de alta resolução (LC-HRMS) e métodos quimiométricos (PLS-DA e SVM).

1.2 OBJETIVOS ESPECÍFICOS

- ✓ Construir modelos de PLS-DA e SVM para detectar câncer de próstata em amostras de urina analisadas por LC-HRMS;

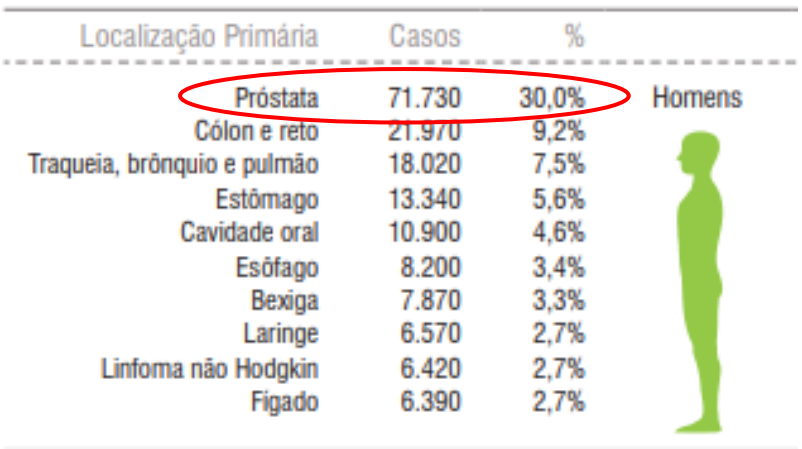
- ✓ Validar os modelos construídos.

2. REVISÃO BIBLIOGRÁFICA

2.1 Câncer de Próstata

O câncer é o principal problema de saúde pública no mundo, figurando como uma das principais causas de morte e, como consequência, uma das principais barreiras para o aumento da expectativa de vida em todo o mundo (INCA, 2022). O câncer de próstata é o segundo tipo de câncer mais frequente no Brasil (sem considerar os tumores de pele não melanoma) e o primeiro entre a população masculina. Mundialmente é o quarto mais frequente totalizando 7,3% dos casos (INCA, 2022).

Figura 1 – Tipos de câncer mais incidentes pela população masculina estimados para 2023



Fonte: INCA, 2022

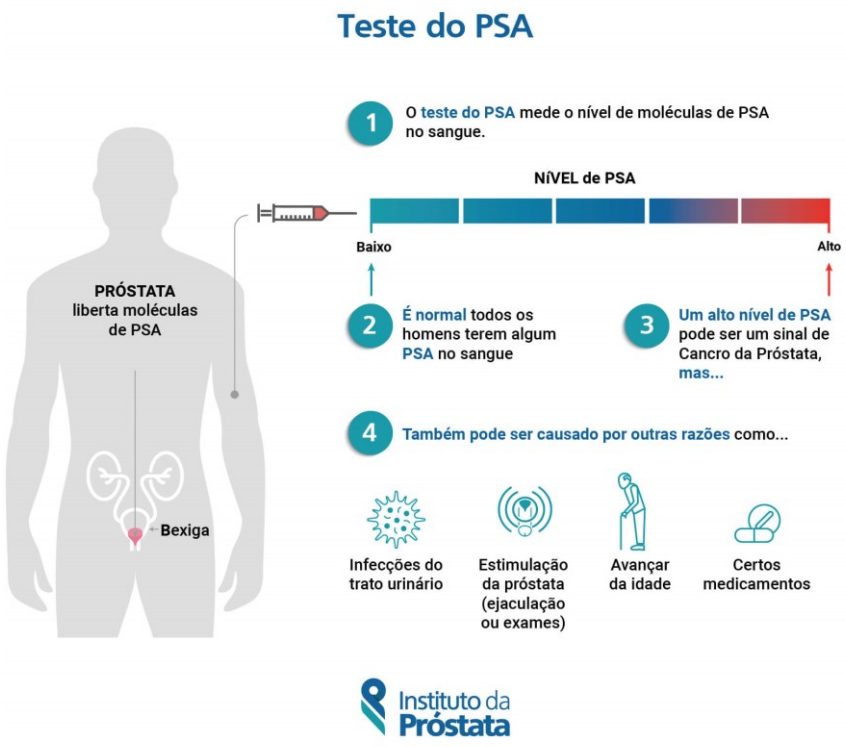
Um dos principais fatores para desenvolvimento da doença é a idade, pois a próstata aumenta naturalmente com o avançar da idade, homens acima de 50 anos, com excesso de peso e obesidade estão mais propensos a doença. Segundo Damião *et al.* (2015, p.81) "aspectos étnicos e geográficos também são fatores de risco".

> O câncer de próstata é aproximadamente duas vezes mais comum em homens negros se comparados aos brancos. Os estadunidenses, jamaicanos e caribenhos com ascendência africana apresentam as mais altas taxas de incidência do câncer de próstata do mundo, o que pode ser atribuído, em parte, à hereditariedade (cerca de 5% a 10%). Apesar disso, é possível que essa diferença entre negros e brancos se dê também em razão do estilo de vida, fatores dietéticos ou por diferenças no acesso ao diagnóstico da doença (DAMIÃO *et al*., 2015, p.81).

Para investigação da doença existem exames clínicos e laboratoriais, sendo, o exame do toque retal realizado pelo médico urologista. “Ao realizar o exame, o médico tenta definir algumas características da próstata: tamanho, consistência, superfície, forma, limites e sensibilidade” (SARRIS *et al*., 2018, p.142). É um exame que apresenta limitações por ser bastante desconfortável e invasivo, além de não ser palpável em estágios iniciais do câncer.

O exame do Antígeno Prostático Específico (PSA) é uma glicoproteína produzida pela próstata, cuja função é liquefazer o sêmen após a ejaculação. Os níveis séricos de PSA têm sido correlacionados ao câncer, porém também podem estar elevados em doenças benignas como a prostatite e a hiperplasia prostática benigna (LIMA; SILVA; ALVES, 2017, p.12). Um nível baixo de PSA é normal pois todos os homens possuem PSA no sangue, um alto nível de PSA pode ser um sinal de câncer de próstata, porém também pode ser causado por uma infecção de urina, estimulação da próstata, avançar da idade, certos medicamentos, como representado na Figura 2. Desse modo, é um teste que apresenta altas taxas de falsos positivos e negativos, além de não ser específico para o câncer de próstata.

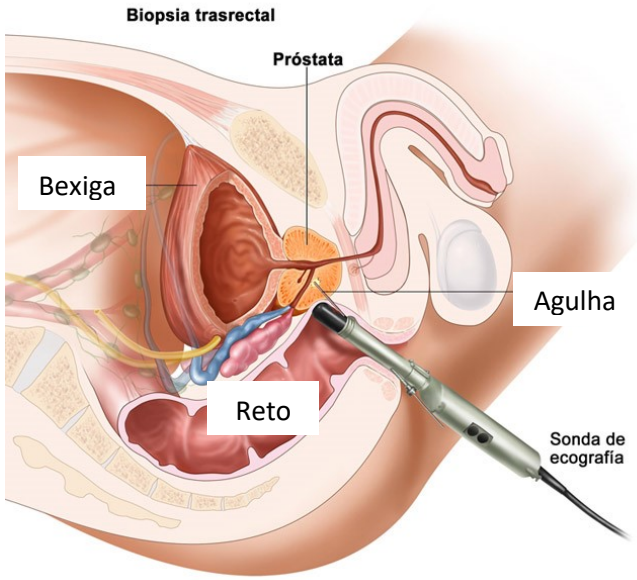
Figura 2 – Teste do PSA



Fonte: Instituto da próstata, s/d.

A biopsia prostática trans-retal é o único exame atualmente que permite o diagnóstico definitivo do câncer de próstata. Quando se realiza uma biopsia, utiliza-se uma sonda de ecografia onde é retirado de 12 a 14 fragmentos da glândula para análise em regiões padronizadas, previamente definidas de acordo com o protocolo do paciente (INSTITUTO DA PRÓSTATA, s/d). Se os fragmentos alcançados não forem totalmente esclarecedores ou se a suspeita da presença de um tumor for elevada e o resultado das biopsias forem negativos, há casos que são necessários repetir duas ou mais vezes para se ter um resultado definitivo. As complicações que podem decorrer da biópsia são a ocorrência de hemorragia e o desenvolvimento de uma infecção (prostatite).

Figura 3 – Biopsia da próstata



Fonte: Grupo Eiger, s/d.

Outros exames como tomografia computadorizada, ressonância magnética, ultrassonografia via abdominal, também podem ser solicitados para rastreamento da doença (INCA, 2022).

2.2 Biomarcadores

Biomarcadores ou marcadores biológicos são moléculas biológicas encontradas no sangue ou nos tecidos corporais que podem ser usados para indicar uma condição biológica ou uma doença. Eles podem ser de diversos tipos, fisiológicos (funções de órgãos), histológicos (amostras de tecido obtidas por biopsia) e anatômicos. Geralmente se refere a proteínas, genes e outras moléculas que influenciam na detecção de células cancerígenas, como elas crescem, multiplicam, morrem e respondem ao tratamento.

De acordo com Dr. Roberta Galvão Campos, especialista em oncogenética, um biomarcador pode ajudar a identificar cânceres em estágio inicial, prever o quão agressivo pode ser um câncer ou prever o quão bem o paciente responderá ao tratamento. O crescente conhecimento dos mecanismos moleculares juntamente com a aplicação de técnicas de *high-throughput* (alto rendimento) nas análises ômicas vêm fornecendo a possibilidade de identificação de potenciais biomarcadores específicos para diferentes doenças (ALBERICE, 2014).

O uso de biomarcadores urinários é utilizado na prática clínica para diagnóstico há considerável tempo. Por exemplo, os testes rápidos de gravidez utilizam-se a urina como fonte de biomarcador. Durante muito tempo a glicose foi detectada em urina avaliando se formigas eram atraídas por ela. Assim, estudos para identificar biomarcadores de doenças na urina têm sido um componente fundamental da medicina investigativa ao longo dos séculos XX e XXI (ALBERICE, 2014).

De um modo geral, da mesma forma que os biomarcadores se correlacionam com suas respectivas neoplasias, os biomarcadores de próstata exibem algumas ou todas as seguintes habilidades: diagnóstica, prognóstica, preditiva e farmacodinâmica. Biomarcadores de diagnóstico representam uma estratégia de triagem (*screening*) ou confirmação do diagnóstico de uma neoplasia. Os biomarcadores prognósticos preveem o curso natural do câncer para distinguir o desfecho do tumor. Eles também ajudam a determinar quem tratar, a intensidade do tratamento, quais candidatos provavelmente responderão a um determinado medicamento e a dose mais eficaz. Biomarcadores preditivos avaliam o provável benefício de um tratamento específico (PAIVA, 2020). Já os biomarcadores farmacodinâmicos avaliam os efeitos iminentes do tratamento de um medicamento em um tumor e, possivelmente, podem determinar a dosagem nos estágios iniciais do desenvolvimento clínico de um novo medicamento anticâncer (PAIVA, 2020).

Um candidato ideal a biomarcador tem que superar limitações, tais como sensibilidade, especificidade e robustez insuficiente, além de baixo poder preditivo. Ele ainda deve ter característica não invasiva e ser extraído preferencialmente de fontes como sangue e urina. As principais características de um biomarcador tumoral ideal são sua especificidade para um determinado tipo de tumor e sua sensibilidade, sendo que seus níveis devem sinalizar com precisão a progressão e a regressão do tumor (PAIVA, 2020).

2.3 Metabolômica

A metabolômica é o conjunto de todos os metabólitos presentes em um sistema biológico seja na célula, tecido e/ou no organismo. Pode ser entendida como o estudo dos metabólitos que, por sua vez, são produtos intermediários ou finais do metabolismo em um sistema biológico (CANUTO *et al*., 2018). Dessa forma, a metabolômica abrange uma ampla gama de compostos químicos, tais como, aminoácidos, lipídios, ácidos nucleicos, carboidratos, vitaminas, dentre outros (AGIN *et al*., 2016).

Assim, a metabolômica faz parte do conjunto das ciências "ômicas" que incluem (genômica, transcriptômica e proteômica), sendo que o foco destas é, respectivamente, os genes, RNAm (ácido ribonucleico do tipo mensageiro), proteínas e metabólitos. Alguns aspectos que diferem a metabolômica das outras ômicas são: o número de metabólitos é menor que o de genes e proteínas, o que reduz a complexidade de seu estudo; a obtenção de um perfil metabolômico é mais barato e rápido que as análises em proteômica ou transcriptômica (PAIVA, 2020). Entretanto, tais características não desmerecem a utilidade e complexidade de análises metabolômicas.

Particularmente na área médica, a metabolômica vem demonstrando um papel promissor à medida que pode proporcionar biomarcadores para diagnóstico e prognóstico de doenças, promover a predição da eficácia e segurança de intervenções farmacêuticas e fornecer percepções sobre os mecanismos bioquímicos de doenças e a modulação por drogas (PAIVA, 2020).

As técnicas analíticas aplicadas em metabolômica destaca-se a ressonância magnética nuclear (RMN), cromatografia líquida (LC) e gasosa (GC) acoplada à espectrometria de massas (MS), dentre outras. Cada uma com sua particularidade e estratégia, em especial, as técnicas acopladas à espectrometria de massas vêm ganhando campo pela sensibilidade que tais técnicas podem oferecer, além do menor custo quando comparado a um equipamento de RMN de alta resolução (PAIVA, 2020). Desta forma, a escolha da técnica analítica está diretamente

relacionada com as propriedades físicas e químicas dos metabólitos a serem analisados, por exemplo, tamanho, volatilidade, presença de grupos ionizáveis, compostos ácidos ou básicos, e etc. (DUNN, 2011) Dependendo da questão biológica, são necessárias diferentes demandas de desempenho analítico, por exemplo, limites de detecção, precisão, exatidão, quantificação, dentre outros parâmetros (GOULART, 2018).

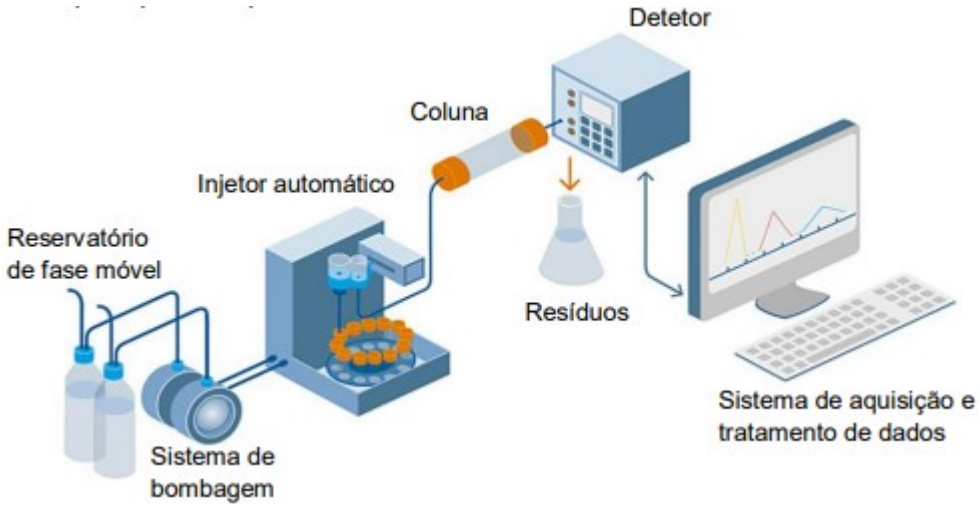
Com relação a análise de dados, ao projetar experimentos em metabolômica, a escolha do método de análise estatística multivariada deve ser conduzida de acordo com o tipo dos dados (espectros de massas, concentrações e etc.) e pelos objetivos experimentais (GOULART, 2018).

2.4 Cromatografia Líquida

A cromatografia líquida de alta resolução constitui uma das técnicas mais utilizadas em química analítica, quer para análises quantitativas ou qualitativas (SARGAÇO, 2013). Seu princípio básico é a separação e quantificação dos componentes de uma mistura através de interações com duas fases: móvel e estacionária.

Um sistema de cromatografia líquida é composto pelo reservatório da fase móvel (solventes), uma bomba, um injetor de amostras, uma coluna cromatográfica (fase estacionária), um detector, descarte de resíduos e sistema de aquisição para tratamentos dos dados, como representado na figura 4.

Figura 4 – Esquema representativo por análise de cromatografia líquida



Fonte: SARGAÇO, 2013.

O sistema de bombeamento opera sob alta pressão para forçar a passagem da fase móvel até a coluna cromatografia, com um fluxo específico em mililitros por minuto (mL/min). Este sistema pode ser programado para variar a proporção dos componentes da fase móvel ao longo da corrida cromatográfica, caso se pretenda uma eluição em gradiente ou para fixar uma proporção desses componentes para eluições isocráticas (SARGAÇO, 2013).

O tipo de eluição utilizado durante a análise afeta consideravelmente não somente a capacidade de pico, mas também a ortogonalidade, devido ao alargamento de banda que o composto sofre através do sistema (MONGOLLÓN, 2014). O método de eluição isocrática utiliza a mesma força cromatográfica do início ao fim da análise (KONZEN, 2015). Ou seja, é utilizado um único solvente ou uma mistura de solventes cuja proporção utilizada permanece a mesma. A eluição por gradiente é aquela em que são utilizados dois ou mais sistemas de solventes com diferenças de polaridade durante a separação e assim observa-se um aumento gradativo na força cromatográfica por meio da elevação da porcentagem de solvente orgânico (BRAGA, 2018).

O injetor automático realiza a injeção da amostra no sistema e vai carregar até a coluna cromatografia onde está presente a fase estacionária. A coluna pode possuir diferentes tipos de materiais de enchimento, que, juntamente com a fase móvel vão determinar a separação pretendida (SARGAÇO, 2013). A coluna é considerada o coração de qualquer sistema cromatográfico, pois é nela que se dá a separação dos diferentes constituintes (analitos) de uma amostra (PAZ, 2018). As separações são obtidas através de processos por partição, adsorção ou por troca iônica, dependendo do tipo de coluna utilizada.

À medida que os compostos vão sendo eluídos da coluna é necessário um detector para quantificar e identificar os analitos. O detector está ligado a um computador que registra o sinal emitido gerando um cromatograma, no qual os diferentes analitos são representados por curvas aproximadamente gaussianas com diferentes tempos de retenção (SARGAÇO, 2013). Através dos tempos de retenção é possível identificar os analitos e com base nas áreas dos respetivos picos efetua-se a sua quantificação.

2.5 Cromatografia Líquida Acoplada à Espectrometria de Massas

Espectrometria de massas e cromatografia são técnicas analíticas bastante estudadas e desenvolvidas ao longo das últimas décadas e, por isso, detêm papel de destaque nas ciências de separação. O avanço de ambas, HPLC e MS contribuíram significativamente para as análises metabolômicas (ALBERICE, 2014).

Estas técnicas são comumente utilizadas para a caracterização e obtenção de informação estrutural de compostos; no campo da metabolômica, estas duas técnicas analíticas podem ser combinadas para caracterizar metabólitos endógenos ou exógenos desconhecidos presentes em amostras biológicas complexas (ALBERICE, 2014). Vários nomes têm sido utilizados para denominar esta técnica: alta velocidade, alta pressão, alto desempenho, alta resolução e alta eficiência (GOULART, 2018).

Cromatografia é um método de separação que depende fundamentalmente da diferença de interação/distribuição dos analitos (moléculas de interesse) entre uma fase que permanece móvel e outra fase que permanece estacionária (GOULART, 2018). O espectrômetro de massas é composto basicamente por um sistema de introdução de amostra, ionização, análise de massas, detecção dos íons gerados e processamento dos dados. Após os analitos serem introduzidos no espectrômetro de massas, ocorrerá a produção de íons na fonte de ionização e esses são direcionados ao analisador de massas e analisados de acordo com sua razão massa/carga (m/z) (JANISCH, 2022). O acoplamento de um cromatógrafo com o espectrômetro de massas combina as vantagens da cromatografia (alta seletividade e eficiência de separação) com as vantagens da espectrometria de massas (obtenção de informação estrutural, massa molar e aumento adicional da seletividade). (GOULART, 2018).

A ampla utilização de cromatografia líquida e espectrometria de massas em análises metabolômicas deve-se, além das vantagens analíticas como robustez, reprodutibilidade e alta resolução, à versatilidade do método de separação intrínseco do método de HPLC (ALBERICE, 2014). Esta permite a separação de compostos de uma vasta gama de polaridade, através de eluição isocrática ou de gradiente de eluição. A eluição isocrática é preferida para amostras simples (ou seja, menos de 10 componentes), enquanto o gradiente de eluição proporciona um conjunto mais rápido de análises (ALBERICE, 2014).

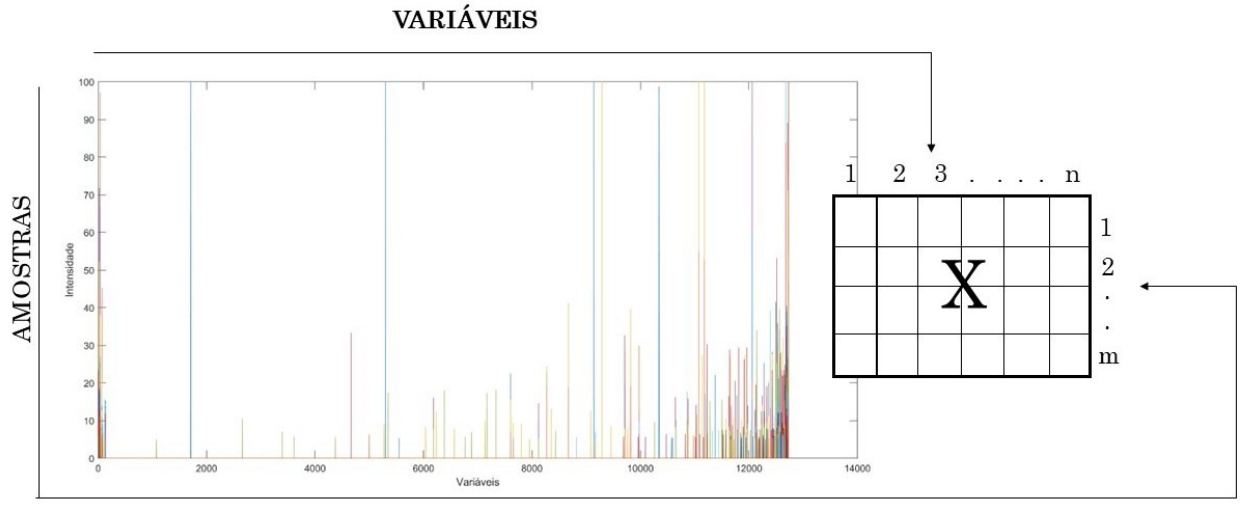
2.6 Métodos Quimiométricos

A aquisição de dados principalmente na área de química analítica, atingiu um ponto bastante sofisticado com o interfaceamento de instrumentos aos computadores produzindo uma enorme quantidade de informação, muitas vezes complexa e variada (FERREIRA *et al*., 1999). Dessa forma, a quimiometria surgiu da necessidade de extrair ao máximo a informação química útil de dados produzidos pela moderna instrumentação, não podendo dessa maneira serem explicados por modelos teóricos da química clássica.

A quimiometria é definida como a aplicação de métodos matemáticos, estatísticos e de lógica formal para o tratamento de dados químicos de forma a extrair uma maior quantidade de informação e otimização dos resultados analíticos (MASSART *et al*., 1998). Dentre os métodos quimiométricos destaca-se três áreas: planejamento e otimização de experimentos, calibração multivariada e classificação, podendo ser subdividida em métodos supervisionados e não supervisionados (MASSART *et al*., 1998).

Os dados multivariados, em geral, são organizados em matrizes de dados **X** de forma que cada amostra "m" represente um vetor linha e cada variável "n" represente uma coluna, na qual, a matriz de dados terá dimensão $\mathbf{X}_{(m \times n)}$ (m linhas por n colunas). A Figura 5 representa a organização de dados espectrais para criação da matriz **X**.

Figura 5 – Organização dos dados para construção da matriz **X**



Fonte: Adaptado, (MÁQUINA, 2017).

Muita ênfase tem sido dada aos sistemas multivariados, nos quais se pode medir muitas variáveis, simultaneamente (ou de forma sequencial, com grande eficiência) ao se estudar uma amostra qualquer. O presente trabalho, utiliza métodos de classificação multivariada, também conhecidos como métodos supervisionados de reconhecimento de padrões. Na análise supervisionada, seleciona-se uma série de amostras representativas de cada classe e para as quais as medidas experimentais são coletadas e o padrão de cada uma delas é definido. Esse conjunto de amostras a classe à qual cada uma delas pertence, é conhecida como "conjunto de treinamento". Utilizando as informações do conjunto de treinamento, é construído um modelo empírico ou uma regra de classificação. Esses métodos de análise são denominados "supervisionados", pois as informações a respeito das classes é que supervisionam o desenvolvimento dos critérios de discriminação que serão utilizados posteriormente para fazer o reconhecimento de novas amostras (FERREIRA, 2015).

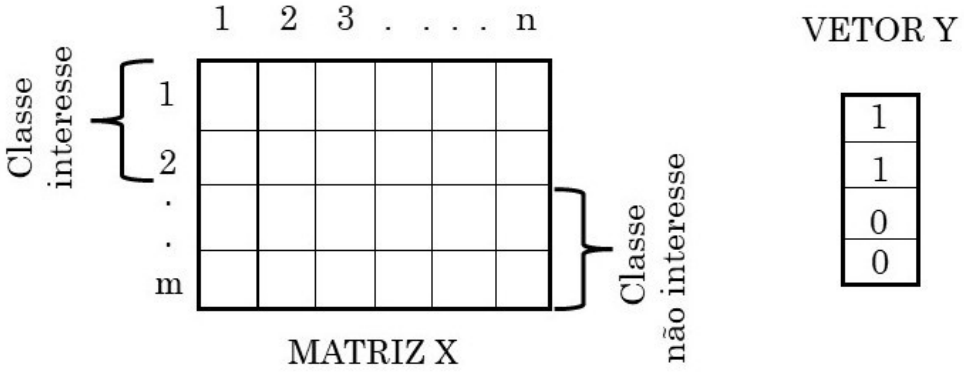
Dado um número de classes, cada uma definida por um conjunto de amostras e os valores das medidas feitas para cada uma delas, deseja-se encontrar as regras para classificar novas amostras para as quais serão adquiridas as mesmas medidas. Antes de sua utilização, o modelo empírico obtido deve ser testado para verificar sua capacidade de prever com sucesso a classe de novas amostras. Para isso, recomenda-se a utilização de um conjunto de amostras externas, denominado de "conjunto de teste" ou "conjunto de validação". Se o resultado for satisfatório, ele pode, então, ser utilizado para identificar a propriedade de interesse (a classe) de novas amostras (FERREIRA, 2015).

2.7 Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA)

O método PLS-DA é uma variação do algoritmo de regressão por quadrados mínimos parciais (PLS), onde as variáveis da matriz **X** (dados espectrais) são correlacionadas com a propriedade de interesse categórica (discreta) contidas em uma matriz **Y**, como mostra a Figura 6. Nos modelos PLS1-DA (onde a matriz **Y** é uma matriz linha, ou seja, um vetor), o vetor de correlação **y** é constituído por valores inteiros (0 e 1), onde 0 é usado para classe não de interesse e 1 para a classe de interesse. Idealmente, os valores previstos pelo modelo PLS-DA seriam 0 ou 1, mas na prática os valores previstos são próximos de 0 e 1. Tendo em conta que os valores previstos não são 0 ou 1, faz-se necessário o cálculo de um valor limite para separar as classes, denominado de *threshold*.

O valor de limiar *threshold* é previsto entre 0 e 1, com base no teorema de Bayes, segundo o qual, o limiar assume que os valores de **y** previstos seguem uma distribuição semelhante ao que vai ser observado para as amostras futuras. Quando as duas distribuições estimadas se cruzam, é selecionado um limite, isto é, um valor de **y** em que o número de falsos positivos e falsos negativos deve ser minimizado para previsões futuras (ALMEIDA *et al*., 2013). Quando a amostra apresentar um valor previsto acima do *threshold*, ela é considerada como pertencente à classe 1 e as amostras abaixo deste valor limite são pertencentes à classe 0.

Figura 6 – Matriz **X** de dados e o vetor **y** aplicados no modelo PLS-DA



Fonte: a autora.

Para a construção do modelo PLS-DA, procura-se encontrar a quantidade de variáveis latentes (VL) que descrevam a covariância nas amostras das matrizes e que tenham a correlação máxima com a classe de valores conhecidos, dando menos peso para a classe irrelevante ou a variância do ruído (BARKER; RAYENS, 2003). O número de VL é determinado no processo de validação cruzada onde geralmente se usa o método "*leave one out*" (deixar uma de fora por vez), onde na etapa de treinamento uma amostra é deixada de fora no processo de construção do modelo e a seguir essa amostra é prevista pelo modelo construído. Na sequência, esta amostra retorna ao conjunto de calibração, e outra é retirada e prevista pelo modelo construído pelas restantes, e esse processo é repetido até que todas as amostras tenham sido retiradas do modelo uma vez (MASSART et al., 1998).

Outro método de validação cruzada é recomendado quando se tem acima de 20 amostras no conjunto de calibração (treinamento), pelo critério de veneziana que é realizado em blocos de amostras, isto é, um número determinado de amostras é deixado de fora no processo de construção do modelo e a seguir essas amostras são previstas pelo modelo

construído, seguindo o mesmo processo que o do método *leave one out*. A eficiência da validação é expressa pelo Erro Quadrático Médio de Validação Cruzada (RMSECV), calculada pela Equação 1.

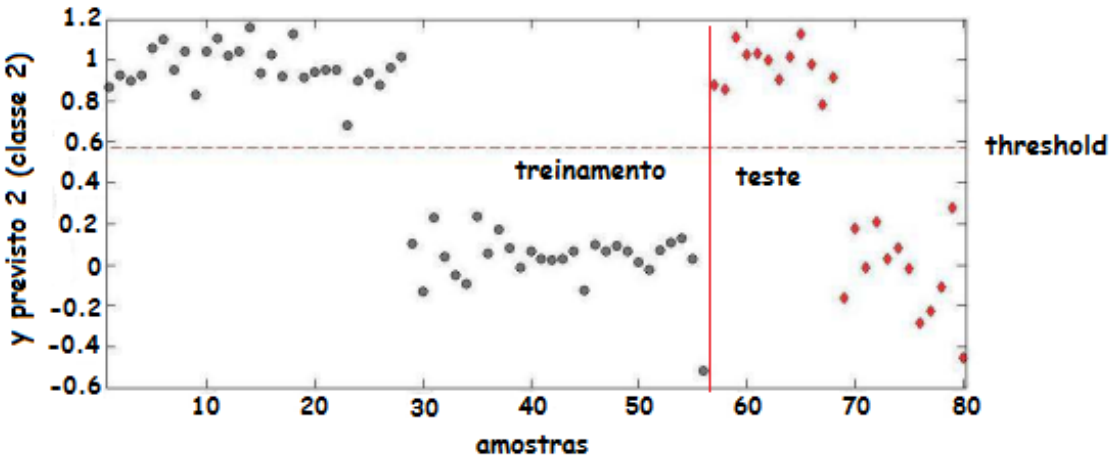
$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

Onde yi , $\hat{y}i$ e n são, os valores reais, valores previstos pelo modelo PLS-DA e número de amostras do conjunto de validação cruzada, respectivamente.

A Figura 7 mostra um exemplo do gráfico de estimativas para um modelo PLS-DA, onde as amostras do conjunto de treinamento (●) posicionadas acima da linha tracejada (*threshold)* pertence a classe 1 (um), no caso a classe de interesse e, as amostras posicionadas abaixo da linha *threshold* pertence a classe 0 (zero), a classe de não interesse.

Da mesma forma, as amostras do conjunto de teste (♦) posicionadas acima e abaixo da linha *threshold,* pertence a classe 1 (classe de interesse) e classe 0 (classe de não interesse).

Figura 7 – Representação de um gráfico de estimativas para um modelo PLS-DA



Fonte: Durans (2022, p.45).

Após a construção do modelo PLS-DA, pode-se fazer a análise do gráfico dos pesos, onde é possível verificar quais variáveis foram responsáveis para separação das classes das amostras de interesse e não interesse.

2.7.1 Detecção de Amostras Anômalas (*outlier*)

Amostras anômalas (*outliers*) são aquelas em que seu perfil é diferente de todas as outras, podendo interferir inadequadamente no modelo proposto. O processo de detecção de *outliers* em um conjunto de dados é realizado geralmente na etapa de construção do modelo (calibração ou treinamento) (SMITI, 2020). Podem ocorrer por várias razões, entre elas:

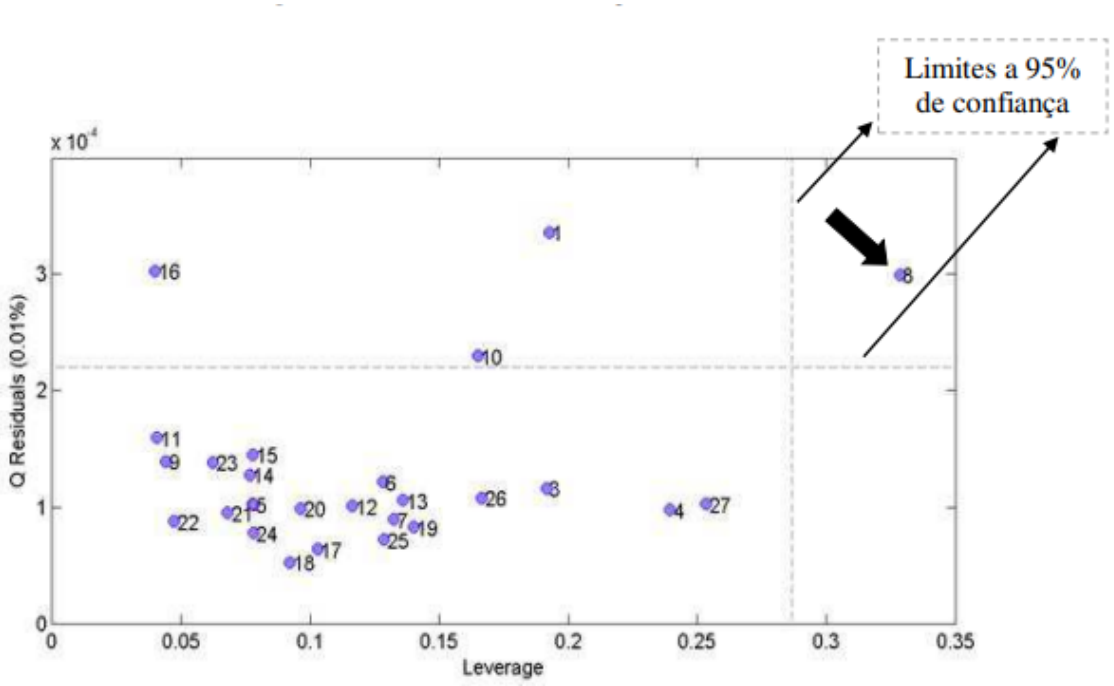
- Presença de ruídos;
- Erros de digitação;
- Erros de laboratório;
- Erros instrumentais.

Uma vez que este tipo de amostra pode afetar a qualidade global do modelo quimiométrico, sujeitando a valores altos de erro, faz-se necessário à sua retirada do conjunto de calibração, possibilitando a criação de modelos mais eficazes e precisos com bom poder preditivo (VALDERRAMA, 2009).

As amostras anômalas na calibração são geralmente examinadas com referência na *leverage* extremo, que contém resíduos fora do modelo no conjunto de dados espectrais e na variável dependente. O *leverage* caracteriza o quanto uma determinada amostra está longe da média do grupo de dados (SANTOS, 2019). Ao usar o conjunto de informações espectrais de exemplo, este parâmetro avalia o quanto o espectro de determinada amostra está em desacordo com os espectros das outras amostras contidas no grupo de dados (VALDERRAMA, 2009).

De acordo com a Figura 8, observa-se que a amostra de nº 8 está fora dos limites de confiança a 95% (linhas tracejadas), isto é, apresenta um alto valor de *leverage* e alto valor de resíduos. Sendo assim, é necessário à sua retirada do conjunto de calibração, pois sua permanência acarreta em informações errôneas, podendo diminuir a eficiência do modelo, caso a informação seja correlacionada com as amostras do conjunto de previsão.

Figura 8 – Gráfico de *Leverage* versus Q *Residuals*



Fonte: Máquina (2017, p.40).

2.8 Máquina de Vetores de Suporte (SVM)

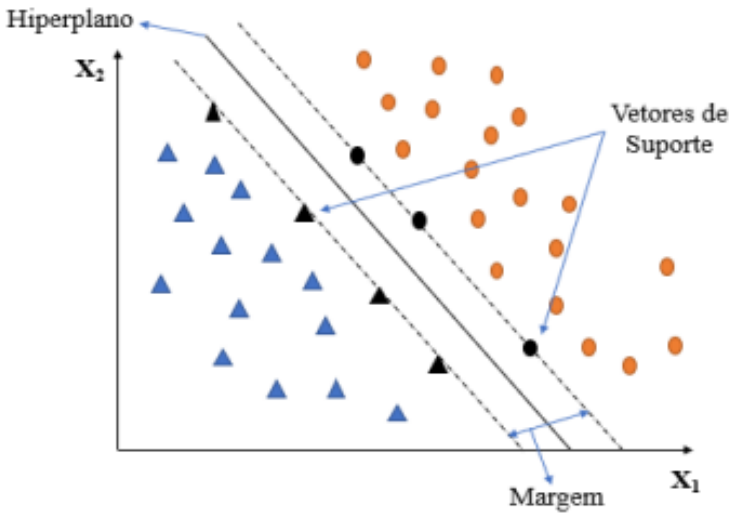
O algoritmo Máquina de Vetores de Suporte (SVM) é embasado pela Teoria de Aprendizado Estatístico proposta por Vapnik e Chervonenkis (1974) que estabelece vários princípios para o seu desenvolvimento. Esses princípios podem ser utilizados na obtenção de classificadores com boa generalização, ou seja, é a habilidade de prever novos casos a partir das hipóteses geradas durante o processo de aprendizagem. (BELLINI, 2020).

A proposta da SVM é a construção de um hiperplano de separação ótima. Tendo-se dados de treinamento rotulados, o algoritmo gera um hiperplano ideal que classifica novos dados. Em um espaço bidimensional, esse hiperplano é uma linha que separa em duas classes, onde cada classe fica em cada lado.

Um classificador SVM separa um conjunto de vetores de treinamento para duas classes diferentes (x_1, y_1), (x_2, y_2), ... , (x_m , y_m), onde $x_i \in R^d$ denota vetores em um espaço de características *d* -dimensional e $y_i \in \{-1, +1\}$ é um rótulo de classe. O modelo SVM é gerado mapeando os vetores de entrada em um novo espaço de recursos de dimensão superior denotado como $\Phi: R^d \rightarrow H^f$ onde $d < f$. Então, um hiperplano de separação ótimo no novo espaço de recursos é construído por uma função kernel $K(x_i, y_j)$, que é o produto dos vetores de entrada x_i e x_j e onde $K(x_i , x_j) = \Phi(x_i) \cdot \Phi(x_j)$ (HUANG *et al.*, 2017).

A Figura 9 representa o hiperplano de separação para um conjunto de treinamento linearmente separável, com os vetores de suporte e margem.

Figura 9 – representação de um conjunto com hiperplano linearmente separável



Fonte: Soares *et al* (2017, p.7).

A distância entre o hiperplano e o ponto mais próximo de cada classe refere-se a margem e, portanto, a superfície de decisão. Assim, todos os vetores situados em um lado do hiperplano são rotulados como -1 e, todos os vetores situados em outro lado são rotulados como +1. Os chamados vetores de suporte são as amostras do conjunto de treinamento que estão mais próximas do hiperplano, sendo assim as mais relevantes para classificação do modelo. O número desses vetores de suporte geralmente é pequeno em comparação com o tamanho do conjunto de treinamento (HUANG *et al*., 2017).

2.8.1 SVM com margens rígidas

As SVMs lineares com margens rígidas definem fronteiras lineares a partir de dados linearmente separáveis (LORENA; CARVALHO, 2007). Seja T um conjunto de treinamento com n dados $x_i \in X$ e seus respectivos rótulos $y_i \in Y$, em que X constitui o espaço dos dados e $Y = \{-1, +1\}$. T é linearmente separável se é possível separar os dados das classes +1 e −1 por um hiperplano (SMOLA; SCHOLKOPF, 2002).

Classificadores que separam os dados por meio de um hiperplano são denominados lineares (CAMPBELL, 2000). A equação de um hiperplano é apresentada na Equação 2, em

que w. x é o produto escalar entre os vetores w e x, w ∈ X é o vetor normal ao hiperplano descrito e $\frac{b}{\|\mathbf{w}\|}$ corresponde à distância do hiperplano em relação à origem, com b ∈ $\Re$.

$$f(x) = w \,.\, x + b = 0 \tag{2}$$

Essa equação divide o espaço dos dados X em duas regiões: w·x+b>0 e w·x +b<0. Uma função sinal g(x) = sgn(f(x)) pode então ser empregada na obtenção das classificações, conforme ilustrado na Equação 3 (SMOLA et al., 1999).

$$g(\mathbf{x}) = \operatorname{sgn}(f(\mathbf{x})) = \begin{cases} +1 \;\text{ se }\; \mathbf{w}\cdot\mathbf{x}+b>0 \\ -1 \;\text{ se }\; \mathbf{w}\cdot\mathbf{x}+b<0 \end{cases} \tag{3}$$

A partir de $f(x)$, é possível obter um número infinito de hiperplanos equivalentes, pela multiplicação de w e b por uma mesma constante (PASSERINI, 2004). Define-se o hiperplano canônico em relação ao conjunto T como aquele em que w e b são escalados de forma que os exemplos mais próximos ao hiperplano w · x + b = 0 satisfaçam a Equação 4 (MULLER et al., 2001).

$$|w.\, x_i + b| = 1 \tag{4}$$

Essa forma implica nas inequações 5, resumidas na Expressão 6.

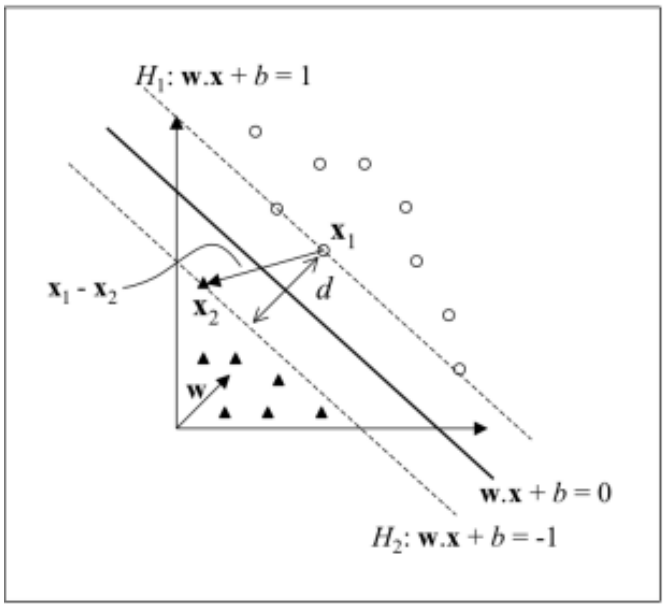
$$\begin{cases} \mathbf{w}\cdot\mathbf{x}_i + b \geqslant +1 & \text{se } y_i = +1 \\ \mathbf{w}\cdot\mathbf{x}_i + b \leqslant -1 & \text{se } y_i = -1 \end{cases} \tag{5}$$

$$y_i\,(w \,.\, x_i + b) - 1 \geq 0, \;\; \forall\, (x_i \,,\, y_i) \in T \tag{6}$$

Seja x_1 um ponto no hiperplano H_1: w · x + b = +1 e x_2 um ponto no hiperplano H_2: w · x + b = −1, conforme ilustrado na Figura 10. Projetando $x_1 - x_2$ na direção de w, perpendicular ao hiperplano separador w · x + b = 0, é possível obter a distância entre os hiperplanos H_1 e H_2 (CAMPBELL, 2000). Essa projeção é apresentada na Equação 7.

$$(\mathbf{x}_1 - \mathbf{x}_2)\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\cdot\frac{(\mathbf{x}_1 - \mathbf{x}_2)}{\|\mathbf{x}_1 - \mathbf{x}_2\|}\right) \tag{7}$$

Figura 10 – Distância entre os hiperplanos H_1 e H_2



Fonte: (LORENA; CARVALHO, 2007, p.54).

Tem-se que $w \cdot x_1 + b = +1$ e $w \cdot x_2 + b = -1$. A diferença entre essas equações fornece $w \cdot (x_1 - x_2) = 2$ (HEARST et al., 1998). Substituindo esse resultado na Equação 7, obtém-se:

$$\frac{2\,(\mathbf{x}_1 - \mathbf{x}_2)}{\|\mathbf{w}\|\,\|\mathbf{x}_1 - \mathbf{x}_2\|} \tag{8}$$

Como deseja-se obter o comprimento do vetor projetado, toma-se a norma da Equação 8, obtendo:

$$\frac{2}{\|\mathbf{w}\|} \tag{9}$$

Essa é a distância d, ilustrada na Figura 9, entre os hiperplanos H_1 e H_2, $\frac{1}{\|w\|}$ é a distância mínima entre o hiperplano separador e os dados de treinamento. Essa distância é definida como a margem geométrica do classificador linear (CAMPBELL, 2000). A partir das considerações anteriores, verifica-se que a maximização da margem de separação dos dados em relação a $w \cdot x + b = 0$ pode ser obtida pela minimização de $||w||$ conforme Equação 10

(BURGES, 1998). Dessa forma, recorre-se ao seguinte problema de otimização (SMOLA; SCHOLKOPF, 2002):

$$\underset{\mathbf{w},b}{\text{Minimizar}} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{10}$$

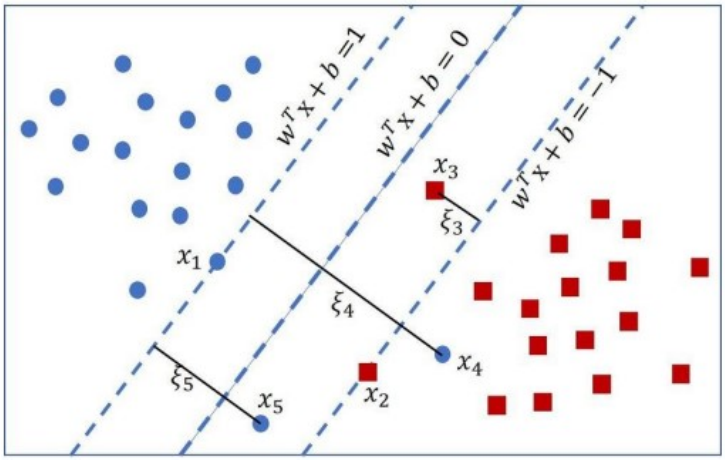
Com as restrições: $y_i\,(w \,.\, x_i + b) - 1 \geq 0, \;\; \forall_i = 1, \ldots, n$ (11)

As restrições são impostas de maneira a assegurar que não haja dados de treinamento entre as margens de separação das classes. Por esse motivo, a SVM obtida possui também a nomenclatura de SVM com margens rígidas (LORENA; CARVALHO, 2007).

2.8.2 SVM com margem suave

Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis. Isso se deve a diversos fatores, entre eles a presença de ruídos e *outliers* nos dados ou à própria natureza do problema, que pode ser não linear (LORENA; CARVALHO, 2007). Para amenizar tal problema, Cortes e Vapnik (1995), introduziram o conceito de variáveis de folga ξ_i, um erro associado a classificação de algumas amostras, essas variáveis relaxam as restrições impostas ao problema de otimização (LORENA; CARVALHO, 2007). A Figura 11 ilustra um possível caso de variável de folga.

Figura 11 – Ilustração de um possível caso das variáveis de folga em uma SVM de margem suave



Fonte: BOCCATO; ATTUX, s/d.

A aplicação desse procedimento suaviza as margens do classificador linear, porém com tolerância às violações da hipótese de separação (LORENA; CARVALHO, 2007). Dessa forma, as SVMs obtidas neste caso também podem ser referenciadas como SVMs com margens suaves.

Um erro no conjunto de treinamento é indicado por um valor de ξi maior que 1. Logo, a soma dos ξi representa um limite no número de erros de treinamento (BURGES, 1998). Para levar em consideração esse termo, minimizando assim o erro sobre os dados de treinamento, a função objetivo da Equação 10 é reformulada como:

$$\text{Minimizar} = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right) \tag{12}$$

Existe um parâmetro de custo (C), ou termo de regularização, que permite flexibilizar a separação das classes que controlam o compromisso entre possibilitar erros de treinamento ou forçar margens rígidas. Através desse parâmetro é possível criar uma margem flexível que torna possíveis alguns erros de classificação (BELLINI, 2020, p.23).

Para valores maiores de C, uma margem menor será aceita se a função de decisão for melhor em classificar corretamente todos os pontos de treinamento. Uma margem menor C, encorajará uma margem maior, portanto, uma função de decisão mais simples, ao custo da precisão do treinamento. Em outras palavras, um baixo valor de C admite classificação incorreta de muitas amostras e, um alto valor de C admite classificação correta das amostras, gerando um maior número de vetores de suporte.

2.8.3 Funções de Kernels

Outros parâmetros são introduzidos para classificação por meio do uso de SVMs, as chamadas funções de *Kernel*, que têm a finalidade de projetar os vetores de características de entrada em um espaço de características de alta dimensão para classificação de problemas que se encontram em espaços não linearmente separáveis (JUNIOR, 2010). Existem vários tipos de funções *Kernels*, porém as mais utilizadas são as Polinomiais, Gaussianos ou RBF (*Radial Basis Function*) e Sigmoidais, apresentados na Tabela 1.

Tabela 1 – Funções de Kernel mais utilizadas em SVM

Tipos de kernel	**função K (X_i, X_j)**	**Parâmetros**
Polinomial	$(\gamma\ (x_i\ .\ x_j) + K)^d$	γ, K e d
Gaussiano ou RBF	$Exp\ (\ -\sigma\ \|\|\ x_i - x_j\ \|\|^2)$	σ
Sigmoidal	$Tanh\ (\gamma\ (x_i\ .\ x_j) + K)$	γ e K

Fonte: Adaptado, (LORENA; CARVALHO, 2007, p.62).

Em um *Kernel* polinomial para SVM, os dados são mapeados em um espaço de dimensão superior usando uma função polinomial. O produto escalar dos pontos de dados no espaço original e a função polinomial no novo espaço é então obtido. Ao mapear os dados em um espaço de dimensão superior, o *Kernel* polinomial pode às vezes encontrar um hiperplano que separa as classes (SIDHARTH, 2022).

O *Kernel* utilizado no presente trabalho para o desenvolvimento do modelo de SVM por classificação foi o tipo RBF, é bastante utilizado para resolução de problemas de aprendizagem, inclusive é usado computacionalmente como padrão em muitas bibliotecas de linguagens de programação que utilizam o algoritmo SVM (JUNIOR, 2010). Na máquina RBF, é possível resolver problemas, originalmente, não linearmente separáveis, através do mapeamento para um espaço de maior dimensão.

3. METODOLOGIA

O presente trabalho foi desenvolvido em parceria com o Professor Dr. Frederico Garcia Pinto, da Universidade Federal de Viçosa, Campus Rio Paranaíba, que, durante seu Pós-doutorado supervisionado pelo Professor Dr. Timothy J. Garrett da Universidade da Flórida nos Estados Unidos, coletou as amostras de urina e realizou as análises por Cromatografia Líquida Acoplada a Espectrometria de Massas de Alta Resolução (LC-HRMS). Os dados dos espectros de massas foram encaminhados ao Laboratório de Quimiometria do Triângulo (LQT-UFU) para os tratamentos quimiométricos, com o consentimento e créditos a todos os pesquisadores participantes.

3.1 Amostras Biológicas

Amostras de urina de 40 pacientes com câncer de próstata foram obtidas no *Biospecimen* Core da rede SPORE em *Prostate Cancer* na Northwestern University (P50 CA180995). Os pacientes nesse grupo representaram vários estágios de progressão da doença. Amostras de urina de 40 indivíduos saudáveis sem histórico médico prévio de câncer foram obtidas no Life Study (University of Florida, Gainesville, FL). O conselho de revisão institucional (IRB) do Hospital da Flórida aprovou o uso das amostras. As amostras de urina foram coletadas usando tubos de preservação de urina (NorgenBioteck, Thorold, ON, Canadá) e armazenadas a -80°C até uso posterior para análise por (LC-HRMS).

3.2 Análise LC-HRMS

As análises LC-HRMS foram realizadas em um espectrômetro de massas de alta resolução da marca Thermo Scientific modelo Q Exactive interfaciado com um sistema LC Dionex Ultimate 3000 de alto desempenho. O espectrômetro de massas operou no modo negativo de ionização, em uma faixa de massa de 70-1000 m/z e uma resolução de massa de 35000. Usou-se uma coluna ACE Excel 2 C18-PFP (100 × 2,1 mm, 2,0 μm), com 2 μL de ácido fórmico 0,1% em H_2O (v/v) para solvente A e Acetonitrila para solvente B. A taxa de fluxo foi de 350 μL/min. Um gradiente linear foi iniciado após 3,0 min de 100% A a 20% A ao longo de 10,0 min e mantida a 20% A por 3,0 min antes de retornar às condições iniciais após 0,5 min. A coluna foi equilibrada nas condições iniciais por 4,0 min antes da próxima injeção. A temperatura da coluna era mantida a 25°C.

3.3 Análises Multivariadas

Os dados espectrais das amostras de urina dos pacientes foram representados numa matriz de dados **X** como ilustrado na Figura 5. Logo após foram tratados e pré-processados usando métodos e rotinas computacionais desenvolvidas no Laboratório de Quimiometria do Triângulo. Os métodos quimiométricos PLS-DA e SVM foram construídos usando o programa MATLAB versão R2021a (Mathworks, Inc.) e PLS_Toolbox versão 9.1 (Eigenvetor Research). Os dados foram normalizados pelo sinal de maior intensidade e os picos com intensidade menor que 5% dos sinais normalizados foram retirados, em seguida os dados da matriz **X** foram pré-processados utilizando o método de centragem na média, ou seja, centrar cada variável em relação a sua média.

A partir dos espectros de massas de 80 amostras de urina, sendo 40 de pacientes com câncer de próstata (CP) e 40 de paciente sadios, um modelo PLS-DA foi construído usando 48 amostras no conjunto de treinamento (24 com câncer e 24 sadios) e 32 amostras no conjunto teste (16 com câncer e 16 sadios). Já no modelo SVM, foram utilizados os mesmos dados dos espectros de massas de 80 amostras de urina, porém para o conjunto de treinamento foram usadas 50 amostras (25 com câncer e 25 sadios) e no conjunto teste 30 amostras (15 com câncer e 15 sadios). Ambos os modelos foram validados conforme os parâmetros de mérito da Tabela de Confusão.

3.4 Parâmetros de validação de modelos de classificação

Sempre que um modelo é proposto, existe o método de validação para comprovar e avaliar os valores gerados pelo modelo em relação a cada amostra predita. Dessa forma, os resultados são organizados numa Tabela de Confusão, onde os parâmetros de verdadeiro positivo (*vp*), verdadeiro negativo (*vn*), falso positivo (*fp*) e falso negativo (*fn*), sensibilidade, especificidade, eficiência e coeficiente de correlação de Matthew's, são calculados para avaliação do desempenho do modelo em relação às classes preditas (XIAOBO et al., 2010).

De acordo com as Equações 13 e 14 são calculadas as taxas de *fp* e *fn*, respectivamente, o percentual de amostras negativas classificadas como amostras positivas e o percentual de amostras positivas classificadas como amostras negativas.

$$\text{FP} = \frac{fp}{fp+vn} \text{ x } 100 \quad (13)$$

$$\text{FN} = \frac{fn}{fn+vp} \text{ x } 100 \quad (14)$$

Onde:

fp é o número de amostras falso positivas, ou seja, a probabilidade de uma amostra negativa ser classificada como uma amostra positiva;
vn é o número de verdadeira negativas, ou seja, a probabilidade de uma amostra negativa ser classificada como uma amostra negativa;
fn é o número de falso negativas, ou seja, a probabilidade de uma amostra positiva ser classificada como uma amostra negativa;
vp é o número de verdadeira positivas, ou seja, a probabilidade de uma amostra positiva ser classificada como uma amostra positiva.

Os parâmetros de sensibilidade e especificidade do modelo é definido de acordo com as Equações 15 e 16, verificando a efetividade do modelo em identificar amostras verdadeira positivas e amostras verdadeira negativas (XIAOBO et al., 2010). O coeficiente de correlação de Matthew's é calculado a partir da Equação 17, resulta em valores como (+1), zero e (-1), que representa classificação perfeita, aleatória e inversa, respectivamente. A eficiência é calculada de acordo com a Equação 18. Os resultados dos parâmetros ditos, são apresentados na Tabela de Confusão.

$$Sensibilidade = \frac{vp}{vp+fn} \quad (15)$$

$$Especificidade = \frac{vn}{vn+fp} \quad (16)$$

$$Coeficiente\ de\ Matthew\ = \frac{(vp\ x\ vn) - (fp\ x\ fn)}{\sqrt{(vp+fp)x\ (vp+fn)x\ (vn+fp)x\ (vn+fn)}} \quad (17)$$

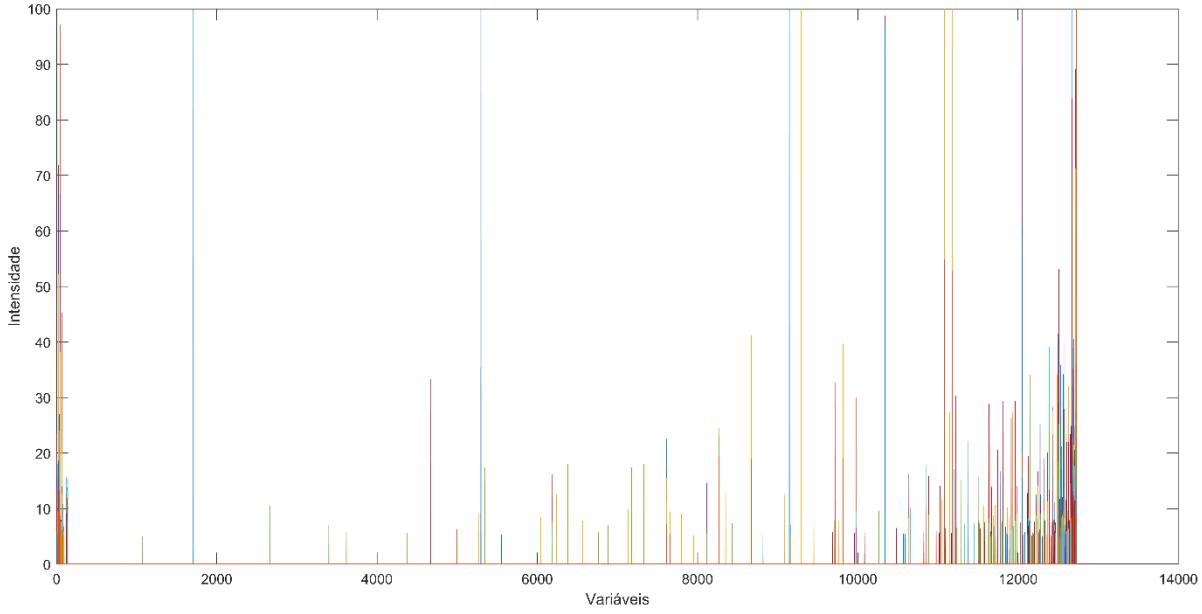
$$Acurácia = \frac{vp+vn}{vp+vn+fp+fn} \quad (18)$$

4. RESULTADOS E DISCUSSÕES

4.1 Espectros de massas das amostras analisadas por LC-HRMS

Os espectros de massas normalizados apresentados na Figura 12 representam as 80 amostras de urina, sendo 40 amostras de indivíduos sadios e 40 de indivíduos com câncer de próstata. Os mesmos foram usados para a construção dos modelos PLS-DA e SVM nas etapas de calibração (treinamento) e validação (teste). Os espectros de cada amostra, realizados em modo de ionização negativo, mostraram os picos em razão massa/carga na faixa de 70-1000 m/z.

Figura 12 - Gráfico dos espectros de massas obtidos pela técnica LC-HRMS das amostras de urina de pacientes com CP e sadios



Fonte: a autora.

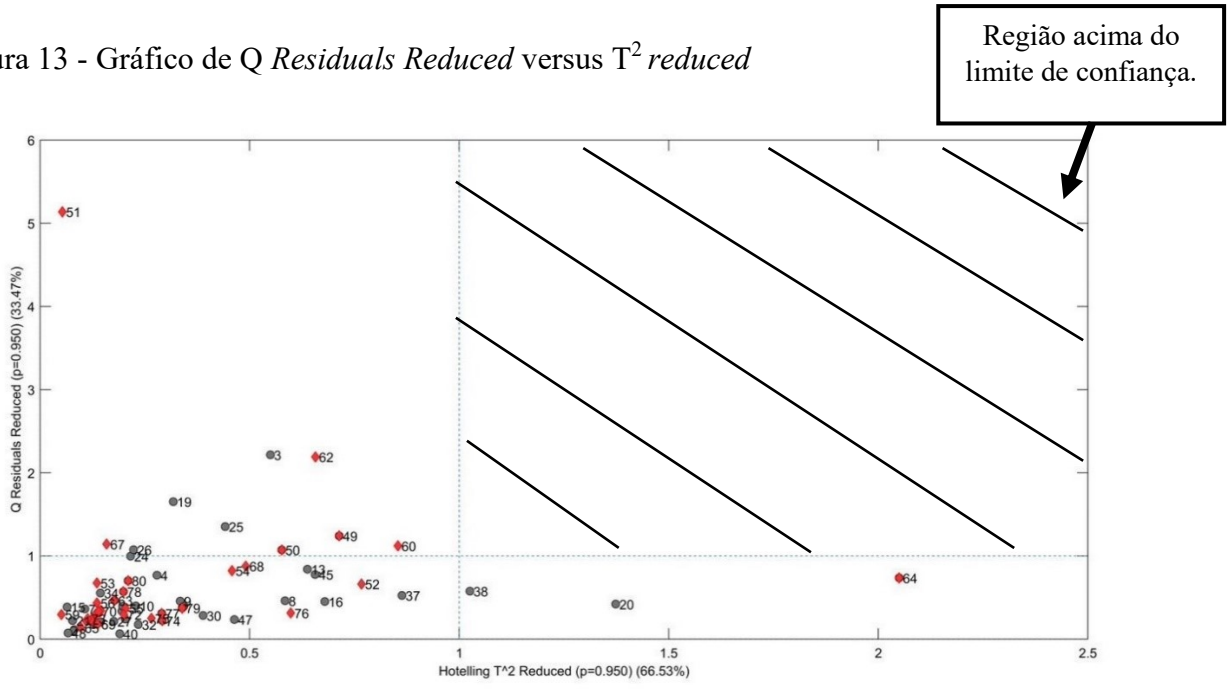
Observa-se a complexidade dos espectros, dificultando a análise visual e, consequentemente, a separação dos pacientes sadios com os pacientes de câncer. A partir daí os métodos quimiométricos são aplicados para extrair o máximo da informação química relevante, com a análise dos dados.

4.2 Modelo PLS-DA

O modelo foi construído da seguinte forma: classe 1 (classe de interesse) encontram as amostras de urinas de pacientes com câncer de próstata, classe 0 (classe de não interesse) as amostras de urinas de pacientes sadios. O modelo foi construído com 5 variáveis latentes (VL) e capturou 66,53% e 33,47% de variância explicada nos blocos **X** e **y**, respectivamente. A validação cruzada foi realizada por janelas venezianas (*venetian blinds*) utilizando 18 janelas com 1 amostra.

A detecção de amostras anômalas (*outlier*) foi realizada através do gráfico Q *Residuals Reduced* versus T^2 *reduced* (Figura 14). O resíduo Q é empregado na avaliação da conformidade de cada amostra no modelo calibrado, medindo a diferença dessa amostra no dado original e no dado que foi calculado no modelo. O resíduo de *Hotelling,* T^2, baseia-se na soma do quadrado dos escores normalizados, medindo a distância ou variação de cada amostra em relação ao centro do modelo. Dessa forma, quando o resíduo tem um valor alto, indica que a amostra não está conforme (no caso onde está essa janela robusta). Para que seja possível a exclusão das amostras com alto resíduo, que exercem uma influência no modelo, um limite de confiança deve ser estabelecido para ambos os parâmetros, geralmente sendo de 95% e, portanto, excluídos. Com base na Figura 13, nenhuma amostra foi considerada anômala em toda a região espectral avaliada.

Figura 13 - Gráfico de Q *Residuals Reduced* versus T^2 *reduced*

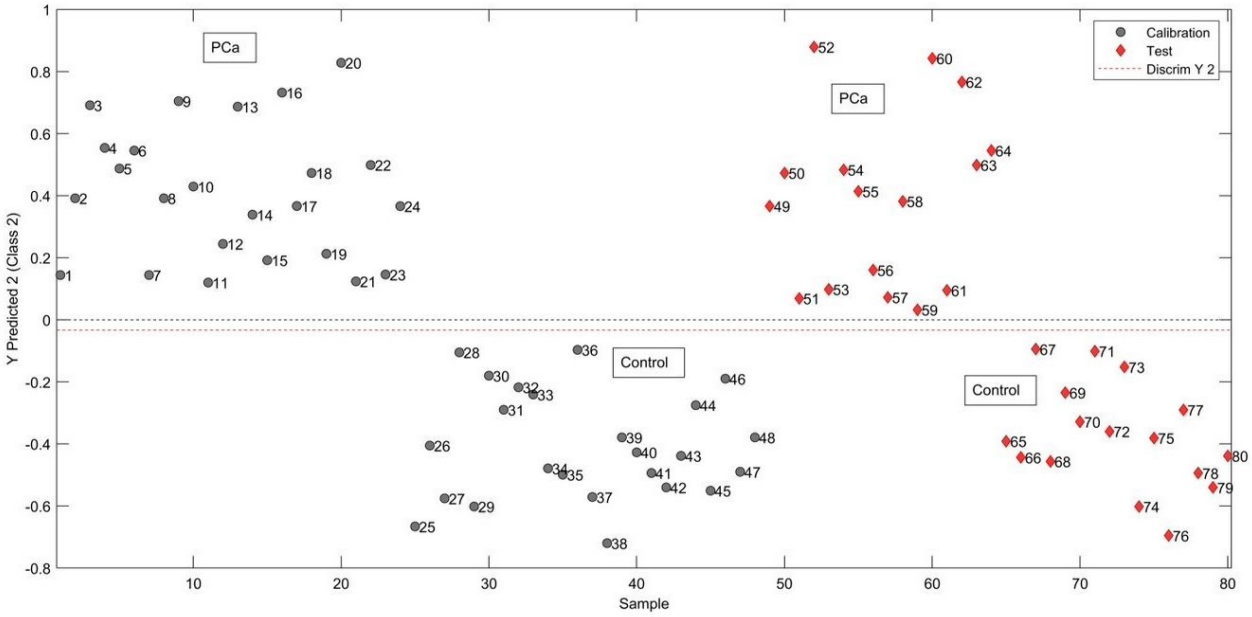


Fonte: a autora.

A Figura 14 apresenta o gráfico de estimativas obtido no modelo PLS-DA, onde, as amostras de 1 a 48 são as amostras da etapa de treinamento (●), isto é, amostras que foram utilizadas para a construção do modelo. As amostras de 49 a 80 (totalizando 32) são as amostras do conjunto de teste (♦), para avaliar se o modelo construído consegue prever amostras desconhecidas.

Observa-se claramente a classificação das amostras por classe, sendo classe 1 (classe de interesse) as amostras de pacientes com câncer de próstata posicionadas acima da linha *threshold* (linha pontilhada vermelha) e classe 0 (classe de não interesse) amostras de pacientes sadios posicionadas abaixo da linha *threshold*.

Figura 14 - Gráfico de estimativas para amostras de urinas de pacientes com câncer de próstata e sadios PLS-DA



Fonte: a autora.

Dessa forma, avaliando a exatidão do modelo através dos valores de erros quadráticos médios de calibração (RMSEC), validação cruzada (RMSECV), previsão (RMSEP), especificidade e sensibilidade, apresentados na Tabela 2, o modelo apresenta uma exatidão satisfatória. O valor do erro de calibração, foi menor que o erro de previsão, transmitindo assim uma confiabilidade na calibração.

Tabela 2 – Parâmetros de classificação obtidos pelo modelo PLS-DA para as amostras de urina dos pacientes saudáveis e com câncer.

PARÂMETROS	**VALORES OBTIDOS (%)**
RMSEC	0,54
RMSECV	0,64
RMSEP	0,57
Sensibilidade (Cal, CV, Prev)	1,0
Especificidade (Cal, CV, Prev)	1,0
Variância em **X**	66,53
Variância em **y**	33,47

Cal = calibração; CV = validação cruzada; Prev = previsão.

Fonte: a autora.

Segundo os parâmetros de mérito presentes na Tabela de Confusão apresentada na Tabela 3, observa-se que o modelo apresentou uma classificação de 100% das amostras, mostrando sua eficiência. Tanto na etapa de treinamento, quanto na etapa de teste, as amostras pertencentes a classe 1 e classe 0 foram classificadas corretamente. Ou seja:

- As 24 amostras de pacientes com câncer de próstata na etapa de treinamento foram previstas na classe 1 e nenhuma amostra de pacientes sadios foi prevista na classe 1, assim como todas as 24 amostras de sadios foram previstas na classe 0 e nenhuma amostra de pacientes com câncer de próstata foi prevista na classe 0;
- As 16 amostras de pacientes com câncer de próstata na etapa de teste foram previstas na classe 1 e nenhuma amostra de pacientes sadios foi prevista na classe 1, assim como todas as 16 amostras de sadios foram previstas na classe 0 e nenhuma amostra de pacientes com câncer de próstata foi prevista na classe 0;

- A eficiência em ambas as etapas, apresentaram valores igual a 1 (um) onde representa um ótimo desempenho do modelo;
- O coeficiente de correlação Matthew's também apresentou valores igual a +1 (um) em ambas as etapas, representando classificação perfeita do modelo.

Tabela 3 – Tabela de Confusão do modelo PLS-DA

Tabela de confusão						
		VP	**FP**	**VN**	**FN**	**N**
Treinamento	Classe 1	100	0	100	0	24
	Classe 0	100	0	100	0	24
Eficiência = 1 Coeficiente de Correlação Matthew's = +1						
		VP	**FP**	**VN**	**FN**	**N**
Teste	Classe 1	100	0	100	0	16
	Classe 0	100	0	100	0	16
Eficiência = 1 Coeficiente de Correlação Matthew's = +1						

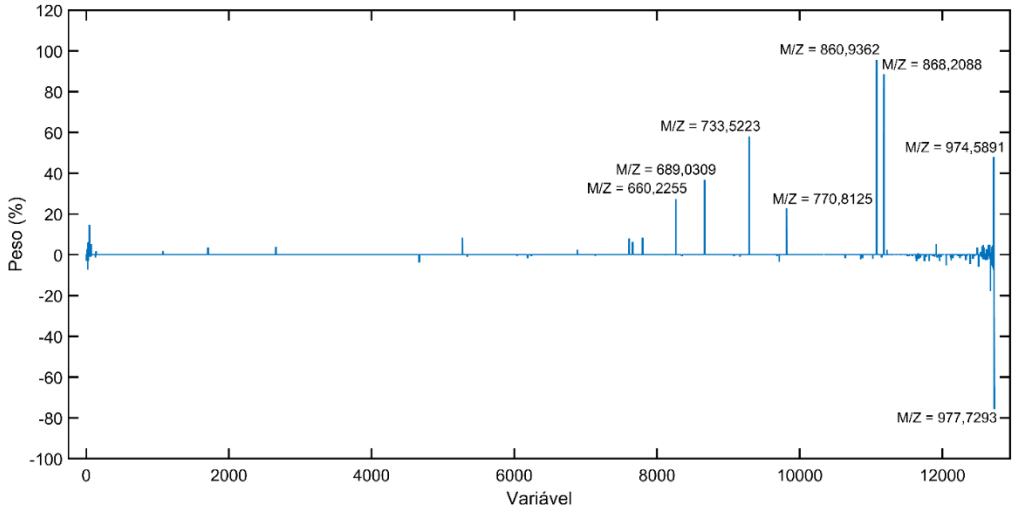
VP = taxa de verdadeiro positivo; FP = taxa de falso positivo;

VN = taxa de verdadeiro negativo; FN = taxa de falso negativo; N = número de amostras.

Fonte: a autora.

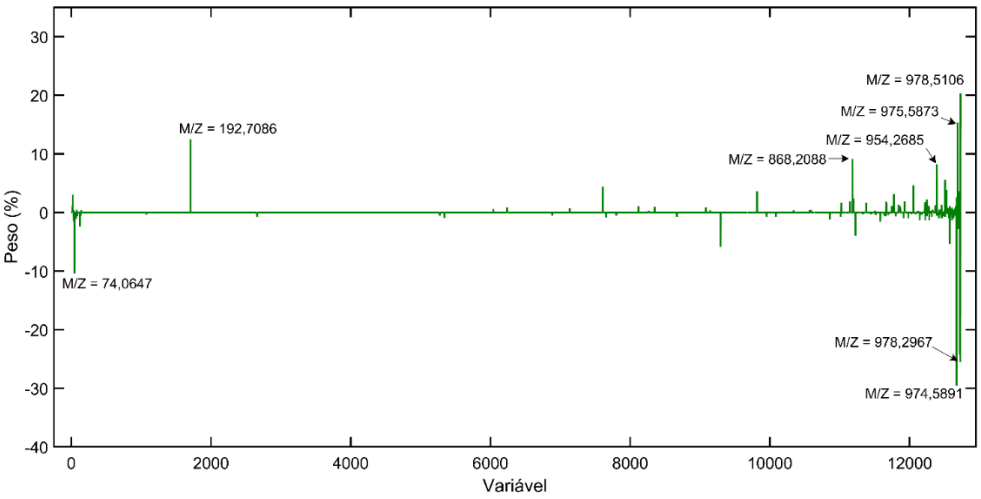
A análise dos gráficos dos pesos mostra as variáveis que foram mais significativas para a classificação das amostras, onde se diferenciou as dos pacientes com câncer de próstata e as dos pacientes sadios. As razões m/z das variáveis mais significativas para as amostras de urina com câncer de próstata foram: 680,2255; 689,0309; 733,5223; 770,8125; 860,9362; 868,2088; 974,5891 e 977,7293; como representadas na Figura 15. Dessa forma, as variáveis mais significativas para as amostras de urina de pacientes sadios foram: 74,0647; 192,7086; 868,2088; 954,2685; 974,5891; 975,5873; 978,2967 e 978,5106; representadas na Figura 16. Assim, esses íons podem levar as moléculas dos compostos a serem testados como biomarcadores para o câncer de próstata, onde serão passados para os professores parceiros Dr. Frederico Garcias e Dr. Thimoty J. Garret para os possíveis testes.

Figura 15 – Gráfico das variáveis mais significativas dos pacientes com câncer de próstata



Fonte: a autora.

Figura 16 – Gráfico das variáveis mais significativas dos pacientes sadios



Fonte: a autora.

A partir da análise dos gráficos dos pesos representados na Figura 15 e 16, observa-se diferenças entre as intensidades das variáveis mais significativas dos pacientes com câncer de próstata e pacientes sadios.

Note que, a variável com razão m/z = 868,2088 tanto no gráfico de pacientes com câncer de próstata tanto de pacientes sadios, apresentam intensidades e pesos diferentes. Enquanto o peso sobre a variável no gráfico de pacientes com câncer de próstata ficou em torno de 80 a 100%, em pacientes sadios a intensidade da variável foi menor, com um peso de aproximadamente 10%.

O mesmo acontece com a variável razão m/z = 974, 5891 no gráfico de pacientes com câncer apresentou um peso em torno de 40 a 60%, enquanto no gráfico de pacientes sadios a intensidade da variável foi negativa, apresentando um peso de aproximadamente -30%. Assim, não importa se o peso apresentou valores positivos e/ou negativos, o que importa é a intensidade.

Dessa forma, as variáveis dadas como exemplo, como as outras significativas destacadas nos gráficos, irão ser testadas como possíveis biomarcadores para detectar câncer de próstata através da urina.

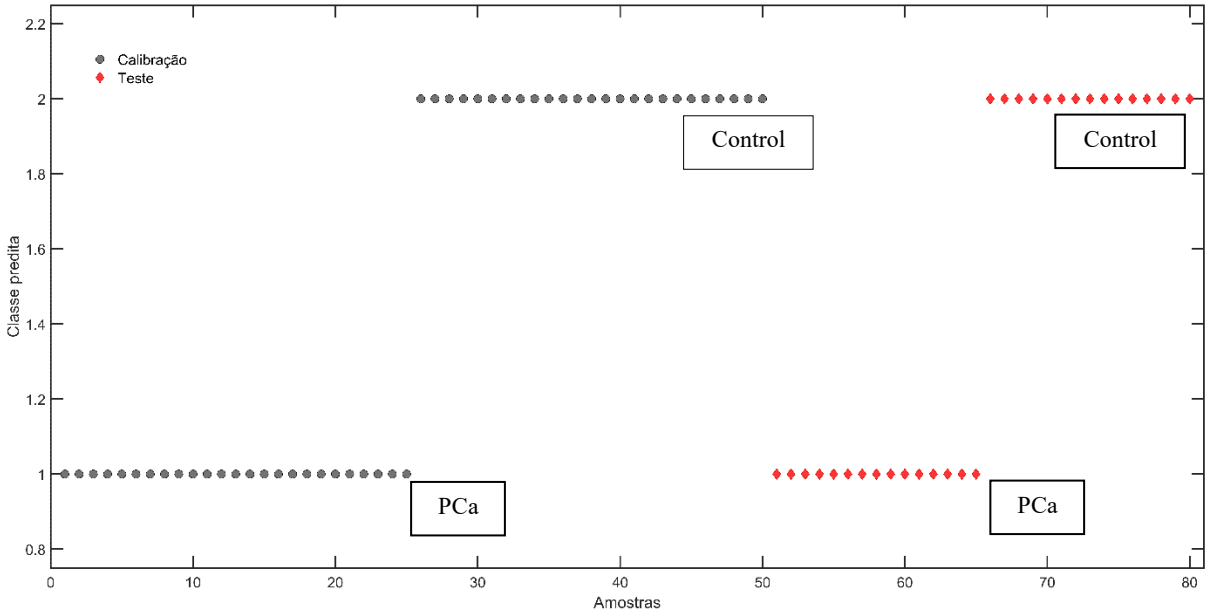
4.3 Modelo SVM

O modelo foi construído da seguinte forma: classe 1 (classe de interesse) encontra-se as amostras de urina de pacientes com câncer de próstata, classe 2 (classe de não interesse) as amostras de urina de pacientes sadios. Possui 23 vetores de suporte (SVs), onde desempenham um papel importante na classe de máquinas de aprendizagem, pois representam os pontos em que os dados se encontram mais próximos da superfície de decisão. A validação cruzada foi realizada por janelas venezianas (*venetian blinds*) utilizando 20 janelas com 1 amostra.

A Figura 17 apresenta o gráfico de estimativas obtido no modelo, onde, as amostras de 1 a 50 são as amostras de treinamento (●), isto é, amostras que foram utilizadas para a construção do modelo. As amostras de 51 a 80 (totalizando 30) são as amostras do conjunto de teste (♦) para avaliar se o modelo construído consegue prever as amostras de classe desconhecidas.

Observa-se claramente a classificação das amostras por classe, sendo classe 1 (classe de interesse) as amostras de pacientes com câncer de próstata e classe 2 (classe de não interesse) amostras de pacientes sadios.

Figura 17 - Gráfico de estimativas para amostras de urinas de pacientes com câncer de próstata e sadios SVM



Fonte: a autora.

A Tabela 4 apresenta os parâmetros obtidos pelo modelo SVM. O *Kernel* utilizado foi o tipo RBF, para isso, existem dois parâmetros em busca de um melhor resultado para o aprendizado do classificador: γ (gamma) e C (custo) onde seus valores estão representados na Tabela 4. Pode-se observar um alto valor da constante C (custo), essa constante controla os erros de classificação do algoritmo SVM, com esse valor alto de C a função de classificação torna-se menos suave para que as amostras sejam classificadas erroneamente, podendo gerar um maior número de vetores de suporte.

Tabela 4 - Parâmetros de classificação obtidos pelo modelo SVM

PARÂMETROS	**VALORES OBTIDOS (%)**
Especificidade (Cal, CV, Prev)	1,0
Custo (C)	100
Gamma (γ)	$3{,}1623 \times 10^{-6}$

Cal = calibração; CV = validação cruzada; Prev = previsão.

Fonte: a autora.

Segundo os parâmetros de mérito presentes na Tabela de Confusão apresentada na Tabela 5, observa-se que o modelo apresentou uma classificação de 100% das amostras, mostrando sua eficiência. Tanto na etapa de treinamento, quanto na etapa de teste, as amostras pertencentes a classe 1 e classe 2 foram classificadas corretamente. Ou seja:

- As 25 amostras de pacientes com câncer de próstata na etapa de treinamento foram previstas na classe 1 e nenhuma amostra de pacientes sadios foi prevista na classe 2, assim como todas as 25 amostras de sadios foram previstas na classe 2 e nenhuma amostra de pacientes com câncer de próstata foi prevista na classe 2;
- As 15 amostras de pacientes com câncer de próstata na etapa de teste foram previstas na classe 1 e nenhuma amostra de pacientes sadios foi prevista na classe 2, assim como todas as 15 amostras de sadios foram previstas na classe 2 e nenhuma amostra de pacientes com câncer de próstata foi prevista na classe 2;
- A eficiência em ambas as etapas, apresentaram valores igual a 1 (um) onde representa um ótimo desempenho do modelo;
- O coeficiente de correlação Matthew's também apresentou valores igual a +1 (um) em ambas as etapas, representando classificação perfeita do modelo.

Tabela 5 – Tabela de Confusão do modelo SVM

Tabela de confusão						
		VP	**FP**	**VN**	**FN**	**N**
Treinamento	Classe 1	100	0	100	0	25
	Classe 2	100	0	100	0	25
Eficiência = 1 Coeficiente de Correlação Matthew's = +1						
		VP	**FP**	**VN**	**FN**	**N**
Teste	Classe 1	100	0	100	0	15
	Classe 2	100	0	100	0	15
Eficiência = 1 Coeficiente de Correlação Matthew's = +1						

VP = taxa de verdadeiro positivo; FP = taxa de falso positivo;

VN = taxa de verdadeiro negativo; FN = taxa de falso negativo; N = número de amostras.

Fonte: a autora.

A aplicação de cromatografia líquida acoplada a espectrometria de massas associada aos métodos quimiométricos PLS-DA e SVM permitiu o desenvolvimento de métodos de classificação para o diagnóstico não invasivo do câncer de próstata utilizando informações relevantes dos espectros de massas das amostras de urina dos pacientes. A eficiência de tais métodos foi avaliada com base nos parâmetros de exatidão, sensibilidade e especificidade que apresentaram classificação correta das amostras do conjunto de treinamento e conjunto de teste de 100% para os dois modelos. A investigação de biomoléculas e alterações moleculares relacionadas ao desenvolvimento do câncer é fundamental na busca de novos biomarcadores e métodos alternativos que possam complementar os exames preventivos e auxiliar na detecção precoce da doença.

5. CONCLUSÃO

A técnica de cromatografia líquida acoplada a espectrometria de massas (LC-HRMS) associada aos métodos quimiométricos PLS-DA e SVM, mostrou-se eficiente para detecção do câncer de próstata através da urina, extraindo de maneira satisfatória informações complexas contidas nos espectros de massas.

O uso dos métodos quimiométricos de análise discriminante por quadrados mínimos parciais (PLS-DA) e máquina de vetores de suporte (SVM) para extrair as informações contidas nos espectros de massas permitiu o desenvolvimento de modelos estáveis, robustos e de fácil interpretação, capazes de discriminar amostras de pacientes com câncer de próstata e pacientes sadios com eficiência de 100%.

O método quimiométrico PLS-DA, devido fornecer através do gráfico dos pesos as variáveis mais significativas, se torna o melhor a ser aplicado na busca de novos biomarcadores e métodos alternativos na detecção precoce da doença.

Os métodos analíticos desenvolvidos, aliando a técnica de LC-HRMS aos métodos quimiométricos, tornam promissores para atender a demanda quanto ao diagnóstico do câncer de próstata usando uma amostragem não invasiva, no caso a coleta da urina do paciente.

A partir dos íons identificados na análise dos pesos torna-se possível descobrir as fórmulas químicas de compostos que poderão ser usados como possíveis biomarcadores, a partir daí após testes poderemos ter o desenvolvimento de testes rápidos a serem comercializados em farmácias, semelhantes aos exames de gravidez e glicosímetros.

6. REFERÊNCIAS BIBLIOGRÁFICAS

AGIN, A.; HEINTZ, D.; RUHLAND, E.; BARCA, J. M. C.; ZUMSTEG, J.; MOAL, V.; GAUCHEZ, A. S.; NAMER, I. J. Metabolomics - an overview. From basic principles to potential biomarkers (Part 1). Médecine Nucéaire. v. 40, p. 4-10, 2016. https://doi.org/10.1016/j.mednuc.2015.12.006

ALMEIDA, M. R.; FIDELIS, C. H. V.; BARATA, L. E. S.; POPPI, R. J. Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation. Talanta, v. 117, p. 305-311, 2013. https://doi.org/10.1016/j.talanta.2013.09.025

ALBERICE, J. V. Avaliação analítica de potenciais biomarcadores para câncer de bexiga em urina. Tese apresentada ao Programa de Pós-Graduação em Química do Instituto de Química de São Carlos - USP, como requisito para obtenção de título de Doutor em Química, São Carlos, 2014.

BARKER, M.; RAYENS, W. Partial least squares for discrimination. Journal of Chemometrics, Chichester, v. 17, p. 166-173, 2003. https://doi.org/10.1002/cem.785

BELINI, R. Aplicação de máquinas de suporte vetorial na classificação textual. Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias da Universidade de Caxias do Sul. Caxias do Sul, 2020.

BOCCATO, L.; ATTUX, R. Máquina de Vetores-Suporte (SVMs): Parte II. São Paulo, s/d. Apresentação em slide share. 4 slides. Disponível em:< https://www.dca.fee.unicamp.br/~lboccato/topico_7.2_SVM.pdf>. Acesso em: 27 de jun. 2023.

BRAGA, S. L. GUIA PARA DESENVOLVIMENTO E OTIMIZAÇÃO DE MÉTODOS POR CROMATOGRAFIA A LÍQUIDO DE ALTA EFICIÊNCIA PARA QUANTIFICAÇÃO DE FÁRMACOS. Trabalho de conclusão de curso apresentado como parte dos requisitos para a obtenção do grau de Bacharel em Farmácia ao curso de Farmácia da Universidade Federal de Ouro Preto. Ouro Preto, 2018.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining, v. 2, p. 1-43, 1998.

BUSTILLOS, O. V. A cromatografia líquida acoplada à espectrometria de massas em Tandem HPLC-MS/MS, Revista Analytica, v. 18, ed. 106, p. 34-35, 2020.

CAMPBELL, C. An introduction to kernel methods. In R. J. Howlett and L. C. Jain, editors, Radial Basis Function Networks: Design and Applications, Springer Verlag, p. 155-192, Berlin, 2000.

CANUTO, G. A. B.; COSTA, J. L.; CRUZ, P. L. R.; SOUZA, A. R. L.; FACCIO, A. T.; KLASSEN, A.; RODRIGUES, K. T.; TAVARES, M. F. M. Metabolômica: Definições, Estado-da-Arte e Aplicações Representativas. Quimica Nova, v. 41, p. 75-91, 2018. https://doi.org/10.21577/0100-4042.20170134

CORTES, C.; VAPNIK, V. N. Support-Vector Networks. Machine Learning. v. 20, p. 273-297, 1995. https://doi.org/10.1007/BF00994018

DAMIÃO, R.; FIGUEIREDO R. T.; DORNAS, M. C.; LIMA, D. S.; KOSCHORKE, M. A. B. Câncer de próstata. Revista Hospital Universitário Pedro Ernesto, [S.l.], v. 14, ago. 2015. https://doi.org/10.12957/rhupe.2015.17931

DIAS, J. S. Próstata, o que é e qual sua função? Instituto da próstata, s/d. Disponível em: <https://www.institutodaprostata.com/pt/blog/prostata-o-que-e-e-qual-a-sua-funcao>. Acesso em: 21 jun. 2023.

DUNN, W. B.; BROADHURST, D.; BEGLEY, P.; ZELENA, E.; FRANCIS-MCLNTYRE, S.; ANDERSON, N.; BROWN, M.; KNOWLES, J. D.; HALSALL, A.; HASELDEN, J. N.; NICHOLLS, A. W.; WILSON, I. D.; KELL, D. B.; GOODACRE, R. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat. Protocols, v. 6, p. 1060-1083, 2011. https://doi.org/10.1038/nprot.2011.335

DURANS, A. F. F. APLICAÇÃO DE MÉTODOS QUIMIOMÉTRICOS E ESPECTROMETRIA DE MASSA PARA DETECTAR CÂNCER DE PRÓSTATA ATRAVÉS DA URINA. Dissertação apresentada ao Programa de Pós-Graduação em Química do Instituto de Química da Universidade Federal de Uberlândia, como requisito para obtenção de título de Mestre em Química. Uberlândia, 2022.

FERREIRA, M. T. C. Uso do FT-MIR e calibração multivariada por MCR-ALS e SVR na determinação do teor de bioquerosene de macaúba e palmiste em misturas com querosene de aviação. Dissertação apresentada ao Programa de Pós-graduação em Química do Instituto de Química da Universidade Federal de Uberlândia, como requisito para obtenção de Título de Mestre em Química. Uberlândia, 2021.

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. Quimiometria I: Calibração multivariada, um tutorial. Quím. Nova, v. 22, n. 5, 724-731, 1999. https://doi.org/10.1590/S0100-40421999000500016

GOULART, V. A. M. METABOLÔMICA APLICADA NA PESQUISA DE BIOMARCADORES PARA ACIDENTE VASCULAR ENCEFÁLICO ISQUÊMICO, INFARTO DO MIOCÁRDIO E ESQUIZOFRENIA. Tese apresentada ao Programa de pós-graduação em Biologia Celular do Departamento de Morfologia, do Instituto de Ciências Biológicas, da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de doutor em Ciências. Belo Horizonte, 2018.

HEARST, M. A.; SCHOLKOPF, B.; DUMAIS, S.; OSUNA, E.; PLATT, J. Trends and controversies - support vector machines. IEEE Intelligent Systems, v. 13, n. 4, p. 18-28, 1998. https://doi.org/10.1109/5254.708428

HUANG, M. W.; CHEN, C. W.; LIN, W.C.; KE, S.W.; TSAI, C. F. SVM e SVM Ensembles in Breast Cancer Prediction. PLoS ONE, v. 12, n. 1, 2017. https://doi.org/10.1371/journal.pone.0161501

INCA, Instituto Nacional do Câncer. Disponível em:<https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/prostata>. Acesso em: 15 de jun. 2023.

INCA, Instituto Nacional de Câncer (Brasil). Estimativa 2023: incidência de câncer no Brasil/ Instituto Nacional de Câncer. Rio de Janeiro: INCA, 2022. Disponível em:< https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil>. Acesso em: 15 de jun. 2023.

JANISCH, B. D. DETERMINAÇÃO MULTICLASSE DE AGROTÓXICOS EM CAMAS BIOLÓGICAS EMPREGANDO CROMATOGRAFIA LÍQUIDA ACOPLADA A ESPECTROMETRIA DE MASSAS. Dissertação apresentada ao Programa de Pós-Graduação em Química, da Universidade Federal de Santa Maria, com requisito parcial para obtenção do título de Mestre em Química. Rio Grande do Sul, 2022.

JUNIOR, G. M. O. Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado. Trabalho apresentado como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco. Pernambuco, 2010.

KONZEN, R. A. VALIDAÇÃO DE PARÂMETROS DE MÉRITO PARA QUANTIFICAÇÃO DE MICROCISTINA-LR POR MEIO DE CROMATOGRAFIA LÍQUIDA DE ALTA EFICIÊNCIA ACOPLADO A DETECTOR DE ARRANJO DE DIODOS. Trabalho de conclusão do curso de Bacharelado em Química do Departamento de Química e Biologia - DAQBi, da Universidade Tecnológica Federal do Paraná - UTFPR. Curitiba, 2015.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. RITA, v. 14, n. 2, p. 43-67, 2007. https://doi.org/10.22456/2175-2745.5690

MADSEN, R.; LUNDSTEDT, T.; TRYGG, T. Chemometrics in metabolomics - A review in human disease diagnosis. Analytica Chimica Acta. v. 659, p. 23-33, 2010. https://doi.org/10.1016/j.aca.2009.11.042

MÁQUINA, A. D. V. ESPECTROMETRIA NO INFRAVERMELHO MÉDIO E MÉTODOS QUIMIOMÉTRICOS PLS-DA e PLS: CLASSIFICAÇÃO E PREVISÃO DO TEOR DE BIODIESEL NA MISTURA DE BIODIESEL/DIESEL DE MAFURRA, MORINGA E ALGODÃO. Dissertação apresentada ao Programa de Pós-Graduação em Química do Instituto

de Química da Universidade Federal de Uberlândia, como requisito para obtenção de título de Mestre em Química. Uberlândia, 2017.

MASSART, D. L.; VANDEGINSTE, B. G. M.; BUYDENS, L. M. C.; JONG, S.; LEWI, P. J.; SMEYERS-VERBEKE, J. Handbook of Chemometrics and Qualimetrics. Amsterdã: Elsevier Science, v. 20, 1998.

MONGOLLÓN, N. G. S.; LIMA, P. F.; GAMA, M. R.; FURLAN, M. F. State of the art two-dimensional liquid chromatography: fundamental concepts, instrumentation, and applications. Química Nova, v. 37, n.10, p. 1680-1691, 2014. https://doi.org/10.5935/0100-4042.20140261

MULLER, K. R.; MIKA, S.; RATSCH, G.; TSUDA, K.; SCHOLKOPF, B. An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks, v. 12, n. 2, p. 181-201, 2001. https://doi.org/10.1109/72.914517

NASCIMENTO, D. S. DETERMINAÇÃO SIMULTÂNEA DE Cu, Pb, Cd, Ni, Co e Zn EM ETANOL COMBUSTÍVEL POR VOLTAMETRIA DE REDISSOLUÇÃO ADSORTIVA E CALIBRAÇÃO MULTIVARIADA. Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal da Paraíba, como requisito para obtenção de título de Mestre em Química. João Pessoa, 2013.

PAIVA, Rodrigo Minuto. AVALIAÇÃO DE MICRORNAS COMO BIOMARCADORES MOLECULARES NO CÂNCER DE PRÓSTATA. Tese apresentada ao Programa de Pós-Graduação em Ciências Biológicas: Fisiologia do Instituto de Ciências Básicas da Saúde da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de doutor em Fisiologia. Porto Alegre, 2020.

PASSERINI, A. Kernel Methods, multiclass classification and applications to computational molecular biology. Tese PhD, Università Degli Studi di Firenze, 2004.

PAULA, R. O. MÁQUINAS DE SUPORTE VETORIAL COMO INSTRUMENTO DE PRIORIZAÇÃO DE INVESTIMENTOS APLICADO AO PROGRAMA DE INVESTIMENTO E LOGÍSTICA - AEROPORTOS. Dissertação apresentada como requisito para obtenção do grau de Mestre pelo Programa de Mestrado em Economia do Setor Público, do Departamento de Economia da Faculdade de Administração, Contabilidade e Economia, Universidade de Brasília e Escola de Administração Fazendária do Ministério da Fazenda. Brasília, 2016.

PAZ, R. A. D. DETERMINAÇÃO SIMULTÂNEA, POR CROMATOGRAFIA LÍQUIDA DE ALTA RESOLUÇÃO, DAS VITAMINAS A, D e E EM AMOSTRAS COMPOSTAS POR DIFERENTES MATRIZES ALIMENTARES. Dissertação apresentada para obtenção do grau de Mestre em Engenharia Química e Biológica do Instituto Superior de Engenharia de Lisboa. Lisboa, 2018.

SARGAÇO, B. R. OTIMIZAÇÃO E VALIDAÇÃO DE UM MÉTODO DE CROMATOGRAFIA LÍQUIDA DE ALTA RESOLUÇÃO (HPLC) PARA A DETERMINAÇÃO DO EDULCORANTE CICLAMATO: OCORRÊNCIA EM ADOÇANTES DE MESA. Trabalho final para obtenção do grau de Mestre em Engenharia Química e Biológica do Instituto Superior de Engenharia de Lisboa. Lisboa, 2013.

SARRIS, A. B.; CANDIDO, F. J. L. F.; FILHO, C. R. P.; STAICHAK, R. L.; TORRANI, A. C. K.; SOBREIRO, B. P. Câncer de próstata: uma breve revisão atualizada. Visão Acadêmica, [S.l.], v. 19, n. 1, 2018. https://doi.org/10.5380/acd.v19i1.57304

SIDHARTH. SVM Kernels: Polynomial Kernel - From Scratch Using Python. PycodeMates.com, dez. 2022.

SMITI, B. A critical overview of outlier detection methods. Computer Science Review, v. 38, 2020. https://doi.org/10.1016/j.cosrev.2020.100306

SMOLA, A. J.; BARLETT, P.; SCHOLKOPF, B.; SCHUURMANS, D. Introduction to large margin classifiers. In SMOLA, A. J.; BARLETT, P.; SCHOLKOPF, B.; SCHUURMANS, D, editors, Advances in Large Margin Classifiers, p. 1-28, 1999. https://doi.org/10.7551/mitpress/1113.001.0001

SMOLA, A. J.; SCHOLKOPF, B. Learning with Kernels. Massachusetts: The MIT Press, 2002. 625 p.

VALDERRAMA, P. CALIBRAÇÃO MULTIVARIADA DE PRIMEIRA E SEGUNDA ORDEM E FIGURAS DE MÉRITO NA QUANTIFICAÇÃO DE ENANTIÔMEROS POR ESPECTROSCOPIA. Tese apresentada ao Programa de Pós-Graduação em Química da Universidade Estadual de Campinas, Instituto de Química, para obtenção do grau de Doutorado em Química. Campinas, SP, 2009.

VAPNIK, V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995. https://doi.org/10.1007/978-1-4757-2440-0

XIAOBO, Z.; JIEWEN Z.; POVEY M. J.; HOLMES M.; HANPIN M. Variables selection methods in near-infrared spectroscopy. Analytica Chimica Acta, v. 667, n. 1-2, p. 14-32, 2010. https://doi.org/10.1016/j.aca.2010.03.048