Universidade de Brasília
Departamento de Estatística

Aplicações em Quimiometria

Alex Rodrigues do Nascimento

Projeto apresentada para obtenção do título de Bacharel em Estatística.

Brasília
2016

Alex Rodrigues do Nascimento

Aplicações em Quimiometria

Orientador:
Prof. **Bernardo Borba de Andrade**

Projeto apresentada para obtenção do título de Bacharel em Estatística.

Brasília
2016

DEDICATÓRIA

À minha Família

AGRADECIMENTOS

Agradeço à minha Família por todo o Apoio dado durante toda minha graduação.

À todos que me incentivaram e estiveram ao meu lado em todos os bons e difíceis momentos.

Ao meu orientador pelo tempo e paciência despendidas.

SUMÁRIO

RESUMO

Aplicações em Quimiometria

A quimiometria é definida como a aplicação de técnicas matemáticas e estatísticas à análise de dados relativos aos processos químicos tanto na área espectroscópica como cromatográfica. Essa área, usualmente, faz utilização de técnicas específicas para problemas de regressão e classificação.

Tendo em vista todo o aparato de técnicas da recente área de Aprendizagem Estatística (do inglês *Statistical Learning* - SL) que se encaixam aos problemas químicos, se tem como objetivo do presente trabalho: aplicar técnicas de SL em dados químicos, algumas já utilizadas na esfera da quimiometria e outras não, ampliando o aparato tecnológico da área e tentando gerar modelos com igual ou melhor previsibilidade.

Em uma primeira etapa do trabalho, realizou-se a análise, sob diferentes perspectivas, de dados amplamente utilizados na indústria farmacêutica para prever variáveis complexas, utilizando técnicas de Regressão com penalização, Mínimos Quadrados Parciais, Regressão com Componentes Principais, entre outras.

Em uma segunda, realizou-se um estudo de dados fornecidos pelo Grupo de Automação, Quimiometria e Química Ambiental (AQQUA) do Instituto de Química da UnB. A problemática dessa etapa é de muita importância para áreas de fiscalização ambiental, mais especificamente o controle e exploração da espécie de madeira Mogno. Essa etapa trata de um problema de classificação e utilizou-se técnicas de *Support Vector Machine*, Árvore de Decisão, Regressão com Penalização, entre outras.

Palavras-chave: Aprendizagem Estatística, Quimiometria, Mogno, Regressão, SVM, PLS, PCR, Regressão Penalização

ABSTRACT

In this study I have implemented several techniques from Statistical Learning to two problems in the field of Chemometrics.

The first problem was an exercise in predictive modeling and dimension reduction where I have performed the analysis of publicly available data used for Quantitative Structure-Activity Relationship (QSAR) modeling. the data consists of 228 chemical predictors used to model solubility. The techniques compared were Partial Least Squares, Principal Components Regression, Ridge Regression and LASSO.

In the second problem I analyzed data provided by the Automation, Chemometrics and Environmental Chemistry Group (AQQUA) from UnB with statistical techniques not commonly used in the field of Chemometrics. The dataset consists of near infrared spectroscopy measurements used to classify four types of wood with special interest in one type, Mogno. I compared the method typically used for this kind of chemical classification problem, Partial Least Squares, with Support Vector Machine, Decision Trees, Bagging, Ridge Regression and LASSO.

Keywords: Statistical Learning, Chemometrics, Mahogany, regression, SVM, PLS, PCR, Penalty Regression

1 INTRODUÇÃO

A Aprendizagem Estatística (do inglês *Statistical Learning - SL*) se trata de um conjunto de ferramentas de modelagem e entendimento de complexas bases de dados. Esse recente campo de estudos se mistura com ciência da computação, mais especificamente, Aprendizado de Máquinas (do inglês *Machine Learning* - ML), para gerar modelos com alto grau de acurácia para previsões.[2]

A SL é bastante utilizada no contexto de *Big Data*, base de dados com grandes dimensões e muita complexidade, advinda da continua evolução tecnológica com facilidade de obtenção e armazenamento de dados. As aplicações de técnicas de SL não se restringem a nenhuma área do conhecimento, sendo utilizada em finanças, biologia, saúde, química, marketing, entre outras.

A aprendizagem estatística trouxe novos e importantes conceitos para a área, como: a troca entre interpretabilidade e poder de previsão. Algumas técnicas possuem, como resultados, modelos de difícil compreensão da relação entre as variáveis explicativas e respostas, mas um alto poder de previsão.

A SL divide as técnicas em supervisionados, quando para cada observação de um conjuntos de variáveis preditoras está associado uma variável resposta, ou não supervisionado, em que não existe a distinção de variável resposta e preditora e se tem a intenção de estudar a relação entre um grupo de variáveis e observações. A terminologia supervisionada é utilizada por causa da presença de uma variável resultado que pode orientar o processo de aprendizagem do modelo, especificando a qualidade das previsão e, no caso não supervisionado, não existe essa possibilidade. Podemos citar como exemplos para estas técnicas clássicas a regressão linear e logística, além de técnicas mais modernas como *support vetor machines*, modelos generalizados aditivos e *boosting*. Para técnicas não supervisionadas podemos citar análise de cluster e componentes principais.

É interessante entender a forma como os métodos da área são avaliados, diminuindo a relevância de propriedades e conceitos da inferência clássica e incorporando medidas que avaliam a capacidade de previsão.

Como citado, as técnicas de SL são aplicáveis na química, que possui uma área específica para construção de modelos, a quimiometria é definida como a aplicação de técnicas matemáticas e estatísticas à análise de dados relativos aos processos químicos tanto na área espectroscópica como cromatográfica.[11]

Existem diversos trabalhos publicados da aplicação de técnicas estatísticas multivariadas em dados químicos, como apresentado em Pastore et al, 2011, em que se utilizou análise de mínimos quadrados parciais para discriminar o verdadeiro mogno de espécies com características físicas parecidas; Tsuchikawa et al. (2003), a distância de Mahalanobis para discriminar nove espécies de madeira; Schwanninger et al. (2004b) análise de Cluster para diferenciar 11 espécies de madeiras europeias.

Nesse contexto nasce o objetivo do presente trabalho, aplicar técnicas de SL em dados químicos, com variáveis físico-químicas e outro proveniente da realização de espectroscopia de infravermelho em alguns materiais, algumas já utilizadas na esfera da quimiometria e outras não, ampliando o aparato tecnológico da área e tentando gerar modelos com igual ou melhor previsibilidade. O presente trabalho faz utilizações de técnicas de Aprendizagem de Máquinas à análise de dados relativos aos processos químicos.

Na primeira etapa deste trabalho, realizou-se a análise, sob diferentes perspectivas, de dados amplamente utilizados na indústria farmacêutica para prever variáveis complexas como lipossolubilidade (melhor penetração através das membranas), lipofilia (tendência a acumular em tecidos adiposos), condicionamento da atividade antimicrobiana, etc. No presente trabalho a variável a ser estimada é lipossubilidade.

Em uma segunda etapa, realizou-se um estudo de dados fornecidos pelo Grupo de Automação, Quimiometria e Química Ambiental (AQQUA) do Instituto de Química da UnB. A problemática dessa etapa é de muita importância para áreas de fiscalização ambiental, mais especificamente o controle e exploração da madeira Mogno (*SwieteniamacrophyllaKing*).

2 MÉTODOS EM APRENDIZAGEM ESTATÍSTICA

Um conceito utilizado na área de *Statistical Learning* é o conjunto de treino e o conjunto de teste. Com o objetivo de avaliar a capacidade de previsão do modelo, divide-se o conjunto de dados em duas partes denominadas treino e teste, a primeira será responsável pelo treinamento do modelo e a outra pela avaliação. Por exemplo, se o conjunto de dados possuir $n = 100$ observações divide-se em um conjunto com $n_1 = 80$ e $n_2 = 20$, será aplicado, por exemplo, MQO nas n_1 observações e utilizará os coeficientes estimados para prever as respostas das n_2 observações e, assim, avaliar a capacidade de previsão do modelo.

2.1 Validação Cruzada

Validação Cruzada é um método para selecionar o melhor modelo dentre um determinado conjunto de modelos. Modelos não se refere apenas a diferentes métodos mas também o mesmo método com diferentes parâmetros, pois diversos desses a serem descritos possuem parâmetros que influenciam de maneira significativa os resultados e validação cruzada é uma forma de estima-los.

O método $K - fold$ implica em dividir aleatoriamente o conjunto de treino em k grupos de aproximadamente igual tamanho, na primeira iteração se retira um dos k grupos, chamado de conjunto de validação, e ajusta-se o modelo com os $k - 1$ grupos restantes. O erro médio de previsão (EMP_1) é calculado para o grupo que foi deixado de fora. Repete-se o processo k vezes e em cada iteração um grupo de observações é tratado como conjunto de validação. Ao final se tem k $EMP'S$ e o a estimativa $K - fold$ é dada pela média desses. Dessa forma, é possível selecionar o modelo/parâmetro baseado na estimativa $K - fold$.[2]
É interessante salientar sobre da escolha de k. Na literatura, é bastante comum se ver $k = 5$ ou $k = 10$. Todos os valores são possíveis, até mesmo 1 sendo denominado no inglês como *Leave one Out*, mas vale ressaltar que existe uma troca de viés e variância nesse caso, quanto maior o k maior o viés mas menor a variância das estimativas, e, quanto menor, a relação se inverte.

2.2 *Support Vector Machine*

O método de *Support Vector Machine* (SVM) é uma técnica desenvolvida pela comunidade de ciência da computação na década de $1990's$ e vem ganhando cada vez

mais espaço no campo científico devido as suas boas performances. A intuição do método é sistematizado a partir da técnica de máxima margem e, consequentemente, de classificação linear.
O procedimento de classificação linear constrói um limite de decisão linear que, de maneira clara, separa o conjunto de dados em diferentes classes.

> Segundo "Gareth James"em um espaço $p - dimensional$ um hiperplano afim é um sub-espaço linear de dimensão $(p-1)$ podendo ou não passar pela origem. Em $p = 2$ é uma linha e em $p = 3$ um plano.

Pode-se pensar que um hiperplano divide o espaço p-dimensional em duas metades, assim, é possível a construção de um classificador intuitivo: a classificação da observação de teste depende apenas de qual lado do hiperplano ela está localizada e, para essa identificação, utiliza-se conceitos de distancia de ponto ao plano e regra do cosseno para o produto escalar. Vale ressaltar que tal técnica é aplicável apenas para dados linearmente separáveis. Assim, se faz necessário a determinação de qual lado do hiperplano a observação está:

$$\operatorname{sgn}(\cos\theta) = \operatorname{sgn}\left(\langle \mathbf{x}, \mathbf{v}_h \rangle\right), \theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$$

No qual V_h é o vetor normal ao hiperplano, θ o ângulo formado entre o vetor normal e a observação e x o ponto a ser classificado.

Assim se define a regra do classificador linear $sgn(\langle \mathbf{x}, v_h \rangle)$ dado que existe uma hiperplano separador, ou seja, se o sinal da equação é positivo a classe é 1 se for negativo a classe é -1.

Existem uma infinidade de hiperplanos que dividem as observações em duas classes, geralmente se pode vizualizar deslocando-o um pouco para cima, para baixo ou mesmo o rotacionando, sem se chocar com nenhuma observação. Do questionamento de qual

escolher se tem a idealização de máxima margem, no qual se escolhe o hiperplano que possui a maior distancia entre as observações de treino. Sendo mais explicito, é possível calcular a distancia de cada observação ao hiperplano separador e a menor distancia entre todas é chamado de margem.

O hiperplano de máxima margem é o hiperplano separador que possui a maior margem, o qual ficará mais distante possível das observações de cada lado. O classificador irá determinar a classe de acordo qual lado a observação de teste ficou do hiperplano de máxima margem.

Ao se definir o classificador de máxima margem, nasce outro conceito, o de vetores suporte que são as observações que definem o hiperplano separador. Como para construção do hiperplano de máxima margem utiliza-se apenas as observações que apresentaram a menor distancia ao mesmo, as restantes não vão afetar a sua definição, logo, o hiperplano separador será diretamente depende dos vetores suportes que são exatamente essas observações que apresentaram as menores distancias.
Baseado em um conjunto de treinamento $x_i, \ldots, x_n \in \mathbb{R}^p$ associado a classes $y_1, \ldots, y_n \in \{-1, 1\}$, A construção do classificador de máxima margem se resume ao problema de otimização:

$$\begin{aligned} & \text{Max M} \\ s.e \sum_{j=1}^{p} \beta_j^2 &= 1 \\ y_i(\beta_0 + \beta_1 x_i 1 + \beta_2 x_i 2 + \ldots + \beta_p x_{ip}) &\geq M, \forall i = 1, \ldots, n \end{aligned}$$

No qual M é a margem a ser maximizada, $(x_i^{\mathrm{T}}\beta + \beta_0)$ define o hiperplano e $y_i, \in (-1, 1)$. As restrições presente no problema descrito garantem que cada observação fique do lado correto do hiperplano e pelo menos um M de distância do mesmo[2]. O problema pode ser reformulado para que tenhamos um problema de otimização convexa[1].

Toda a sistematização apresentada faz duas suposições: que os dados sejam perfeitamente separáveis e que seja de forma linear. O método de *Support Vector Machine*

flexibiliza essas suposições e proporciona aplicabilidade ao método.

A técnica explicada até o momento não possui solução quando os dados não são perfeitamente separáveis e, mesmo que exista, o método é extremamente sensível a poucos observações pois, ao adicionar um ponto que se posicione do lado contrário do hiperplano separador, muda completamente a construção do mesmo e essa dependência forte de uma única observação sugere sobre-ajuste. Por essa perspectiva é válido permitir que algumas observações não sejam responsáveis pela definição do hiperplano ou mesmo não se classifiquem de maneira correta para que, assim, se tenha um estimador mais robusto a observações individuais e se construa um melhor classificador para a maior parte do conjunto de treino.

Como solução se faz uma pequena modificação nas restrições do problema de otimização e se tem os seguintes resultado:

$$\begin{aligned} & \max M \\ s.e & \sum_{j=1}^{p} \beta_j^2 = 1 \\ y_i \left(\beta_0 + \beta_1 x_i 1 + \beta_2 x_i 2 + \ldots + \beta_p x_{ip}\right) & \geq M\left(1 - \epsilon_i\right) \\ \epsilon_i \geq 0, \sum_{j=1}^{p} \epsilon_i & \leq \mathrm{C} \end{aligned}$$

Onde ϵ_i é definida como variável de folga, a qual é a quantidade proporcional que a predição está do lado errado da margem. Sendo mais explicito, a variável de folga ϵ_i diz onde a observação ith está localizada em relação ao hiperplano e a margem. Se $\epsilon_i = 0$ então a observação esta do lado correto da margem, Se $\epsilon_i > 0$ a observação está do lado errado da margem e se $\epsilon_i > 1$ a observação está do lado errado do hiperplano. Como $\sum_{j=1}^{p} \epsilon_i \leq C$, C determina o número e a gravidade das as violações a margem que vamos tolerar, esse é determinado utilizando validação cruzada[1,2]. Dessa forma, é possível construir um classificador robusto mesmo para observações que não são perfeitamente separáveis.

Em alguns conjuntos de dados, a fronteiras de classificação não é necessariamente linear, logo, existe uma necessidade de flexibilizar essa restrição. Utiliza-se o mesmo princípio de regressão linear, quando os dados não se relacionam de maneira linear, existe a possibilidade de inclusão de termos quadráticos ou cúbicos.

O método de *Support Vector Machine* é um extensão dessa ideia. Como apresentado, o problema de classificação pode ser representado em termos de produtos internos, mais detalhes em [3]. Pode-se aplicar uma transformação nesse produto interno com a intenção de flexibilizar a restrição linear, utilizando Kernel's.

2.3 Árvore de Decisão

O método de Árvore de decisão (AD) consiste em particionar, segmentar o espaço preditor em números de regiões simples e a regra de decisão é dada pela maior frequência da classe que a região a qual as observações de treinamento pertencem. O espaço é particionado de acordo com os possíveis valores de $X_1, X_2, \ldots, X_P$ dentro de J distintos e não sobrepostas regiões $R_1, R_2, \ldots, R_p$. Para toda observação que pertencer a região R_j se faz a mesma predição da classe.

Para construção das regiões $R_1, R_2, \ldots, R_p$, se particiona o espaço em todos os valores presentes de $X_1, X_2, \ldots, X_P$ e é selecionado a região que gere o menor erro de previsão. Sendo mais específico, no primeiro passo seleciona-se o preditor X_J e o ponto de corte s tais que a divisão do espaço nas regiões $X|Xj < s$ e $X|Xj \geq s$ leve a maior redução possível do erro médio de previsão.Em seguida, repete-se o processo, procurando o melhor preditor e melhor ponto de corte , a fim de dividir os dados de modo a minimizar o erro médio de previsão dentro cada uma das regiões já criadas no primeiro passo.[2]

O método de árvore de decisão é simples e possui alto poder de interpretação, podendo identificar as variáveis preditoras mais importantes para classificação mas costuma gerar previsões menos precisas que SVM e regressão com penalização.

2.4 *Bagging*

A técnica de *Bagging* pode ser utilizada em conjunto com diversos métodos aprendizagem estatística e, no presente trabalho, foi utilizado no contexto de árvore de decisão.

O método é um procedimento geral que tem como objetivo a redução da variância das estimativas produzidas pelos métodos de aprendizagem. A técnica de Árvore de Decisão produz estimativas com baixo viés mas grande variância, ou seja, se dividirmos de forma aleatória o conjunto de treinamento e aplicarmos a técnica em cada grupo poderíamos ter estimativas muito diferentes.

Bagging utiliza conceitos de *Bootstrap*. O processo se resume a retirar B amostras *Bootstrap* do conjunto de treinamento, aplicar Árvore de Decisão em cada **bth** amostra e gerar a estimativa $\widehat{f}^{*b}(x)$. No final do processo se tem B estimativas e a final é a média das $B's$ calculadas[2]. Assim, temos a estimativa de *Bagging*:

$$\widehat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{*b}(x)$$

Para o contexto de classificação, é computada a estimativa para cada amostra B e a previsão geral é dada pela classe mais frequente presente nas B amostras.

2.5 Florestas Aleatórias

Florestas Aleatórias é um aperfeiçoamento da técnica de *Bagging*, também utiliza o conceito de retirar amostras com reposição do conjunto de treinamento e aplicar o método em cada uma das B amostras, mas o que aquela se diferencia dessa é o fato de reduzir o espaço de preditores realizando uma amostra aleatória das variáveis.

Sendo mais específico, o processo implica em retirar B amostras com reposição do conjunto de teste e aplicar o método em cada B amostra, mas em cada aplicação não será considerado o conjunto p de preditores e sim um conjunto m , com $m \leq p$ que é determinado aleatoriamente para cada uma das B amostras.

O objetivo do método é diminuir a correlação entre as B estimativas. No método de *Bagging* as variáveis que tem maior poder de explicação irão dominar as B

estimativas e, assim, implicar em uma alta correlação entre elas. A média dessas estimativas correlacionadas não irá gerar o efeito desejável de redução da variância das estimativas, mas esse aperfeiçoamento gerado por Florestas Aleatórias irá permitir a influencia de outros preditores e gerar estimativas menos correlacionadas.

2.6 Regressão *Ridge*

Um dos pressupostos para se ter as propriedades de Mínimos quadrados ordinários (MQO) é a ausência de multicolinearidade, ou seja, é necessário que a matriz $\mathbf{X}^T\mathbf{X}$ seja não singular. Para problemas atuais onde se tem grandes dimensões com alguns casos que $p \geq n$, tal pressuposto é dificilmente atendido e, se tal fato ocorrer, o problema se torna numericamente instável e se tem como consequência um aumento da variância das estimativas. Uma alternativa é se ajustar o modelo contendo os p preditores usando técnicas que restringem ou regularizam as estimativas dos coeficientes.

Regressão *Ridge* é uma modificação de mínimos quadrados que tem como objetivo gerar estimativas mais robustas quando $\mathbf{X}^T\mathbf{X}$ é quase singular. A solução desse método para um problema de regressão linear é dado por: [1]

$$\widehat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I})^{-1}\mathbf{X}^T y$$

Dado que λ é o parâmetro de ajuste e $\mathbb{I}$ matriz identidade.

A inclusão do termo λ gera uma perturbação na matriz, alterando o número de condição e assim solucionando o problema de não inversão. A estimativa dos coeficientes da regressão *Ridge* é dado pelo seguinte problema de otimização.

$$\widehat{\beta}^{ridge} = \arg\min_{\beta}\{\|y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\}$$

$\|\beta\|_2^2$ pode ser interpretado com termo de penalização, introduz ao problema um custo pela adição variáveis e contrai os coeficientes estimados. O parâmetro λ é

selecionado via validação cruzada.

2.7 Regressão LASSO

A regressão LASSO é uma outra alternativa para o caso onde se tem um problema com alta dimensão e variáveis repostas correlacionadas, se assemelha ao método *Ridge* alterando o termo de penalização. As estimativas para regressão LASSO é dado pelo seguinte problema de otimização:

$$\widehat{\beta}^{lasso} = \arg\min_{\beta}\{\|y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1\}$$

na qual,

$$|\beta|_1 = \sum_{j=1}^{P} |\beta_j|$$

O termo de penalização que em *Ridge* era dado por β_j^2, agora se torna $|\beta|$. Em termos práticos, a consequência dessa alteração é que os valores dos coeficientes em *Ridge* são diminuídos mas dificilmente serão iguais a zero e a penalização em LASSO força de maneira efetiva os coeficientes assumirem valor nulo. Dessa forma, LASSO além de ser um método de regularização é, também, um de seleção de variáveis, pois só permanecerá no modelo as variáveis que forem realmente relevantes para predição, LASSO gera o que se chama de resultado esparsos. As informações apresentadas são justificadas matematicamente mas fogem do escopo do trabalho, para um melhor detalhamento[1].

Um aspecto interessante ao se analisar os dois tipos de regressão citados acima é o *trade-off* entre viés e variância. A inclusão do termo λ introduz viés ao estimador mas reduz sua variância, logo há uma troca entre viés e variância que em alguns casos é interessante por aumentar o poder de previsão do modelo (diminuição do erro quadrático médio de previsão)[2].

2.8 Regresão com PCA

Como citado no caso de regressão com penalização (LASSO e *Ridge*), as bases de dados atualmente possuem grande dimensões e métodos tradicionais começam a perder propriedades desejáveis. Uma técnica clássica de redução de dimensão pode ser aplicado a problemas de modelagem preditiva, componentes principais.

O método basicamente aplica componente principais nas variáveis preditoras e utiliza-se para gerar os modelos os componentes que expliquem uma relevante parte da variância dessas. Assim, ao invés de utilizar um grande número de variáveis p, aplica-se mínimos quadrados em um número de componentes d com $d \leq p$.

A ideia-chave é que muitas vezes um pequeno número de componentes principais é suficiente para explicar a maior parte da variabilidade dos dados , bem como a relação com a resposta (não sendo garantida). O número de componentes geralmente é selecionado via validação cruzada e para previsão de novas observações é utilizado os *loadings* do conjunto de treinamento.

2.9 Mínimos Quadrados Parciais

Uma possível falha do método de regressão com componentes principais (PCR) é o fato dos componentes serem derivados apenas das variáveis respostas e, assim, abre a possibilidade dessas, que explicam a maior parte da variância das preditoras, não incorporarem a relação com a resposta. O método de Componentes Principais é não-supervisionado.

Como uma alternativa de um método supervisionado para PCR se tem Mínimos Quadrados Parciais (PLS) que, como PCR, é um método de redução de dimensão que identifica d componentes que são combinação linear das variáveis originais e, ao invés de utilizar X e Y originais no procedimento de MQO, utiliza-se os componentes gerados e esses por sua vez utilizam informações de Y.

Para *PLS* é comum chamar os componentes de variáveis latentes e essas, ao contrário de PCA, são obtidas de maneira iterativa pelo algoritmo que segue:

Se começa aplicando Decomposição de Valor Singular (SVD) na matrix de produto cruzado $S = X^T Y$, dessa forma se inclui as informações da variância de X e Y e sua correlação. Os primeiros vetores singulares da esquerda e da direita, w e q, são utilizados como peso para X e Y respectivamente e geram os scores t e u:

$$\begin{aligned} t &= Xw \\ u &= Yq \end{aligned}$$

Se utiliza a t normalizado. O próximo passo é encontra os loadings X e Y e para isso se utiliza o vetor t:

$$\begin{aligned} p &= X^T t \\ q &= Y^T t \end{aligned}$$

Finalmente retira-se a informação relacionada a variável latente, na forma do produto tp^T e tq^T subtraindo das matrizes originais X e Y :

$$\begin{aligned} X_{n+1} &= X_n - tp^T \\ Y_{n+1} &= Y_n - tq^T \end{aligned}$$

E assim se repete o processo iterativamente. os vetores w, t, p e q são armazenados nas matrizes W, T, P e Q, respectivamente. Utiliza-se a matriz de scores T para estimar os coeficientes e posteriormente converte-los de volta para a parametrização das variáveis originais pré-multiplicando com a matriz R (desde $T = XR$ e $R = W(P^T W)^{-1}$):

$$B = R(T^T T)^{-1} T^T Y = RT^T Y = RQ^T \tag{1}$$

O algoritmo apresentado é denominado $SIMPLS$.

3 PREDIÇÃO DE LIPOSSUBILIDADE

3.1 Introdução

Em todos os projetos de descoberta de fármacos, um entendimento profundo sobre as relações entre a estrutura e atividade (SAR) é essencial para o rápido desenvolvimento de candidatos a novos fármacos.

Nesta primeira etapa do trabalho foi realizado um estudo sob diferentes perspectivas de uma base de dados que traz 228 atributos (preditores) a serem utilizados na previsão da variável físico-química lipossubilidade. São 208 atributos binários (fingerprints FP1-208) indicando presença de uma particular subestrutura química; 16 atributos de contagem e 4 atributos contínuos.

Esse tipo de base de dados tem grande uso na indústria farmacêutica e é usado para a previsão de atividades biológicas ou outras propriedades observáveis a partir de estudos de relação quantitativa entre estrutura e atividade (QSAR): o objetivo final é prever variáveis complexas como lipossolubilidade (melhor penetração através das membranas), lipofilia (tendência a acumular em tecidos adiposos), condicionamento da atividade antimicrobiana, etc.

3.2 Metodologia

O banco de dados utilizado é advindo da literatura e está disponível.

Foi feito uso do métodos clássico de Regressão Linear.

Regressão *Ridge* e *LASSO*: Com o objetivo de restringir e regularizar as estimativas dos coeficientes de uma regressão linear comum, para cada uma das técnicas penaliza-se a estimativa pela quantidade de p preditores existentes, gerando estimativas subestimadas em relação a regressão linear. Tende a dar pesos maiores aos preditores que possuem maior relação com a variável resposta.

Aplicou-se a técnica de regressão de componentes principais, onde se aplica componente principais nas variáveis preditoras e utiliza-se para gerar o modelos os componentes que expliquem uma relevante parte da variância dessas e, ainda, mínimos quadrados parciais, com também o objetivo de reduzir a dimensão do problema, ajusta-se um modelo com M variáveis que são resultados da combinação de linear dos p preditores e que os coeficientes são

dados ajustando-se regressões lineares simples com a variável resposta.

Para avaliação da capacidade de previsão dos métodos será utilizada a medida de Erro Quadrático Médio (EQM):

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$$

No qual $\widehat{f}(x_i)$ é o resultado da predição da observação i utilizando determinado método.

Para avaliação dos modelos será calculado o erro quadrático médio do conjunto de treino(EQM_{treino}), teste(EQM_{teste}) e validação cruzada(EQM_{cv}). O conjunto de dados foi dividido em 25% para teste e 75% para treino.

As análises foram desenvolvidas no Software R e toda a abordagem teórica foi baseada nos livros An Introduction to Statistical Learning with Applications in R [2] e The Elements of Statistical Learning [1].

3.3 Resultados

Os resultados são referentes a análises do banco de dados da indústria farmacêutica, são 208 atributos binários indicando presença de uma particular subestrutura química e 20 atributos contínuos. Os atributos são descritores moleculares que podem ser de fácil interpretação e mensuração ou extremamente complexos (eg. peso molecular, número de ligações, de rotações, índices topológicos,índices de conectividade, índices de forma molecular, descritores químico-quânticos). O banco de dados possui 1267 observações

3.3.1 Análise Descritiva

Tendo em vista a dimensão do banco de dados, a análise descritiva fica restrita não sendo viável examinar as variáveis uma a uma, dessa forma, realizou-se uma resumida, com foco em aspectos mais relevantes para análises posteriores.

Depreende-se da tabela 1 que a distribuição da variável é levemente assimétrica a esquerda, com média de $-2,74$ e mediana $-2,49$, a diferença entre o intervalo interquartí-

Tabela 1 – Quadra resumo das estatísticas referente a variável Resposta do Modelo

Mínimo	Primeiro Quartil	Mediana	Média	Terceiro Quartil	Maximo
-11,62	-3,96	-2,49	-2,74	-1,36	1,58

lico e a amplitude total sugere, ainda, a presença de valores extremos. O histograma ilustra a leve assimetria e não sugere normalidade, tendo em vista o tamanho da amostra

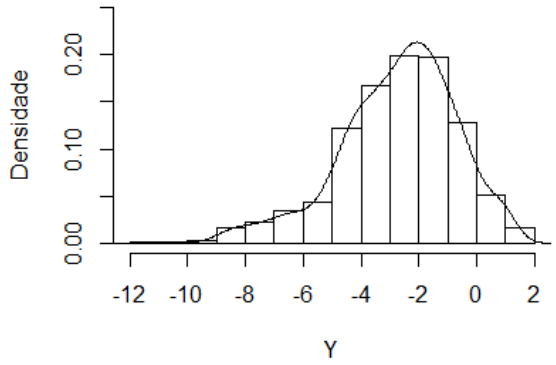


Figura 1 – Histograma da Variável Resposta (Y)

A existência de 208 variáveis binárias dificulta analisar a frequência de cada subestrutura química separadamente. Como alternativa se construiu um boxplot das frequência relativas de cada.

Verifica-se que a maior parte das subestruturas químicas estão presente em menos de 30% da amostra, com a existência de valores extremos. Algumas compostos estão presente em mais de 50% e um, especificamente, está presente em mais de 80% da amostra.

Para as variáveis quantitativas restringiu-se a uma análise de correlações.

Depreende-se da tabela 2 que existem variáveis altamente correlacionadas,com casos de coeficientes de correlação com valor de 0.99(par número de átomos e número de ligações), esse resultado talvez seja interessante e evidente na esfera da química mas, para a finalidade de modelagem com foco em previsão, não é desejável.

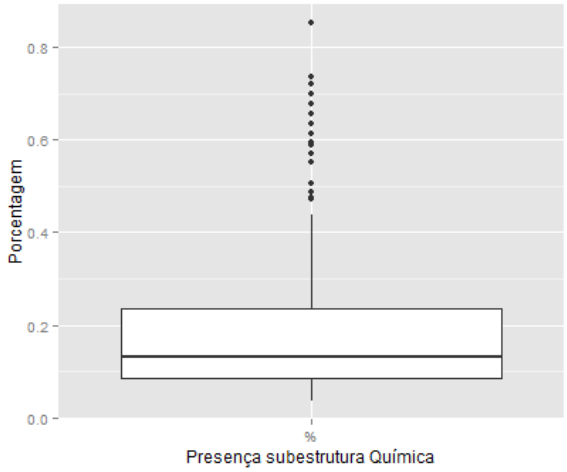


Figura 2 – Box Plot da frequência relativa da presença das 208 subestruturas químicas

Tabela 2 – Quadra resumo das dez maiores intensidades de correlações

	Variável 1	Variável 2	ρ
1	Número de Átomos	Número de Ligações	0,99
2	Número de Átomos de Hidrogênio	Numero de Ligações de Hidrogênio	0,99
3	Área da superfícia 1	Área da superfície 2	0,96
4	Número de Ligações Múltiplas	Número de ligações Aromaticas	0,94
5	Número de Ligações	Número de Carbonos	0,94
6	Numero de Ligações de Hidrogênio	Número de Carbonos	0,93
7	Número de Átomos de Hidrogênio	Número de Carbonos	0,92
8	Número de Átomos	Número de Carbonos	0,91
9	Peso Mol	Número de Átomos de Hidrogênio	0,90
10	Número de Átomos de Hidrogênio	Número de Ligações	0,89

3.3.2 Modelo Linear

Como ponto de partida foi construído um modelo de regressão linear múltipla, regredindo a variável resposta com todas as outras 228 variáveis disponíveis.

O resultado indicou 54 variáveis significativas, um R^2 de 93%, resíduos não-normais (teste de shapiro wilk com p-valor $< 0,05$) e heterocedasticidade (teste de Breusch Pagan com p-valor $< 0,05$), entretanto, a característica do modelo que mais influencia as estimativas encontradas é a multicolinearidade.

Como esperado pela análise de correlações anterior, o modelo indicou presença

de multicolinearidade evidenciado pelos altos valores de inflação da variância (VIF'S).

Tabela 3 – Quadra resumo dos dez maiores VIF's do Modelo Linear

Variaveis	VIFS
NumNonHBonds	7637,61
NumAtoms	7382,23
NumNonHAtoms	6307,98
NumBonds	4984,89
NumNitrogen	1954,53
NumAromaticBonds	1384,57
FP063	1142,31
FP005	657,62
NumSulfer	612,89
NumHalogen	474,09

É aconselhável investigar valores de VIF's maiores que 20, no determinado modelo atingiu-se valores na casa dos milhares com números maiores que 7000, evidenciando de maneira forte a presença de multicolinearidade. Dessa forma, sabe-se que a propriedade de eficiência do estimador não é atendida,o que justifica a utilização de métodos para reduzir a dimensão do problema e , assim, construir modelos com melhor capacidade de previsão.

Para critérios de comparação, calculou-se, sem nenhuma alteração do modelo, o erro quadrático médio de previsão no conjunto de dados teste , tendo como resultado o valor de $0,56$.

Uma proposta simplória utilizada foi retirar do modelo as variáveis que não tiveram coeficientes significantes, restando apenas 54 variáveis. Os Vif's foram reduzidos mas não de maneira suficiente.

Tabela 4 – Resultados Para Modelos Lineares

Modelo Linear	EQM_{teste}	EQM_{treino}
Completo	0,56	0,23
Restrito	0,68	0,43

Verifica-se que a retirada de variáveis do problema não diminuiu o erro quadrático médio, ao contrário, aumentou. O erro de treino para o modelo completo foi de $0,23$ e para o restrito $0,43$, para o conjunto de teste os resultado foram de $0,56$ e $0,68$ respectivamente. A análise mostra o sobre-ajuste em utilizar o modelo linear completo, dado que o

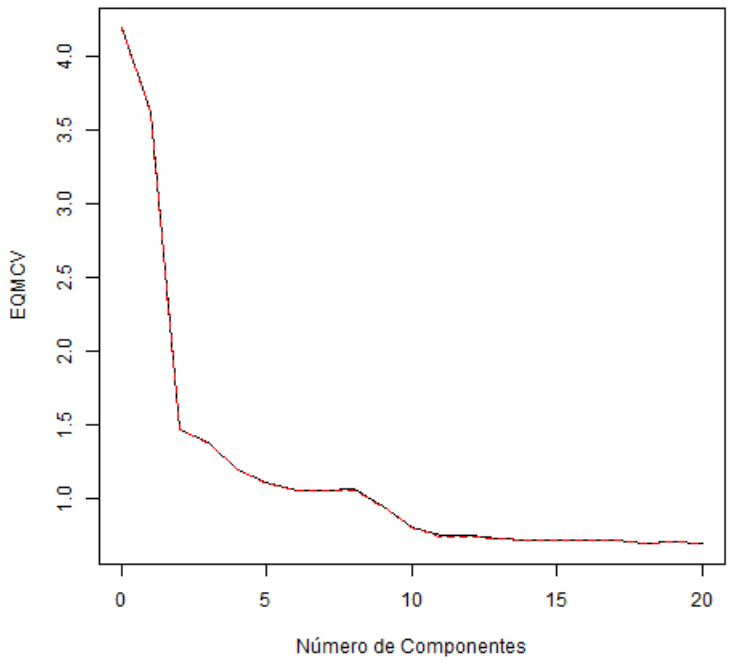
EQM_{teste} foi o dobro que o de EQM_{treino} para esse modelo.

Uma alternativa para uma melhor seleção de variáveis é a utilização de métodos mais robustos e refinados, como stepwise ou *best subsets*, mas, devido à grande dimensão do problema, se teve a dificuldade de custo computacional para realiza-los. Uma solução para tal é a utilização de processamento em paralelo mas esse foge do escopo do trabalho sendo deixado para projetos posteriores.

3.3.3 Regressão Com Componentes Principais

O aplicação de componentes principais é possível apenas em variáveis quantitativas, dessa forma, se aplicou o método de PCR utilizando como variáveis respostas os componentes gerados apenas das variáveis contínuas presentes no problema. Para seleção do número de componentes se utilizou-se Validação cruzada.

Figura 3 – Erro quadrático Médio de Validação Cruzada segundo Número de Componentes - PCR

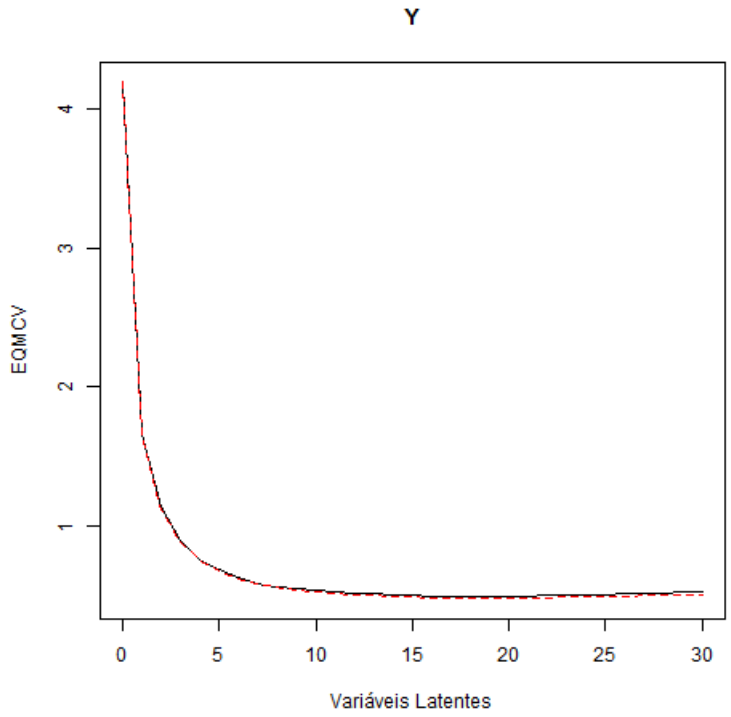


Analisando a figura verifica-se que o EQM_{cv} não diminui de maneira considerável após o 14º componente, dessa forma, se escolheu esse número de componentes para gerar as estimativas.

3.3.4 Mínimos Quadrados Parciais

O primeiro passo para utilização da técnica de PLS é a determinação de quantas Variáveis Latentes serão utilizadas no problema, essa quantidade foi determinada por validação cruzada.

Figura 4 – Erro quadrático Médio de Validação Cruzada segundo Número de Componentes - PLS



Pela figura depreende-se que após o 20º componente o EQM_{cv} se estabiliza, dessa forma, se escolheu essa quantidade para gerar as estimativas.

Tabela 5 – Resultados para Técnicas de Redução de Dimensão

Método de Redução	EQM_{teste}	EQM_{treino}	EQM_{cv}
PCR	0,81	0,69	0,69
PLS	0,54	0,28	0,30

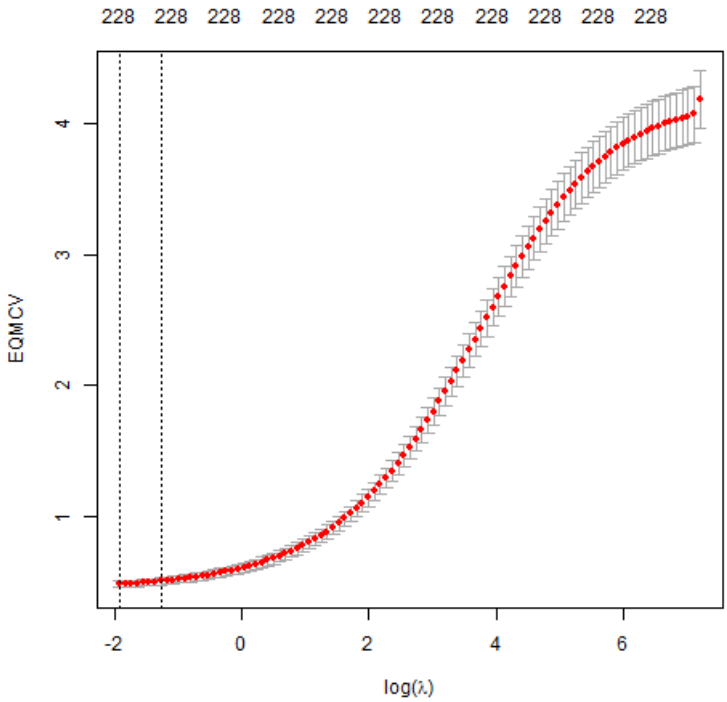
Depreende-se da tabela que a técnica de PCR não diminuiu o erro quadrático médio em relação ao modelo linear completo e os resultados foram parecidos para o modelo linear restrito, essa solução pode ter sido influenciada pela restrição em utilizar as variáveis respostas binárias.

O método de PLS obteve melhores resultados que o modelo linear, com erro quadrático médio para o conjunto de teste de $0,54$.

3.3.5 Regressão Ridge

Para construção do modelo de regressão Ridge é necessário a estimação do parâmetro de penalização λ, o mesmo foi estimado via validação cruzada de 10 grupos.

Figura 5 – Curva do erro de validação cruzada para evolução de $\log(\lambda)$, e as curvas de desvio padrão superiores e inferiores - Regressão Ridge



O λ que gerou o menor valor de EQM_{cv} foi $0,14$.

3.3.6 Regressão LASSO

Assim como em Ridge, para estimação do parâmetro de penalização na regressão LASSO se utilizou validação cruzada de 10 grupos

O lambda que gerou o menor valor de EQM_{cv} para LASSO foi $0,004$. É valido ressaltar a propriedade esparsa de LASSO, que das 228 variáveis do problema 88 tiveram seus coeficientes zerados, evidenciando sua capacidade de seleção.

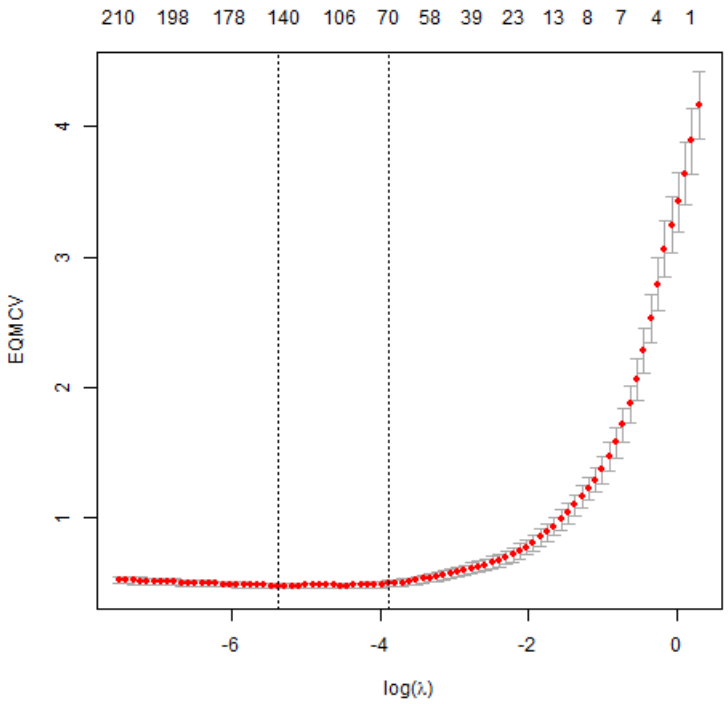


Figura 6 – Curva do erro de validação cruzada para evolução de $\log(\lambda)$, e as curvas de desvio padrão superiores e inferiores - Regressão LASSO

Tabela 6 – Resultados para Regressão com Penalização

Tipo Penalização	EQM_{teste}	EQM_{treino}	EQM_{cv}
Ridge	0,51	0,32	0,49
Lasso	0,49	0,30	0,45

A tabela mostra que as regressões com penalização obtiveram os melhores resultados, levando em consideração o EQM_{teste}. Lasso foi o método que melhor se encaixou ao problema e diminuiu o erro quadrático médio de teste em mais de 14% em relação ao modelo linear completo.

3.4 Conclusão

A primeira etapa do trabalho mostrou a importância da inclusão de métodos de Aprendizagem Estatística para o desenvolvimento de novos fármacos, mais especificamente na construção de modelos (QSAR).Fixando como parâmetro de comparação a técnica clássica de Regressão Linear, os métodos de aprendizagem estatística aplicados diminuíram, em diferentes proporções, o erro quadrático médio.

O problema tratado é considerado de alta de dimensão (mais de 200 variáveis), e, usualmente, o método de regressão linear não conserva suas propriedades nesses casos, sendo necessário a aplicação de outras técnicas em conjunto como, por exemplo, seleção de modelos que possui alto custo computacional. Se fez a tentativa de usar tais técnicas no projeto sendo inviabilizado pelo tempo de processamento, o que mostra a importância das técnicas de SL como possíveis soluções.

O método que mostrou melhor desempenho foi a regressão $LASSO$, diminuindo em mais de 14% o erro quadrático médio. A técnica é extremamente prática, pois além de gerar boas estimativas tem a propriedade intrínseca de esparsidade, ou seja, de seleção de modelos.

4 DISCRIMINAÇÃO DO MOGNO

4.1 Introdução

Nesta segunda etapa analizou-se dados oriundos da aplicação NIRS a espécies de madeira. Por suas excelentes propriedades como: bela cor rosada para avermelhada, densidade específica e boa resistência a fungos e insetos [6]; o mogno foi amplamente explorada para comercialização e, hoje, é considerada uma espécie ameaçada de extinção.

Em situações de fiscalização ou apreensão para controle da exploração e comércio, geralmente dispõe-se apenas da madeira em toras ou tábuas e o método geralmente utilizado para discrimina-la de outras madeiras é o anatômico visual, e, como existem diversas espécies semelhantes ao mogno, o desenvolvimento de métodos instrumentais para a identificação da espécie é uma alternativa que pode tornar a sua fiscalização mais eficiente.[3]

Na Química existe um procedimento clássico para determinação da fórmula molecular de uma substância que possui três passos: análise elementar qualitativa, descobrir quais átomos estão presentes; análise elementar quantitativa, percentagem de de cada átomo na molécula; e determinação do peso molecular. Os procedimentos tradicionais costumam despender tempo e recursos,como por exemplo, a provável necessidade de laboratórios que realizem análises elementares *in loco* [8]. Nesse cenário se desenvolveu estudos na área da espectroscopia.

A técnica de espectroscopia no infravermelho próximo oferece vantagens para o estudo da madeira, uma vez que possibilita análises rápidas e não destrutivas, com simples preparo de amostra e pode ser aplicada a amostras sólidas.[3]

Na utilização da técnica de NIRS é necessário construir modelos de calibração que possam ser posteriormente aplicados a amostras reais. Uma técnica amplamente utilizada na construção de modelos para discriminação de madeira é a regressão por mínimos quadrados parciais para análise discriminante (do inglês Partial Least Squares Discriminant Analysis - PLS-DA).

A técnica de PLS-DA, em determinadas circunstâncias, não se sobressai de maneira efetiva à métodos clássicos como Análise de Discriminante Linear (LDA) ou Distância Euclidiana [7], dessa forma se justifica a sondagem de outras técnicas, tendo em vista a popularidade de PLS-DA no ramo da quimiometria.

4.2 Metodologia

As amostras de andiroba, cedro e curupixá foram obtidas de discos localizados às bases dos troncos das árvores. Cada amostra corresponde a uma árvore (indivíduo) diferente. Essas espécies foram coletadas em áreas de exploração autorizadas no estado do Pará. Os espécimes foram doados pelas empresas Serraria Marajoara Ind. Com. e Exp., Selectas S.A. Ind. e Com.de Madeira Ltda, Madeireira Caingá Ltda e Empresa Juruá Florestal. As amostras de mogno foram obtidas de tábuas apreendidas oriundas do Mato Grosso do Sul que seriam exportadas. Como forma de garantir a identificação das amostras por espécie, essas foram identificadas pela anatomista botânica Dra. Vera Teresinha Rauber Coradin do LPF, SFB.Todas a informações acima supracitadas foram retiradas de [3].

As amostras foram cedidas pelo Laboratório de Produtos Florestais (LPF) do SFB, já previamente identificadas por espécie, secas ao ar livre, cortadas e com as faces orientadas (de acordo com as sessões transversal, radial e tangencial, conforme a Figura 2). As superfícies de todas as amostras foram lixadas com lixa número 80, mantendo-se a uniformidade granulométrica, que é um fator importante para a distribuição da radiação refletida. As amostras de andiroba, cedro e curupixá foram obtidas de discos localizados às bases dos troncos das árvores. Cada amostra corresponde a uma árvore (indivíduo) diferente. Essas espécies foram coletadas em áreas de exploração autorizadas no estado do Pará. Os espécimes foram doados pelas empresas Serraria Marajoara Ind. Com. e Exp., Selectas S.A. Ind. e Com.de Madeira Ltda, Madeireira Caingá Ltda e Empresa Juruá Florestal.

O equipamento de NIRS utilizado foi o PHAZIR que, por se tratar de um equipamento portátil, é uma abordagem inovadora na identificação de mogno. Desenvolvidos para uso em campo que possibilita uma supervisão da exploração mais prática.[3]

Para cada amostra foram obtidos quatro espectros, dois em cada face radial, medidos em pontos distintos da face.Foi calculada a média dos quatro espectros de cada amostra resultando em 111 espectros.[3]

4.2.1 Métodos de SL

Foi feito uso do métodos clássico de Regressão Logística, em conjunto com a técnica de seleção de um subconjuntos de preditores, com a intenção de reduzir a dimensão

do problema,*Stepwise*:começa com um modelo que não contém preditores, e, em seguida, adiciona preditores que dê o melhor ajuste, um-a-um, até que a adição não gere mais ganhos significantes no sentido de diminuir o Erro Médio de Previsão. (*FORWARD*);

Regressão *Ridge* e *LASSO*: Com o objetivo de restringir e regularizar as estimativas dos coeficientes de uma regressão logística comum, para cada uma das técnicas penaliza-se a estimativa pela quantidade de p preditores existentes, gerando estimativas subestimadas em relação a regressão linear, tende a dar pesos maiores aos preditores que possuem maior relação com a variável resposta.

Aplicou-se a técnica de regressão de componentes principais, onde se aplica componente principais nas variáveis preditoras e utiliza-se para gerar o modelos os componentes que expliquem uma relevante parte da variância dessas e, ainda, mínimos quadrados parciais, com também o objetivo de reduzir a dimensão do problema, ajusta-se um modelo com M variáveis que são resultados da combinação de linear dos p preditores e que os coeficientes são dados ajustando-se regressões lineares simples com a variável resposta.

Utilizou-se também SVM, sistematizado a partir da técnica de máxima margem e, consequentemente, de classificação linear. Árvores de Decisão, segmentação do espaço preditor em números de regiões simples e , ainda, os métodos de reamostragem *Bagging* e Florestas Aleatórias aplicados à AD.

4.2.2 Pré-processamento

Os pré-processamentos minimizam as variações nos dados que não estão relacionadas com a propriedade de interesse.[3] Existem alguns pré-processamentos usualmente utilizados na quimiometria e para o presente trabalho foram utilizados os seguintes:

Centrar na média: Calcula-se a média de cada variável $\mathbf{X}$ (comprimentos de onda, por exemplo) e a subtrai de cada observação em cada variável correspondente. Essa operação pode tornar o modelo mais robusto.

Subtração de linha reta (em inglês, straight line subtraction): Ajusta uma linha reta ao espectro e a subtrai deste, corrigindo inclinações e desvios de linha de base. O processo consiste em ajusta uma regressão simples entre a absorbância($\mathbf{Y}$) e o número de onda.

A amostra foi dividida aleatoriamente em um conjunto de teste e outro de treino, 25% e 75% respectivamente. A variável Y foi criada recebendo valor 1 para as observações pertencentes a classe e 0 para as demais, no banco de dados se tem a 28 observações da espécie mogno e 83 das demais

Os modelos serão avaliados em relação ao Erro Médio de Previsão (ou taxa de erro):

$$EMP = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(y_i \neq \widehat{y}_i)$$

Na qual $\widehat{y}_i$ é previsão de classe da observação i para determinada técnica e $\mathbf{I}(y_i \neq \widehat{y}_i)$ é uma função indicadora que recebe valor 1 se $y_i \neq \widehat{y}_i$ e 0 se $y_i = \widehat{y}_i$, ou seja, se a função indicadora resultar em 0 o método classificou corretamente a observação.

Calculou-se o Erro Médio de Previsão para o conjunto de teste(EMP_{teste}), treino(EMP_{treino}) e de validação cruzada(EMP_{cv}). O objetivo é encontrar modelos com boa capacidade de prever novas observações, avaliado pelo EMP_{teste}.

As análises foram desenvolvidas no Software R e toda a abordagem teórica foi baseada nos livros An Introduction to Statistical Learning with Applications in R [2] e The Elements of Statistical Learning [1].

4.3 Resultados

Tendo em vista toda explicação dada a respeito do problema e importância ao se desenvolver e aperfeiçoar métodos para discriminação da madeira mogno para com outras espécies, aplicou-se diversos métodos de SL avaliando a capacidade de previsão dos mesmos e verificando a relevância de se testá-los no contexto supracitado. Foi necessário, em uma primeira etapa, a realização de um pré-processamento e análise descritivas dos dados oriundos de NIRS.

4.3.1 Pré-Processamento

A metodologia utilizada para o pré-processamento dos dados foi replicada de [3], dado que o trabalho citado foi desenvolvido por profissionais da área e, portanto, possui respaldo técnico para se empregar.

O pré-processamento se resumiu em centrar na média e aplicar subtração de linha reta, a Figura X mostra os Espectros médios das amostras.

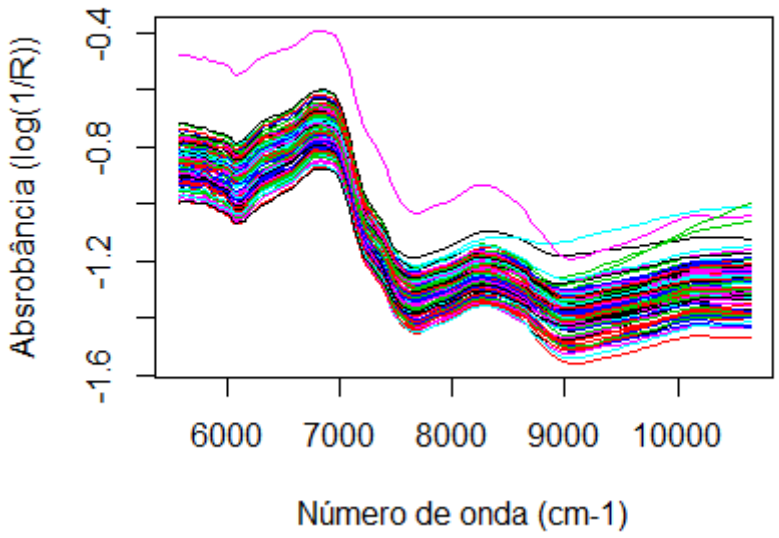


Figura 7 – Espectros médios das amostras de andiroba (azul), cedro (vermelho), curupixá (verde) e mogno (roxo) medidas no equipamento Phazir.

Não se faz necessário identificar as bandas dos espectros para realizar a discriminação pelas técnicas de SL, para uma melhor discussão sobre as origens das principais bandas observadas na região [3].

O pré-processamento pode ter uma influência significativa sobre o resultado, e também se relacionam para o objetivo química ou física da análise.[5]

4.3.2 Análise Descritiva

A análise descritiva é importante para se obter um panorama geral dos dados que irão ser analisados. Com esse tipo de análise identifica-se características importantes para construção dos modelos.

Como o banco de dados utilizado possui 100 variáveis (absorbância em 100 diferente números de ondas) se torna inviável fazer uma análise 1 a 1, dessa forma,restringiu-se em analisar as maiores estatísticas descritivas, a correlação 2 a 2 e, por final, com o objetivo de avaliar classes intrínsecas aos dados, utilizou-se uma técnica não supervisiona, cluster.

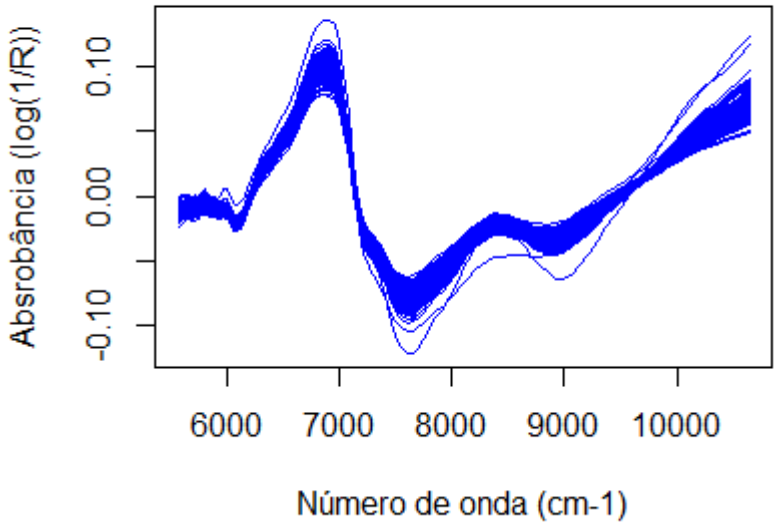


Figura 8 – Espectros do equipamento Phazir pré-processados com subtração de linha reta.

Tabela 7 – Estatísticas descritivas Máximas do conjunto de dados oriundos de NIRS

Estatísticas Máximas	Valores
Média	0,10
Desvio padrão	0,01
Mediana	0,10
Mínimo	-0,12
Máximo	0,13
Desvio Total	0,07

A tabela 7 apresenta os maiores valores de estatísticas descritivas entre todas as variáveis, ou seja, calculou-se as 100 médias e verificou-se o valor máximo, os 100 mínimos e se pegou o menor valor, o maior desvio-padrão entre os 100 e etc. Depreende-se a escala dos dados e que o maior desvio total de absorbância em um número de onda foi de 0.07.

Para se avaliar a relação entre as variáveis, calculou-se a correlação de Pearson 2 a 2, mas pela alta dimensão do problema restringiu-se em apresentar as estatísticas descritivas da intensidade(valor absoluto das correlações), ou seja, calculou-se as estatísticas descritivas das 4950 pares de correlações.

Em média, a intensidade da correlação entre as variáveis foi de 0,57 e 50%

Tabela 8 – Estatísticas Descritivas dos pares de correlações das 100 variáveis dos dados oriundos de NIRS

Mínimo	1st Quartil	Mediana	Média	3rd Quartil	Máximo
0,00	0,32	0,61	0,57	0,83	1,00

possuem correlações acima de 0,61, o que indica que no banco de dados existem variáveis altamente correlacionada, devido a proximidade no número de onda, o que pode gerar complicações em métodos clássicos de discriminação.

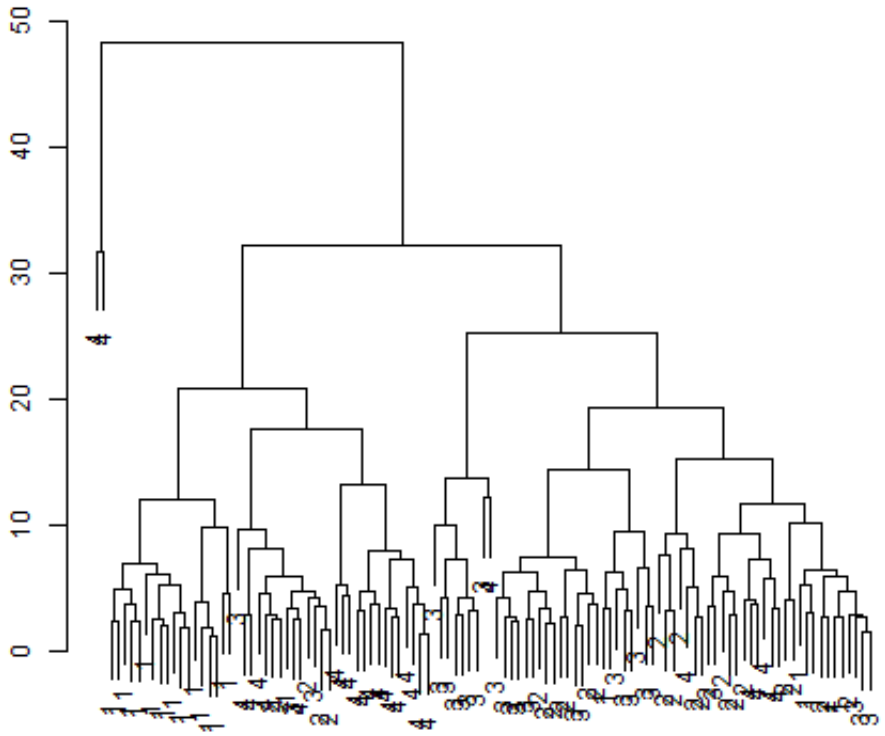


Figura 9 – Dendograma da técnica de Cluster, com distância Euclidiana e Ligação Completa

O dendograma oriundo da técnica de cluster sugere que, por se tratar de uma técnica não supervisionada, os dados já possuem características instríssecas discriminantes dos grupos. Não se agrupou de maneira perfeita as classes mas se verifica alguns agrupamentos

de mesma classe.

4.3.3 Regressão Logística

Regressão Logística é um método clássico de discriminação. Um primeiro modelo foi construído com todas as 100 variáveis presentes no banco, mas devido às já apresentadas altas correlações entre variáveis respostas o resultado evidenciou sobre-ajuste (do inglês *overfitting*).

Incluindo todas as 100 regiões como variáveis respostas o modelo não errou nenhum classificação para o conjunto de treino, mas, em relação ao conjunto de teste, errou 48% das vezes. Isso mostra o sobre-ajuste do modelo, se classifica muito bem para dados dentro da amostra, mas, para observações novas não é um bom discriminante.

Dessa forma, se fez a utilização da técnicas de seleção de variáveis *stepwise* seguindo a direção *Backward* e utilizando como critério de escolha AIC. O modelo final teve bom resultados, apresentando nenhum erro de classificação no conjunto de treino e errou $3,7\%$ no conjunto de teste.

4.3.4 SVM

A técnica de SVM permite a inclusão de kernel's que podem ser determinantes para geração de boas estimativas. Para o conjunto de dados do problema foram aplicados três Kernels: Linear, Radial e Polinomial; cada um com parâmetros específicos estimados por validação Cruzada e, para cada caso, considera-se diferente vetores suportes.

Tabela 9 – Estimativas por validação Cruzada dos Parâmetros de cada Kernel Utilizado e o Número total de vetores suportes selecionados

Kernel	Custo	Grau	γ	coeficiente	Nº Vetores suportes
Polinomial	0,50	1,00	0,50	0,50	19
Linear	0,50	-	-	-	14
Radial	0,50	-	0,50	-	44

A etapa principal da presente análise é avaliar a capacidade dos modelos em discriminar a espécie mogno das demais, se obteve resultados relevantes.

Avaliando os resultados verifica-se que o Kernel Linear obteve o melhor seguido do polinomial, os resultados para o kernel radial sugere que o mesmo não se encaixou ao

Tabela 10 – Erros de Teste, Treino e Validação Cruzada segundo Kernel para o Método de *SVM*

Kernel	EMP_{teste}	EMP_{treino}	EMP_{cv}
Linear	0,000	0,010	0,013
Polinomial	0,000	0,010	0,030
Radial	0,250	0,250	0,220

problema.

Para o Kernel Linear, o impressionante resultado de nenhum erro para o conjunto de teste e o baixo valor de erro de treino e de validação cruzada, 1% e $1,3\%$ respectivamente, demonstram o poder do método. O mesmo pode ser dito para o Kernel Polinomial onde se obteve diferente resultado apenas no erro de validação cruzada, 3%.

Dado que o método foi capaz de prever corretamente todas as observações fora da amostra, mostra-se que o mesmo aliado a NIRS, mesmo com equipamento portátil, é efetivo para discriminar o mogno de outras espécies semelhantes a olho nu e, assim, auxiliar de maneira fundamental o trabalho de fiscalização.

4.3.5 Regressão com penalização

Os modelos de regressão com penalização, Lasso e Ridge, necessitam da estimação do parâmetro de penalização λ, e, para os dois casos, estimou-se por validação cruzada de 10 grupos.

A figura 10 ilustra o processo de estimação de λ. Os valores que deram o menor erro de de validação cruzada foram de $\lambda = 0,003$ para LASSO e $\lambda = 3,14$ para RIDGE.

Tabela 11 – Erros de Teste, Treino e Validação Cruzada segundo Método de Penalização

Método	EMP_{teste}	EMP_{treino}	EMP_{cv}
LASSO	0,00	0,01	0,03
RIDGE	0,14	0,21	0,23

depreende-se da tabela 11 que a regressão ridge não se mostrou tão eficiente para o problema (erro de treino de 14%) mas, por outro lado, a regressão LASSO não errou nenhuma classificação fora amostra e apresentou baixos erros de teste e treino, 1% e 4% respectivamente.

Os resultados demonstram que a regressão LASSO é uma técnica válida para

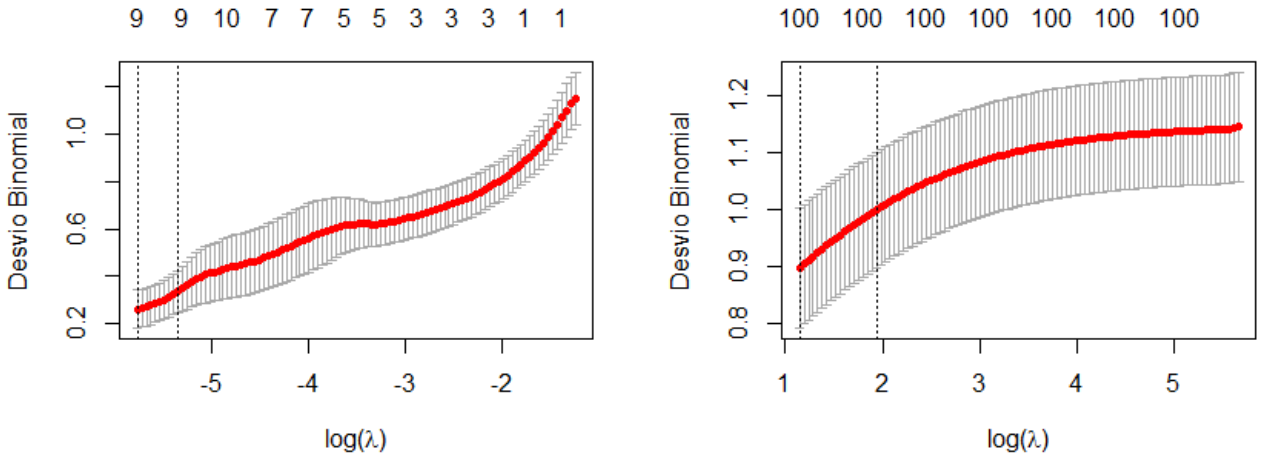


Figura 10 – Curva do erro de validação cruzada para evolução de $\log(\lambda)$, e as curvas de desvio padrão superiores e inferiores - Regressão LASSO à esquerda e Ridge à direita

discriminar mogno de outras espécies com dados oriundos de NIRS.

Vale ressaltar a propriedade esparsa de LASSO; dado que apenas 9 variáveis não tiveram seus coeficientes zerados. O números de ondas que mostraram-se relevantes para o método foram: $5590,965cm^{-1}; 5828,525cm^{-1}; 5913,311cm^{-1};$
$6061,341cm^{-1}; 6122,949cm^{-1}; 7168,459cm^{-1}; 8202,100cm^{-1}; 8922,198cm^{-1}; 10643,960^{-1}$

4.3.6 Árvore de Decisão

O método de Árvore de decisão tende a não ter um bom poder de previsão comparado a métodos como SVM e regressão com penalização, mas existem método de reamostragem que, aplicados a Árvore de Decisão, aumentam a capacidade de previsão. Dessa forma, com o objetivo de avaliar tais métodos de reamostragem em uma primeira etapa se verificou os resultados de AD.

A figura 11 indica que as regiões de absorbância $6091,989cm^{-1}$; $6122,949cm^{-1}; 5971,575cm^{-1}; 7124,537cm^{-1}e10643,960cm^{-1}$ são importantes para discriminação da classe, essa é a característica do método em possuir uma troca em poder de previsão e capacidade de interpretação. O processo de discriminação do método mostra que se absorbância na região $6091,989cm^{-1}$ for maior que $-0,0187727$ a observação será classificada como 1,pertencente a classe, e, se for menor, será avaliado as outras ramificações: se na

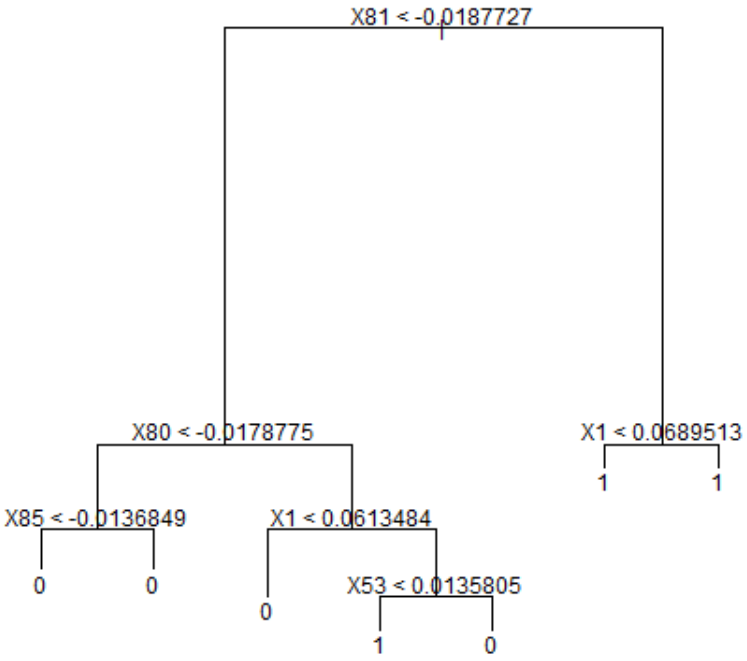


Figura 11 – Resultado da técnica de árvore de decisão, indica 5 regiões para construção dos nós. ($X81 = 6091,989cm^{-1}, X80 = 6122,949cm^{-1}, X85 = 5971,575cm^{-1}, X53 = 7124,537cm^{-1}, X1 = 10643,960cm^{-1}$)

região $6122,949cm^{-1}$ for menor que $-0,0178775$ será classificada como 0 e a interpretação segue para as outras ramificações.

Uma característica interessante no resultado apresentando é que para o caso em que a absorbância na região $6091,989cm^{-1}$ for maior que $-0,0187727$ o próximo nó a ser avaliado é para a região $10643,960cm^{-1}$ mas independente do lado o resultado será classificar em 1. Isso se chama nó pureza, intuitivamente ele não possui grande utilidade mas no processo de construção da árvore ele melhora o índice de Gini e a entropia cruzada, que são outras medidas utilizados para cálculo do erro de previsão durante o processo.[2]

4.3.7 *BAGGING*

O utilização da técnica de *Bagging* aliado a árvore de decisão tende a produzir um classificador com alto poder de previsão. Dessa forma, foi realizado o procedimento com

B (número de árvores) igual a 1000, B não é um parâmetro crítico para análise dado que para altos valores não se gera sobre-ajuste.

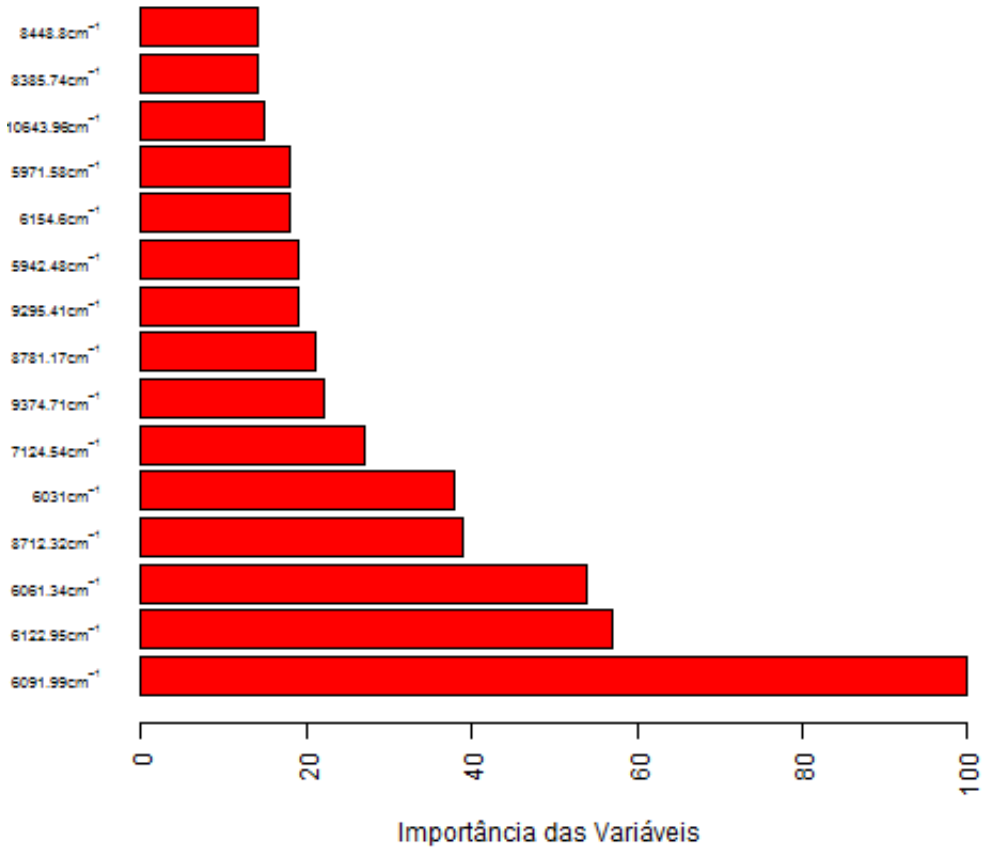


Figura 12 – 15 regiões mais importantes detectadas por *Bagging*, a importância da variável é calculado usando a diminuição média no índice de Gini a cada divisão, e é expressa em relação ao máximo

A figura 12 mostra as 15 regiões mais importantes para discriminação, segundo método de Bagging. Percebe-se que as regiões de Árvore de decisão aparecem novamente como regiões importantes para discriminação, dessa forma se tem indicações de em quais regiões a absorbância da espécie mogno se difere das demais.

A região $6091.99cm^{-1}$ apareceu como a mais importante para diminuição do erro médio de previsão, medido pelo índice de Gini avaliando a presença da mesma em cada nó nas 1000 árvores geradas, ela foi seguida da região $6122.95cm^{-1}$ que apresentou cerca de

60% de importância daquela.

4.3.8 *Floresta Aleatória*

Dada o objetivo de diminuir a correlação entre as estimativas de *Bagging* e, assim, diminuir a variância do resultado total, aplicou-se, também, o procedimento de Florestas Aleatórias. A literatura sugere que a amostra dos p preditores em cada árvore seja de tamanho $\sqrt{p}$, dessa forma, se escolheu $p = 10$. Assim como no método anterior foi considerado $B = 1000$.

Métodos	EMP_{treino}	EMP_{teste}
Árvore de Decisão	0,05	0,11
Bagging	0,00	0,00
Florestas Aleatória	0,00	0,00

Tabela 12 – Resultados Para os Métodos de Árvore de Decisão, *Bagging* e Florestas Aleatória

Como os métodos não utilizaram validação cruzada não se calculou EMP_{cv}. Árvore de Decisão errou 5% no conjunto de treino e 11% no de teste, resultado já esperado pelas suas propriedades. *Bagging* e Florestas Aleatórias não errou nenhuma previsão de classe tanto para o conjunto de treino quanto para o conjunto de teste, evidenciando o poder de previsão dos métodos. E, novamente, os resultados apresentados sugerem que os dois últimos métodos também podem ser considerados no contexto de quimiometria.

4.3.9 Mínimos Quadrados Parciais

Mínimos Quadrados Parciais(PLS) é amplamente utilizado na esfera de quimiometria [6], o método tem uma boa capacidade de previsão o que justifica seu uso mas em alguns casos ele não tem ganho evidentes sobre métodos clássicos e sua utilização deve ser seguida de alguns pressupostos[7]. Como continuação do exercício de se testar métodos de SL no problema de discriminação da espécie mogno, aplicou-se a técnica ao problema.

Analisando a Figura 13 percebe-se que após o componente 15 o EMP_{cv} volta a crescer, evidenciando um possível sobre-ajuste. Selecionou-se o número de componentes que gerou o menor EMP_{cv}, 12.

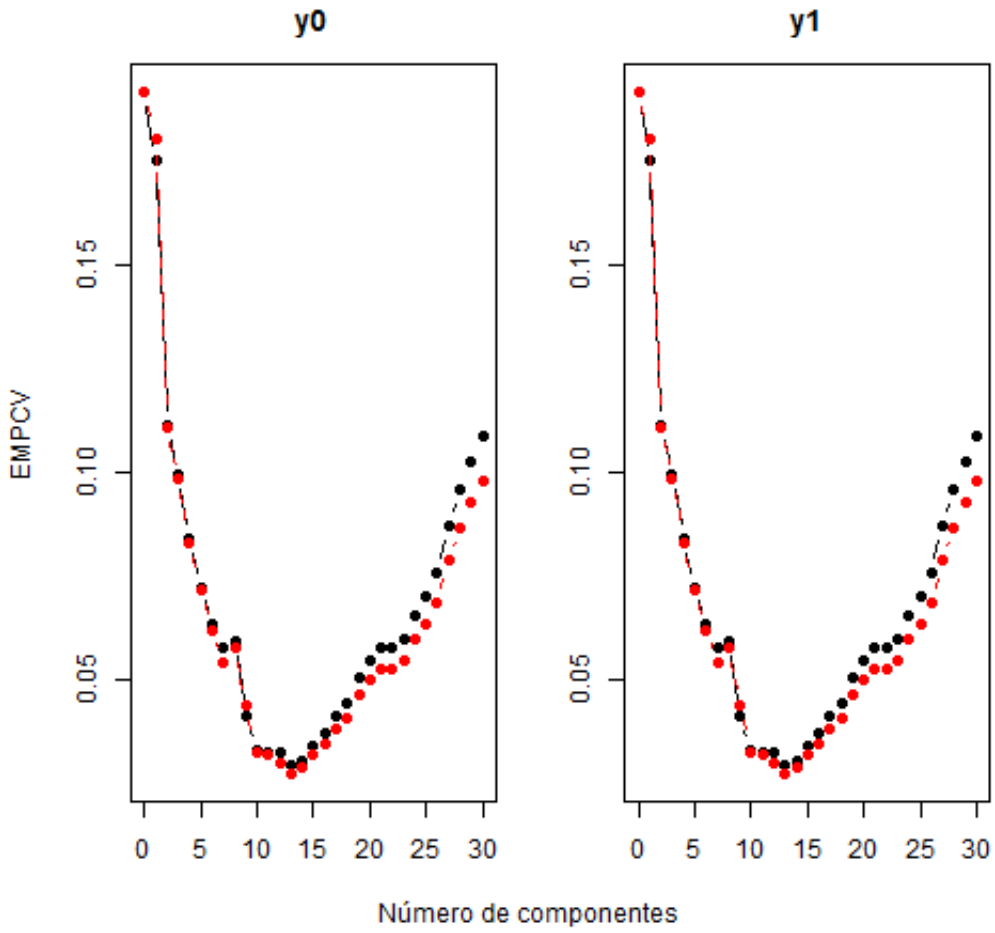


Figura 13 – Validação cruzada para PLS, Erro Médio de Previsão de Validação Cruzada segundo Número de Componentes. A termologia $y0$ e $y1$ é proveniente do tratamento de um problema multivariado

4.3.10 Regressão Com Componentes Principais

O método de Regressão com Componentes Principais (PCR) não é adequado para um problema de classificação, ele interpreta a variável resposta com classes como variável contínua, mas apenas como exercício aplicou-se o método ao problema. Uma maneira mais adequada de se utilizar Componentes Principais é utilizar a técnica juntamente com Discriminante Linear.

Segundo a raiz do erro quadrático médio de previsão, selecionou-se 28 componentes. As previsões de PCR não serão as classes 0 ou 1, mas sim um valor numérico, dessa forma classificou-se como 0 resultados que deram abaixo de $0,5$ e como 1 valores acima

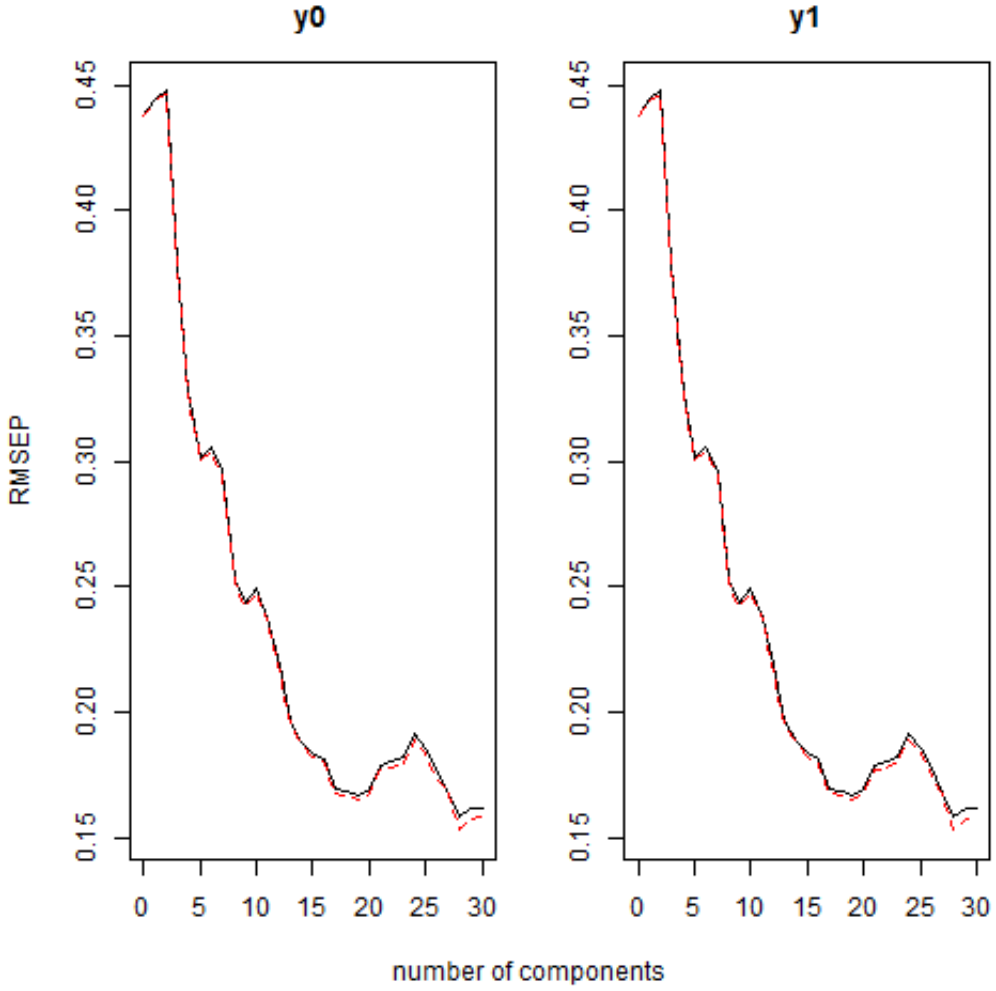


Figura 14 – Validação cruzada para PCR, Erro Quadrático Médio de Previsão de Validação Cruzada segundo Número de Componentes. A termologia $y0ey1$ é utilizada por se tratar de um problema multivariado

Tabela 13 – Erros de Teste, Treino e Validação Cruzada para os Métodos de Redução de Dimensão

Método	EMP_{teste}	EMP_{treino}	EMP_{cv}
PLS	0,00	0,00	0,03
PCR	0,00	0,00	0,03

Verifica-se o poder de predição do método de PLS, o mesmo não errou nenhuma classificação tanto no conjunto de treino e teste e apenas 3% no conjunto de validação. Mesmo não sendo intuitivamente adequado ao problema PCR respondeu muito bem, não errou nenhuma classificação no conjunto de teste e de treino.

O resultado justifica a ampla utilização do método a problemas de quimiometria

4.4 Conclusão

Os resultados mostraram a capacidade dos métodos de aprendizagem em gerar boas estimativas. Essa etapa foi interessante por todo cunho ambiental envolvido: para aperfeiçoar o trabalho de fiscalização da exploração ilegal de um espécie ameaçada de extinção, utilizar técnicas estatísticas de discriminação. Mostra a importância que a área pode exercer em vários ramos práticos.

Para essa segunda etapa se teve resultados satisfatórios para muitas das técnicas testadas, não se errou nenhuma classificação fora da amostra para PLS, PCR, SVM, LASSO, *BAGGING* e Florestas Aleatórias, evidenciando, mais uma vez, a relevância das técnicas de aprendizado a problemas de quimiometria e, mais especificamente, a dados oriundos de NIRS.

A técnica de PLS foi uma das que apresentou melhores resultados, justificando, em partes, a sua popularidade na área quimiométrica. Mas, em determinadas circunstâncias, ela não se sobressai a métodos clássicos como Distância Euclidiana e Análise de Discriminante Linear [7], e os resultados semelhantes à outros métodos de SL evidencia que a restrição em utilizar apenas esse método na área é negligente.

Para esse problema vale ressaltar que as técnicas de LASSO, *BAGGING* e Florestas Aleatórias além de possuírem boa capacidade de previsão ainda possuem boa de interpretação, e, dado que obtiveram a mesmas performances que as outras (Erro de Teste), se pode entender como técnicas mais recomendáveis.

5 CONSIDERAÇÕES FINAIS

Após análise dos resultados verificou-se a aplicabilidade das técnicas de Aprendizagem Estatística em problemas de quimiometria, em muito casos gerando estimativas iguais ou melhores que técnicas ditas clássicas e usuais na área.

Para o problema da variável Lipossubilidade , primeira etapa do trabalho, verificou-se expressivos ganhos na diminuição do erro quadrático médio ao se utilizar técnicas de aprendizado, e, para esse caso, a técnica de regressão LASSO obteve melhor performance. Os resultados sugerem que tais métodos são adequados para construção de modelos para estudos de relação quantitativa entre estrutura e atividade ($QSAR$) amplamente utilizados na indústria farmacêutica.

Na segunda etapa do trabalho, se obteve excelentes resultados para o problema de discriminante. Os resultados estão em linha com o objetivo do trabalho ao se mostrar que várias técnicas de SL se adequaram ao problema e evidenciar que existe um restrição na área ao focar apenas em um método quando se existe uma gama que pode se encaixar melhor em determinados casos.

Em geral, os métodos de aprendizagem realmente se mostraram eficazes tanto para problemas de regressão quanto para de discriminação. Essa área tende a crescer cada vez mais e se mostra como uma grande evolução para várias áreas científicas, sendo importante a sua difusão em ambiente académico.

REFERÊNCIAS

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor. An Introduction to Statistical Learning: With Applications in R. 2014.

Bergo, Maria Cecília Jorge. Transferência de calibração na discriminação de mogno e espécies semelhantes utilizando NIRS e PLS-DA. Diss. Universidade de Brasília, 2014.

Mevik, Bjorn-Helge, Ron Wehrens, and T. N. San Michele all?Adige. "Introduction to the pls Package."(2015).

Brereton, Richard G. Applied chemometrics for scientists. John Wiley & Sons, 2007.

PASTORE, Tereza Cristina Monteiro et al. Near infrared spectroscopy (NIRS) as a potential tool for monitoring trade of similar woods: Discrimination of true mahogany, cedar, andiroba, and curupixá. 2011.

Brereton, Richard G., and Gavin R. Lloyd. "Partial least squares discriminant analysis: taking the magic away."Journal of Chemometrics 28.4 (2014): 213-225.

PAVIA, Donald et al. Introduction to spectroscopy. Cengage Learning, 2008.

LOPES, Wilson Araújo; FASCIO, Miguel. Esquema para interpretação de espectros de substâncias orgânicas na região do infravermelho. Química Nova, v. 27, n. 4, p. 670-673, 2004.

BARBOSA, L.C. Espectroscopia no Infravermelho na Caracterização de Compostos Orgânicos. Ed. da UFV: Viçosa. 189 p., 2007.

TRINDADE, M. M. et al. Espectroscopia no infravermelho por reflexão total atenuada horizontal (HATR) aplicada na identificação de óleos vegetais comerciais.