Universidade Federal de São Carlos – UFSCar
Centro de Ciências Exatas e de Tecnologia
Departamento de Engenharia Química – DEQ

Estudo das Aplicações de Data Science e Machine Learning na Engenharia Química

Sofia Miran Han

Trabalho de Graduação apresentado ao
Departamento de Engenharia Química
da Universidade Federal de São Carlos.

Orientador: Prof. Dr. Antônio Carlos Luperni Horta

São Carlos - SP
2022

BANCA EXAMINADORA

Trabalho de Graduação apresentado no dia 13 de setembro de 2022 perante a seguinte banca examinadora:

Orientador: Prof. Dr. Antônio Carlos Luperni Horta - Departamento de Engenharia Química - Universidade Federal de São Carlos.

Convidado: Prof. Dr. João Paulo Silva Queiroz - Departamento de Engenharia Química - Universidade Federal de São Carlos.

Professor da Disciplina: Prof. Dr. Ruy de Sousa Junior - Departamento de Engenharia Química - Universidade Federal de São Carlos.

AGRADECIMENTOS

Agradeço à minha querida mãe, que sempre esteve ao meu lado durante os momentos mais difíceis da minha jornada e ao meu pai, que me incentivou na busca por desafios.

RESUMO

Atualmente, com o avanço da tecnologia e da capacidade dos computadores em armazenar enormes quantidades de dados, empresas e indústrias tem buscados profissionais capacitados para lidar com o desafio de organizar e extrair valores significativos dos mesmos e que confiram vantagens competitivas como previsão de demandas, detecção de falhas e exploração de padrões. A área que lida com esse ramo é chamada Ciência de Dados, e emprega uma combinação de recursos estatísticos, matemáticos e computacionais para realizar inferências sobre padrões encontrados nos dados, que dificilmente seriam detectados apenas pela análise humana. Aliada a ela, o uso de algoritmos capazes de aprender e predizer padrões chamados de Machine Learning estão impulsionando o ramo ao introduzir inteligência artificial para a resolução desse tipo de problema. O avanço da computação, bem como a popularização da programação e o surgimento de bibliotecas matemáticas completas tem tornado o uso dessas ferramentas cada vez mais cotidianas para os engenheiros químicos, em especial nas áreas de análise e otimização de processos. Nesse contexto, o trabalho aqui apresentado teve como objetivo realizar uma revisão bibliográfica e teórica do atual contexto dos dados na engenharia química, elucidar a teoria por trás de algoritmos de aprendizado de máquina supervisionado, suas vantagens, limitações e aplicações perante 3 estudos de caso pertinentes a engenharia química. Os resultados obtidos demonstraram que a acurácia dos modelos estudados variam de acordo com a aplicação que lhes é conferida, tendo, de modo geral, algoritmos mais flexíveis (Modelos de Árvore e Redes Neurais Artificiais) tido desempenho melhor para problemas não lineares e de classificação, enquanto algoritmos mais rígidos (Regressão Linear Simples e Múltipla) obtiveram melhor desempenho em experimentos no qual a natureza da relação entre as variáveis de entrada e saída já era conhecidamente linear. O conhecimento de relações físico-químicas dos experimentos em análise também contribuíram para a construção de modelos mais acurados.

Palavras – chave: Engenharia química, Machine Learning, Data Science, Aprendizado Supervisionado.

ABSTRACT

Today, with the advancement of technology and the ability of computers to store huge amounts of data, companies and industries have sought out professionals trained to deal with the challenge of organizing and extracting meaningful insights from those data and conferring competitive advantages such as demand prediction, fault detection and exploitation of standards. The area that deals with this branch is called Data Science, and employs a combination of statistical, mathematical and computational resources to make inferences about patterns found at the data, which would hardly be detected only by human analysis. Combined with it, the use of algorithms capable of learning and predicting patterns called Machine Learning are driving the industry by introducing artificial intelligence to solve this type of problem. The advancement of computing, as well as the popularization of programming and the emergence of complete mathematical libraries has made the use of these tools increasingly everyday for chemical engineers, especially in the areas of process analysis and optimization. In this context, the work presented here aimed to perform a bibliographic and theoretical review of the current context of data in chemical engineering to elucidate the theory behind supervised machine learning algorithms, their advantages, limitations and applications before 3 case studies relevant to chemical engineering. The results showed that the accuracy of the models studied vary according to the application given to them, having, in general, more flexible algorithms (Tree Models and Artificial Neural Networks) performed better for nonlinear and classification problems, while more rigid algorithms (Simple and Multiple Linear Regression) performed better in experiments in which the nature of the relationship between the input and output variables was already known to be linear. The knowledge of physical-chemical relationships of the experiments under analysis also contributed to the construction of more accurate models.

Keywords: Chemical Engineering, Machine Learning, Data Science, Supervised Learning.

SUMÁRIO

1. INTRODUÇÃO……...10

2. REVISÃO BIBLIOGRÁFICA……11

2.1 DADOS E MACHINE LEARNING NA ENGENHARIA QUÍMICA.....11

2.2 CONCEITOS DE APRENDIZADO ESTATÍSTICO.....17

2.3 MÉTODOS DE AMOSTRAGEM.....22

2.4 REDES NEURAIS ARTIFICIAIS.....29

2.5 MÉTODOS DE ÁRVORE.....37

2.6 REGRESSÃO LINEAR.....44

2.7 REGRESSÃO LOGÍSTICA.....57

3. METODOLOGIA……....62

4. RESULTADOS E DISCUSSÕES……...65

4.1 PREDIÇÃO DE CONSUMO DE ELETRICIDADE.....65

4.2 PREDIÇÃO DA CINÉTICA DE DISSOLUÇÃO DE VIDRO DE SILICA.....68

4.3 PREDIÇÃO DA CONTAMINAÇÃO POR FLUOR EM LENÇÓIS FREÁTICOS.....70

5. CONCLUSÃO…….73

6. REFERÊNCIAS……...74

LISTA DE FIGURAS

Figura 2.1 Emprego de ML para problemas de otimizaçã de dados na engenharia…….……14

Figura 2.2. Comparação de modelo de ML híbrido e *black – box*……..……15

Figura 2.3. Base de observações e seus respectivos erros……..…..18

Figura 2.4. Modelo paramétrico ajustado versus modelo não-paramétrico

com overfitting…….....19

Figura 2.5. Comparativo entre modelos hipotéticos, a função real

e o MSE de treino e teste……..20

Figura 2.6. Esquema de uma versão simplificada da validação cruzada…….……23

Figura 2.7. Validação cruzada simples com divisão binária feita

uma vez e repetida dez vezes……..……24

Figura 2.8. Esquema de validação cruzada LOOCV……25

Figura 2.9. Curva de erro do LOOCV para 1 divisão e para 9 divisões……...25

Figura 2.10. Erro MSE real e de teste para dados simulados……...28

Figura 2.11. Esquema de amostragem bootstrap……..32

Figura 2.12. Representação de função de ativação limiar, linear por partes e sigmóide com inclinação variável *a*…….32

Figura 2.13. Esquema de descida do gradiente……33

Figura 2.14. Esquema de MPL com duas camadas intermediárias……..35

Figura 2.15. Esquema de RNN…….37

Figura 2.16. Estrutura de uma árvore de decisão…….38

Figura 2.17. Esquema de um espaço preditor……..39

Figura 2.18. Dados com alto e baixo ganho de informação…….40

Figura 2.19. Esquema de uma floresta aleatória……..44

Figura 2.20. Regressão linear utilizando-se a técnica dos mínimos quadrados……...46

Figura 2.21. Esquema da curva de regressão linear da população e da curva obtida pelos mínimos quadrados……..47

Figura 2.22. Aplicação da Regressão Ridge a um conjunto de dados……..53

Figura 2.23. Troca viés-variância para a regressão Ridge……54

Figura 2.24. Aplicação da regressão Lasso a um conjunto de dados……...55

Figura 2.25. Aplicação de LOOCV para seleção de λ.……56

Figura 2.27. Diferença de uma classificação logística e linear……58

Figura 2.28. Regressão logística múltipla com evidência de *coufounding*……..61

Figura 4.1. Predição da taxa de dissolução do silicato em modelo totalmente a base de dados e com introdução de informação físico-química……..69

LISTA DE TABELAS

Tabela 2.1. Coeficientes de regressão linear simples para os meios

A e B separadamente……50

Tabela 2.2. Coeficientes para regressão linear múltipla aplicados para os meios A, B e C simultaneamente……...51

Tabela 4.1. Comparação dos erros RASE associados a cada método……..…67

Tabela 4.2. Fatores significativos na demanda energética por método……70

Tabela 4.3. Seleção das entradas utilizando-se o teste do chi-quadrado……..70

Tabela 4.4. Acuracidade, sensibilidade e erro para os modelos de ML……...…72

1. INTRODUÇÃO

O presente trabalho teve por objetivo realizar uma revisão bibliográfica para melhor compreensão de como o uso do Aprendizado de Máquina ou *Machine Learning* tem impactado a indústria química, assim como discutir quais são ainda as limitações dos métodos de aprendizagem quando comparados às tradicionais modelagens físico-químicas, assim como os benefícios que esses artifícios podem trazer para o meio. Em um primeiro momento, realizou-se um levantamento teórico por trás de 4 algoritmos de aprendizagem supervisionada bastante empregados atualmente: Redes Neurais, Métodos de Árvore, Regressão Linear e Regressão Logística. A partir do conhecimento sobre seu funcionamento, em um segundo momento, buscou-se exemplificar seus usos por meio de artigos contendo experimentos reais performados com o auxílio dessas ferramentas.

As empresas do setor químico estão passando por mudanças, enfrentando maior pressão no quesito competitividade e demandando processos mais otimizados e ciclos de operação mais curtos (SCHWEIDTMANN et al. , 2021). Convencionalmente, as decisões em engenharia química baseiam-se no desenvolvimento de modelos físicos experimentais para fins de estudo e otimização. No entanto, o desenvolvimento de sistemas químico-físicos é caro e muitos fenômenos não podem ser completamente descritos por modelos computacionais. Nesse sentido, o uso de *Machine Learning* (ML) torna-se uma opção interessante, principalmente no que diz respeito a optimização de alguns processos, dado que seus algoritmos são capazes de aprender comportamentos complexos a custos mais baixos. O ML na engenharia química já experimentou duas ondas entre os anos 1980 e 2008, tendo-se destacado, principalmente, o uso de redes neurais artificiais (ANNs) mais simples (SCHWEIDTMANN et al. , 2021). No entanto, o impacto dessas ondas foi amenizado devido a alguns fatores, como a falta de dados acessíveis, capacidade computacional limitada e o surgimento de tecnologias competitivas na engenharia química, particularmente, modelos mecanicistas, modelos de optimização e modelos de controle preditivo.

Em seu artigo, ALMEIDA (2016) destaca um estudo no qual empregou-se o uso de ML para auxiliar na modelagem de um processo de tratamento de efluentes orgânicos visando a predição da concentração de nitrogênio em um biorreator. Devido a complexidade dos fenômenos que ocorrem no reator e a dificuldade de modelagem, optou-se pelo uso do ML dado a extensa base de dados adquirida. Esse caso fornece um bom exemplo de uso do ML para otimização, dado que o interesse maior não é o entendimento do processo, mas sim

buscar uma correlação entre as variáveis de entrada com uma variável resposta e, dessa forma, tornar o processo mais eficiente. No estudo, optou-se pela modelagem via ANNs definindo-se 11 variáveis de entrada e 1 de saída (concentração final de nitrogênio no reator) formando-se uma arquitetura de 11 neurônios na camada de entrada, 27 na camada oculta e 1 na camada de saída. A performance do algoritmo detectou a realização de um procedimento operacional que resultou em aumentos bruscos na concentração de nitrogênio, no caso, a etapa de adição de antiespumante, sendo possível a partir dessa conclusão, eliminar a ação e melhorar o desempenho do biorreator. Esse exemplo pode ser considerado um caso de ML black-box no qual não se sabe bem que tipo de decisões o algoritmo toma ao longo do processo, porém, suas saídas são utilizadas como insights na compreensão de um problema.

Atualmente, dispõe-se de ferramentas acessíveis e poderosas, além de ambientes amigáveis para programação e comunidades de código aberto que facilitam a troca de informação e a implementação de modelos. A engenharia química passa por um momento de digitalização e automação tanto da pesquisa quanto da indústria (SCHWEIDTMANN et al. , 2021). O uso de ML para detecção de anomalias, como no exemplo do biorreator com excesso de nitrogênio, tem se tornado comuns e aumentado as possibilidades de aplicações comerciais e técnicas na indústria química. SCHWEIDTMANN et al.(2021) cita outras possibilidades já também disponíveis para mercado, como algoritmos de clusterização na detecção de bateladas normais e com falhas, além do uso de ML para construção de mapas auto-organizáveis para monitoramento de estações de tratamento de águas residuárias.

Mostrando-se uma área inovadora, o ML possui uma ampla variedade de aplicações e que em muito podem contribuir para o avanço da engenharia química em uma era no qual os dados tornam-se um recurso precioso e novas tecnologias para seu armazenamento e processamento encontram-se disponíveis, motivando, portanto, a realização do estudo.

2. REVISÃO BIBLIOGRÁFICA

2.1 DADOS E MACHINE LEARNING NA ENGENHARIA QUÍMICA

Segundo SCHWEIDTMANN et al.(2021), o ML subdivide-se em duas categorias: o aprendizado supervisionado e não-supervisionado. Cada uma dessas técnicas é empregada para propósitos distintos. O aprendizado não-supervisionado refere-se a técnicas de investigação de dados não rotulados, ou seja, as variáveis não possuem uma relação de

entrada-resposta. O propósito principal, nesse caso, é encontrar padrões escondidos nos dados, sendo utilizado para fins de agrupamento ou clustering, extração de características comuns e detecção de anomalias. Já o aprendizado supervisionado utiliza-se de modelos de treinamento em variáveis rotuladas visando uma correlação entre as variáveis de entrada (ou preditivas) e uma variável resposta. As técnicas de regressão são um bom exemplo desse tipo de aprendizado e possui amplas aplicações na engenharia de processos e têm sido utilizadas com a finalidade de se desenhar processos de otimização, como por exemplo, sensores e predição de qualidade, utilizando-se de métodos como o Ajuste dos Mínimos Quadrados.

As ANNs têm sido utilizadas para problemas complexos de larga escala tal como processos com unidades de separação de gases, osmose reversa de água salgada e outros processos químicos (SCHWEIDTMANN et al. , 2021). O aprendizado supervisionado também tem sido utilizado há muito tempo para controle de sistemas dinâmicos em operação, dado que uma ampla variedade de modelos são empregados para se descrever dados em tempo discreto ou contínuo. Para tais abordagens, o uso de Redes Neurais Recorrentes tem se mostrado útil em modelos de operação para biorreatores operando em batelada, dado que essa arquitetura específica de Redes Neurais possuem capacidade de memória temporal (SCHWEIDTMANN et al. , 2021).

Além dos exemplos apresentados, SCHWEIDTMANN et al.(2021) destaca algumas categorias de problemas emergentes em ML aplicados a engenharia química que merecem destaque:

- Otimização de decisões:

A otimização é um tópico proeminente na engenharia química para processos de síntese, controle seleção de solventes, catalisadores e adsorventes. Todas essas decisões devem se basear em informações existentes, seja na forma de dados ou modelos mecanicistas já existentes. Como mostrado na figura 2.1, a otimização na tomada de decisão pode ser feita baseada em dados com o treinamento necessário ou modelos híbridos que contam com decisão baseada em dados respaldados por modelos teóricos.

Na engenharia de processos, vários modelos baseados em dados têm sido utilizados para regressão e otimização sequencialmente. A complexidade do modelo à base de dados varia de aproximações lineares a ANNs de aprendizado profundo. Durante muito tempo, a literatura focou em modelos lineares para aproximar dados de simulação e dados experimentais. Desde 1990, ANNs mais simples têm sido usadas para se aproximar modelos

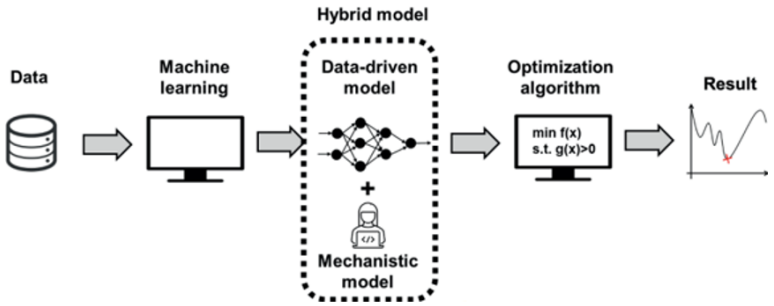
não-lineares dado um número suficiente de neurônios em sua arquitetura. Em muitas aplicações, ANNs são usadas para complementar o processo por meio de uma abordagem *black-box* (ou caixa-preta) e combinadas a modelos físico-mecânicos.

Com a popularização do ML, muitos algoritmos tornaram-se baseados em ANNs e no *Big Data* (bases de dados extremamente grandes). o aprendizado profundo das ANNs provavelmente se tornarão mais importantes nos processos de engenharia química, pois a quantidade de dados disponíveis tenderá a ser cada vez maior através de inovações como manufatura inteligente. No entanto, aplicações de ANNs ainda são limitadas no desenho de processos. Enquanto problemas de otimização com abordagens lineares podem ser resolvidos em larga escala, para problemas de dimensões maiores e não-lineares não conseguem ser "aprendidos" com acurácia. Em contrapartida, modelos de maior complexidade com ANNs são limitados a problemas locais de natureza estocástica (SCHWEIDTMANN et al. , 2021).

THEBELT et al. (2022) destaca um exemplo de uso de ML para otimização de performance em catálise realizado na BASF. A modelagem da eficiência de uma catálise química é uma tarefa morosa que envolve modelos altamente não-lineares e variam de acordo com a aplicação destinada. Modelos de árvores e ensembles (outro algoritmo conhecido de ML a ser aprofundado mais a frente) são populares para modelagem desses sistemas devido a sua excelente acurácia preditiva. Além do mais, esses algoritmos são eficientes e permitem modelagens baratas. Ao combinar técnicas tradicionais com ML para garantir a viabilidade das propriedades catalíticas enquanto se otimiza o poder preditivo com a aplicação de árvores e ensembles treinadas com dados, as indústrias conseguem desenhar processos promissores enquanto reduzem a quantidade de recursos para a construção de modelos.

.

Figura 2.1 Emprego de ML para problemas de otimização de dados na engenharia



Fonte: Adaptado de SCHWEIDTMANN et al.(2021)

●Complementação de modelos físicos com ML:

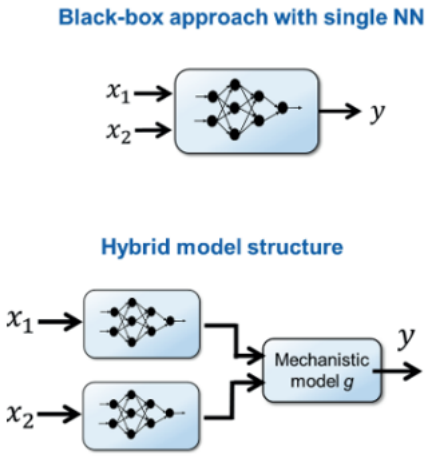
SCHWEIDTMANN et al.(2021) informa que o uso de sistemas físicos complementares são frequentemente necessários para o desenvolvimento de modelos na engenharia química. Em muitas disciplinas, utiliza-se a abordagem *black-box* com aprendizagem supervisionada, ou seja, com dados externos para fins de comparação das respostas obtidas pelos modelos de ML. No entanto, abordagens *black-box* possuem sérios problemas de interpretabilidade, extrapolação, demanda por dados e confiabilidade. Essas limitações podem resultar em erros fatais quando não realizados os devidos checks. Por outro lado, modelos mecânicos e físico-químicos podem prover conhecimento estrutural para serem combinados com modelos baseados em dados.

A combinação de modelos físicos com modelos baseados em dados é chamado de híbrido (ou semi-paramétrico). Dentre suas vantagens, destaca-se melhor interpretabilidade, melhora das propriedades de extrapolação e melhor capacidade de previsão mais acurada (SCHWEIDTMANN et al. , 2021).

No exemplo dado na figura 2.2, os dados de treinamento são analisados em uma abordagem *black-box (*utilizando-se ANNS com dois inputs e um output) e uma abordagem híbrida com com uma função *g(f(x1),f(x2))* no qual *f(x1)* e *f(x2)* são funções desconhecidas e g é uma função física conhecida. Isso permite a construção de um modelo em que cada *black-box* possui apenas uma variável de input e, portanto, é avaliada dentro do alcance permitido pelos dados de treino. Consequentemente, o modelo híbrido evita extrapolações de cada *black-box* dentro do modelo.

Modelos híbridos possuem inúmeras aplicações em engenharia química e biotecnologia desde o início de 1990 em modelagem de reatores, polimerização, cristalização, destilação, processos de secagem e controle (SCHWEIDTMANN et al., 2021).

Figura 2.2. Comparação de modelo de ML híbrido e *black - box*



Fonte: Adaptado de SCHWEIDTMANN et al.(2021)

Em outras palavras, os modelos híbridos tendem a oferecer melhor interpretabilidade comparados com modelos de abordagem *black-box*. Normalmente, variáveis intermediárias em modelos híbridos possuem significado físico que pode facilitar a explicação das predições.

● Armazenamento e representação de dados:

SCHWEIDTMANN et al.(2021) explica que o surto do uso de ML em aplicações de mídias sociais, compras online e serviços de streaming dependem muito da vasta quantidade de dados estruturados disponíveis para uso. No entanto, em campos como a engenharia química, uma fração muito pequena de conhecimento e informações é acessível para o uso em algoritmos de ML, sendo a maioria disponibilizada em formatos não digitalizados. A maioria dos dados de engenharia química provém da computação de dados coletados por sensores e medidores. No entanto, dados moleculares, fluxogramas, publicações científicas, experimentos laboratoriais, dentre outros, não estão prontamente disponíveis para uso no ML.

Portanto, a estruturação e disponibilidade de dados para a engenharia química ainda é um grande obstáculo para o avanço na área.

- Heterogeneidade dos dados:

A heterogeneidade dos dados na engenharia química possui muitas fontes (dados de laboratório, diferentes medidas, fontes de publicação, arquivos de simulação) e o processamento dessas diversas fontes acaba por se tornar um empecilho (SCHWEIDTMANN et al. , 2021). Para descrever dados heterogêneos, representações especializadas de engenharia química são necessárias. Por muito tempo, pesquisadores têm desenhado manuais para descrever esses dados. Moléculas, por exemplo, podem ser descritas por contagens moleculares ou métodos de contribuição de grupos. Porém, esses manuais normalmente requerem conhecimentos específicos e podem levar a modelos enviesados (SCHWEIDTMANN et al. , 2021).

SCHWEIDTMANN et al.(2021) ainda cita que, recentemente, moléculas e cristais têm sido representados na forma gráfica e processados por algoritmos de ML especializados. Esses modelos conseguem diretamente operar na estrutura gráfica e têm mostrado resultados promissores na predição de relações de propriedades estruturais. Dessa forma, o algoritmo consegue interpretar os dados gráficos e aprender sobre as representações moleculares e mapear as propriedades físico-químicas inerentes. Espera-se que, futuramente, dados de diferentes tamanhos e fontes possam ser integrados para serem utilizados em estudos que se usem do ML, juntando, por exemplo, dados experimentais e de simulações.

Por fim, THEBELT et al. (2022) destaca também a questão do impacto da alta variância dos dados e o baixo volume com que chegam para fins de aplicação em ML na engenharia química. Como será explicitado na seção 2.2, a estatística do ML lida com o balanço viés-variância, sendo que o ganho em um acarreta a perda no outro, sendo um bom modelo estatístico aquele que consiga um equilíbrio mútuo. Devido à pobreza de fontes de dados estruturados e somado à alta variabilidade nos dados, tem-se um desafio a ser superado para as aplicações em engenharia. Um exemplo ilustrado são as produções em batelada ou semi-batelada, que normalmente são empregadas quando a demanda é baixa ou o produto resultante é caro. Enquanto indústrias de larga escala possuem alto nível de monitoramento, os dados para processos em batelada são limitados. Sem as limitações do regime estacionário, os procedimentos em em batelada permitem um tratamento dinâmico, ou seja, diferentes perfis de temperatura, para maximizar variáveis chave como taxa de conversão e seletividade.

Por mais que essa dinamicidade permita uma melhor performance comparado aos processos estacionários contínuos, a modelagem e monitoramento do processo se torna mais complexo, normalmente levando ao enfraquecimento de modelos físicos. Nesse sentido, modelos a base de dados podem se tornar mais interessantes para se controlar processos dinâmicos, obter otimizações e poder prever a qualidade do produto final (THEBELT et al.,2022).

2.2 CONCEITOS DE APRENDIZADO ESTATÍSTICO

Com o intuito de se dar embasamento para os estudos de caso citados no presente trabalho, realizou-se uma revisão teórica dos principais conceitos e algoritmos envolvendo o ML, descrito por JAMES et al. (2013) também como aprendizado estatístico.

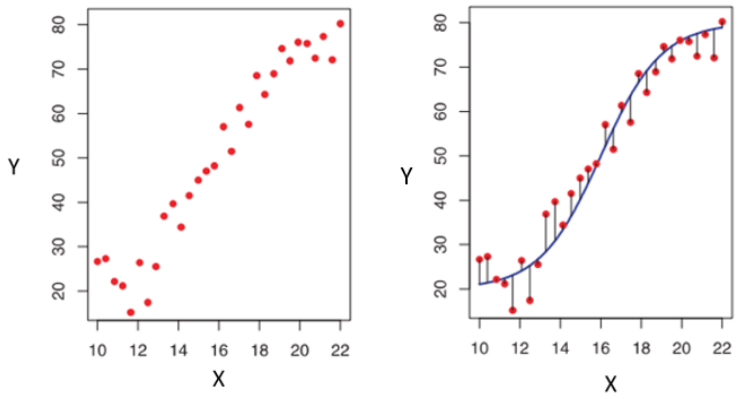
Estimando uma função *f*

O aprendizado estatístico consiste em se desenvolver modelos capazes de realizar tarefas de previsão de padrões. JAMES et al. (2013) exemplifica o conceito denominando variáveis de saída ou dependentes como Y, enquanto as variáveis de entrada são chamadas independentes e denotadas por X. De forma geral, ambas as variáveis se relacionam da seguinte maneira:

$$Y=f(X)+\varepsilon \tag{2.1}$$

A função *f* é uma função fixa, porém desconhecida, que correlaciona Y e X, enquanto ε representa o termo de erro, que é independente de X e possui média zero como mostrado na figura 2.3. Pode-se observar que alguns erros são positivos e outros negativos com relação a *f*, porém, na média, o valor final se aproxima de zero. Em essência, os métodos de aprendizado estatístico referem-se a uma série de métodos com o intuito de se prever a forma da função *f* (JAMES et al.,2013).

Figura 2.3. Base de observações e seus respectivos erros



Fonte: Adaptado de JAMES et al. (2013)

Modelos paramétrico e não-paramétricos

Segundo JAMES et al. (2013), uma dada função pode ser desenhada através de modelos lineares ou não-lineares. Para se chegar em uma estimativa, os métodos de aprendizado estatístico se utilizam dos chamados dados de treinamento para "ensinar" um algoritmo. A maioria dos métodos de aprendizagem se dividem em em duas categorias:

●Métodos paramétricos: Consistem em se assumir, previamente, a natureza da função (seu formato) para então se estimar seus coeficientes. Assumindo-se, por exemplo, que a função seja linear, tem-se o seguinte formato:

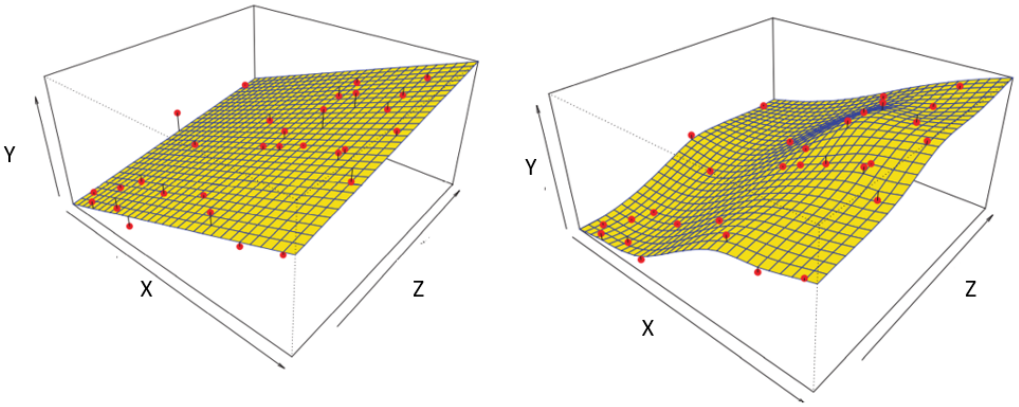
$$f(X) = \beta_{0+}\beta_1 X_1 +_{+}\beta_2 X_2 + ... + \beta_p X_p \qquad (2.2)$$

O trabalho se resumiria, portanto, em se utilizar dados de treino para se estimar os coeficientes B, utilizando-se de técnicas como o Ajuste dos Mínimos Quadrados que minimizam a distância dos pontos para se obter uma reta (JAMES et al.,2013). Nos métodos paramétricos, não há necessidade de se estimar *f* e sim seus parâmetros, o que o torna vantajoso nesse aspecto. Porém, na maioria dos casos, não se tem um conhecimento prévio da natureza de *f*, podendo-se optar por modelos muito diferentes da forma real e que logo, forneceriam uma estimativa pobre empobrecida. Uma alternativa diz respeito a escolha de modelos mais flexíveis que possam se encaixar em uma variedade maior de formatos. O problema de modelos mais flexíveis é que esses dependem de uma estimativa mais complexa

de parâmetros e podem levar ao chamado *overfitting*, ou seja, a curva segue precisamente os padrões dos dados de treino, incluindo erros e gerando os chamados ruídos (JAMES et al.,2013). Tem-se duas abordagens pela qual um determinado modelo estatístico pode ser estudado:

●Métodos não-paramétricos: nessa abordagem, não se assume previamente o formato de uma função, mas estima-se o formato de *f* com base na distribuição dos próprios dados, se aproximando a curva dos pontos. Dessa forma, uma abordagem não-paramétrica pode criar modelos que se encaixam em uma variedade maior de curvas, porém, necessitam também uma base de observações maior com relação aos métodos paramétricos para se obter uma função de maior acurácia (JAMES et al.,2013). Um exemplo de abordagem paramétrica e não paramétrica é mostrado na figura 2.4.

Figura 2.4. Modelo paramétrico ajustado versus modelo não-paramétrico com *overfitting*.



Fonte: Adaptado de JAMES et al. (2013)

Para métodos não-paramétricos, um nível de suavidade deve ser escolhido. Caso a abordagem escolhida seja muito flexível , surgem os problemas de *overfitting* mencionados, no qual a curva se ajusta muito bem aos dados de treino, porém perde sua capacidade de extrapolação para outras observações durante a fase de teste (JAMES et al.,2013).

Acurácia de um modelo

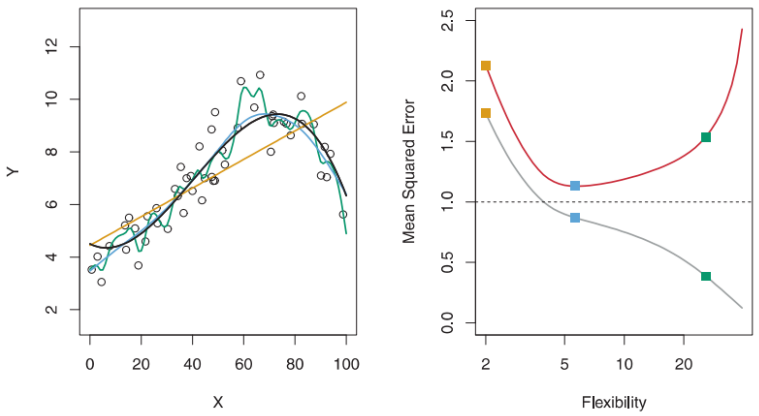
Para se avaliar a performance de um determinado modelo, deve-se determinar uma forma de medição de quão bem as previsões deste se encaixam nos dados observados. Segundo JAMES et al. (2013), para problemas de regressão, a medida mais comumente utilizada é o erro médio quadrático ou MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2.3}$$

Tem-se, pela equação, uma avaliação do erro ocasionado pela diferença entre a predição do modelo para uma dada observação dada por *f(xi)* e o resultado de saída esperado correspondente a essa observação dada por *yi*. Portanto, o MSE será menor se as as saídas do modelo construído forem o mais próximas possível do valor esperado (JAMES et al.,2013).

Deve-se ressaltar que, no aprendizado estatístico, o interesse maior advém da minimização dos erros relacionados aos dados de teste e não de treino. Os dados de treino são utilizados para se ajustar o modelo tendo em mente uma resposta já conhecida, porém, na prática, o interesse é que esse seja aplicado em uma série de observações desconhecidas (dados de teste) e avaliar se o erro para esse também é o mínimo possível. Em outras palavras, o objetivo do experimento é minimizar simultaneamente o MSE tanto para os dados de treino quanto os de teste (JAMES et al.,2013).

Figura 2.5. Comparativo entre modelos hipotéticos, a função real e o MSE de treino e teste



Fonte: Adaptado de JAMES et al. (2013)

A figura 2.5 da esquerda mostra três modelos de função (um modelo paramétrico linear representado em amarelo e dois não-paramétricos representados em verde e azul) construídos para se determinar a forma real da função *f* (curva preta). À direita, a curva vermelha indica o MSE para os dados de treino e a cinza, para os dados de teste. Os quadrados, portanto, indicam o MSE de treino e teste para cada um dos modelos da esquerda.

Uma característica essencial do aprendizado estatístico, independente da abordagem adotada é que, à medida que se aumenta a flexibilidade de um modelo, o MSE para as observações de treino diminuem, porém o mesmo não ocorre necessariamente com as observações de teste como exemplificado (JAMES et al.,2013). Isso ocorre pois o processo de aprendizado do algoritmo está interpretando possíveis erros aleatórios como parte do modelo, ocasionando um processo de overfitting, ou seja, uma adequação muito alta aos dados e não a curva característica propriamente. Para amenizar essas dificuldades, a escolha de métodos apropriados de amostragem são empregados, sendo estes discutidos mais à frente.

Equilíbrio viés-variância

JAMES et al. (2013) mostra em seu livro que, no aprendizado estatístico, duas propriedades competem entre si e são essenciais para a escolha de um modelo apropriado: o viés e a variância.

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon) \tag{2.4}$$

Na equação 2.4, o primeiro termo denota o MSE de teste esperado caso se realizasse múltiplas estimativas para se chegar a um f utilizando-se de uma amostragem expressiva de dados de teste. A equação mostra a partir de seu segundo e terceiro termo que, para se minimizar o erro médio do MSE de teste, o método deve assumir, simultaneamente, uma baixa variância e um baixo viés. A variância se refere ao quanto a função f poderia variar se houvesse uma pequena mudança nos valores dos dados de entrada durante o treino (algo que costuma ocorrer quando se adota modelos mais flexíveis), enquanto o viés refere-se ao erro que é introduzido ao se assumir um modelo muito simplificado (o que normalmente ocorre para modelos mais rígidos como o linear). Como regra geral, o uso de modelos mais flexíveis geram uma maior variância com um menor viés (JAMES et al.,2013).

Aprendizado supervisionado e não-supervisionado

Segundo JAMES et al. (2013), quando se tem um valor esperado de saída, ou seja, para cada observação *x1, x2* e sua respectiva resposta *f(x1), f(x2)* há um valor de comparação *y1, y2*, tem-se em mãos um problema de aprendizagem supervisionada. Os problemas de aprendizagem supervisionada são o enfoque dos métodos aqui apresentados e suas aplicações. Problemas de aprendizagem não-supervisionados, por outro lado, não necessitam um vetor *Y* de saída e, portanto, não se utilizam de uma variável resposta supervisionando o problema. Busca-se, normalmente, para esse tipo de análise, compreender o nível de similaridade entre as informações observadas, sendo portanto, mais aplicável a problemas de clusterização ou agrupamento.

Problemas de classificação e regressão

Pode-se definir, de forma geral, que problemas envolvendo variáveis de saída numéricas são considerados problemas de regressão, enquanto problemas envolvendo variáveis de saída qualitativas ou categóricas são considerados problemas de classificação (JAMES et al.,2013). A distinção entre ambos nem sempre é tão clara, sendo métodos como regressão linear por mínimos quadrados usados para respostas quantitativas, enquanto a regressão logística costuma ser utilizada para problemas de classificação binária. Porém, essa última também pode ser pensada como uma estimativa da probabilidade de classificação, e logo, poderia ser pensada como um problema de regressão (JAMES et al.,2013). Costuma ser de praxe no aprendizado estatístico a codificação das categorias em números (1 para sim e 0 para não, por exemplo) o que torna ambos os métodos aplicáveis para a maioria dos casos.

2.3 MÉTODOS DE AMOSTRAGEM

A escolha de uma metodologia apropriada de amostragem é um dos pontos essenciais no aprendizado estatístico. De forma geral, tais métodos envolvem distribuir e redistribuir amostras de um conjunto de dados de teste e ajustar a função em cada batelada da amostra para se obter o máximo de informação sobre a curva procurada.

Aqui, JAMES et al. (2013) apresenta duas técnicas muito utilizadas. A primeira é a validação cruzada, que pode ser utilizada para se estimar o erro do teste associado com um determinado método estatístico e assim avaliar sua performance ou selecionar um nível de

flexibilidade adequado. A segunda, o *bootstrap*, é usado em vários contextos, mais comumente para se obter uma medida da acurácia de um parâmetro ou também se estimar um método estatístico

Validação cruzada

Como mencionado nos conceitos principais de aprendizagem estatística, o erro associado aos dados de treino comumente são diferentes dos erros associados aos dados de teste e, particularmente, o primeiro pode superestimar o segundo (JAMES et al.,2013).

A técnica de validação cruzada ou CV apresentada por JAMES et al.(2013) possui um princípio simples. Suponha-se que seja preciso estimar o erro de teste associado com uma determinada função a ser modelada e seus dados de observação. O método consiste em dividir aleatoriamente as amostras de observações em duas partes: o conjunto de treino e o conjunto de validação. O modelo é ajustado ao conjunto de treino e seu resultado estimado é utilizado para se predizer as respostas para as observações no conjunto de validação. O resultado associado ao erro da validação provê uma estimativa do erro de teste .

Figura 2.6. Esquema de uma versão simplificada da validação cruzada

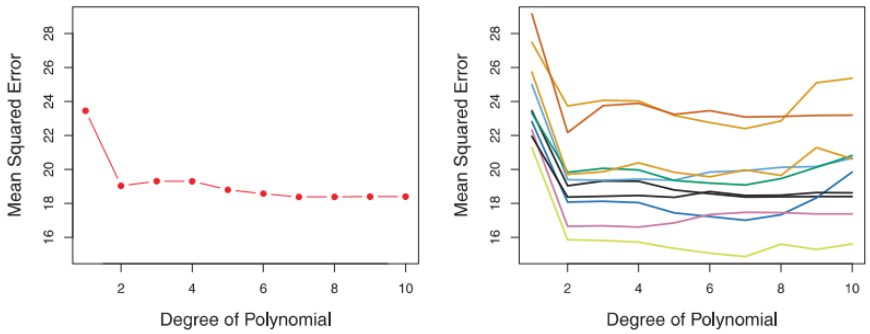
Fonte: Adaptado de JAMES et al. (2013)

Como observado pela figura, um conjunto de *n* observações é aleatoriamente dividido em um conjunto de treino e de validação. A função estatística procurada é ajustada ao conjunto de treino e, posteriormente, sua performance é avaliada no conjunto de validação.

Entretanto, essa abordagem mais simplista de se dividir uma amostragem possui limitações. Nesse caso, utiliza-se apenas um subgrupo de observações (o conjunto de treino) para se estimar o modelo estatístico. Como os modelos tendem a uma performance mais pobre quando treinados com poucas observações, isso sugere que essa abordagem de divisão

binária dos dados pode superestimar o erro de teste para o conjunto completo de dados (JAMES et al.,2013).

Figura 2.7. Validação cruzada simples com divisão binária feita uma vez e repetida dez vezes

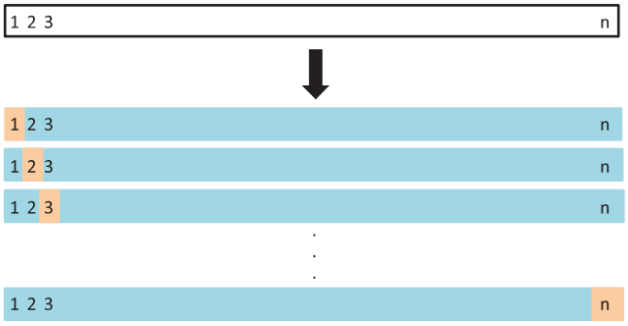


Fonte: Adaptado de JAMES et al. (2013)

A figura 2.7 mostra o erro do conjunto de validação para uma validação cruzada com divisão binária feita uma única vez e repetida 10 vezes, cada vez para uma divisão aleatória em um conjunto de treino e um conjunto de validação. Pode-se observar que para todos os erros desenhados, a variação mostrou-se elevada e portanto, não há um consenso dentre as curvas sobre qual modelo resulta no menor erro associado ao conjunto de validação. A única inferência que se pode fazer é que, ao se aumentar o grau da função para 2, o erro médio associado ao conjunto de validação da função diminui, porém o mesmo não ocorre para os demais graus do polinômio. Isso mostra que, dependendo de qual observação adentrar em um conjunto de treino ou de validação durante a partição binária, o erro associado a performance da função quando ajustada aos dados de validação possui alta variabilidade.

Para contornar tais erros, uma alternativa segundo JAMES et al. (2013) é o uso da validação cruzada *leave-one-out* ou um dado de fora (LOOCV). Na figura 2.8, é mostrado um conjunto de n dados com divisão repetitiva entre um conjunto de treino (em azul) e um de validação (em laranja), assim como na validação cruzada simples, porém, para cada divisão do conjunto de treino, é retirada uma única observação que é alocada no conjunto de validação (este contendo então, uma única amostra). Isso permite que a amostragem possa ser repetida n vezes.

Figura 2.8. Esquema de validação cruzada LOOCV



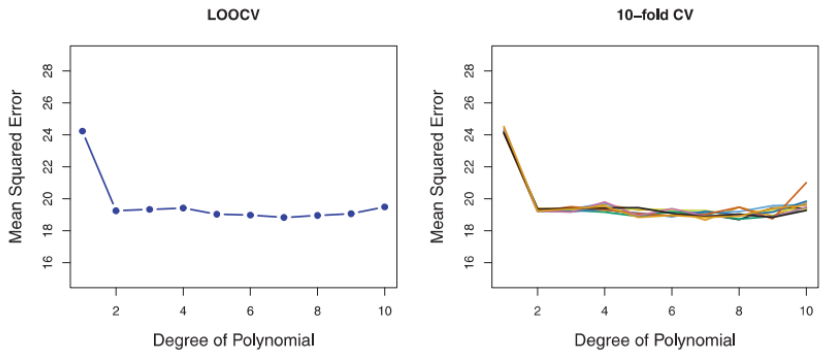
Fonte: Adaptado de JAMES et al. (2013)

A fórmula para LOOCV utiliza a média do MSE para os *n* testes realizados:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i \tag{2.5}$$

O método LOOCV possui algumas vantagens como ter menos viés e, em contraste com a validação cruzada simples, que gera resultados sempre diferentes devido a aleatoriedade na divisão do conjunto de treino/validação, ao se performar o LOOCV múltiplas vezes o resultado será equivalente (JAMES et al.,2013), como mostrado na figura 2.9 abaixo.

Figura 2.9. Curva de erro do LOOCV para 1 divisão e para 9 divisões.



Fonte: Adaptado de JAMES et al. (2013)

JAMES et al. (2013) cita uma alternativa ao LOOCV que é a validação cruzada K-fold ou k-fold CV. Essa abordagem consiste em dividir aleatoriamente os dados

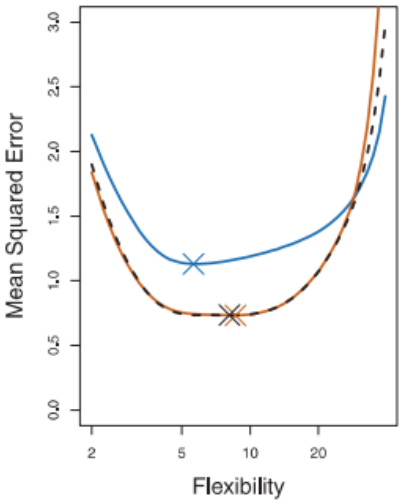
observados em *k* grupos de aproximadamente o mesmo tamanho. O primeiro grupo é tratado como de validação e os *k-1* restantes são ajustados no método estatístico estudado. O MSE1 é então computado para as observações para o grupo de validação e o procedimento é repetido *k* vezes, sendo a cada vez, um grupo diferente é tratado como sendo de validação. O *k*-fold CV é obtido através da média para os *k* MSEs calculados como segue:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i \tag{2.6}$$

O LOOCV é, na realidade, um caso especial do k-fold CV com $k = n$, sendo que a vantagem principal é o menor número de amostragens necessárias e, portanto, menor esforço computacional (JAMES et al.,2013).

Ao se examinar dados reais, não é possível saber o verdadeiro erro MSE de teste e, logo, torna-se difícil mensurar a acuracidade da validação-cruzada estimada. Quando se realiza uma validação-cruzada, o objetivo pode ser determinar quão bem um modelo estatístico irá performar em dados independentes. Nesse caso, a estimativa real do MSE de teste é interessante, porém, em outras situações, o interesse maior se dá apenas na localização do ponto de mínimo da curva do MSE teste (JAMES et al.,2013). Como mostrado na figura 2.10, embora às vezes a validação-cruzada gere resultados subestimados da curva real, o método consegue mostrar qual o melhor nível de flexibilidade para o modelo estatístico em estudo, ou seja, com o menor MSE de teste. Na figura 2.10, pode-se comparar o erro de teste MSE associado ao LOOCV (azul), ao *k* - fold CV (laranja) e o verdadeiro MSE (tracejado preto).

Figura 2.10. Erro MSE real e de teste para dados simulados



Fonte: Adaptado de JAMES et al. (2013)

Bootstrap

Segundo JAMES et al. (2013), esse método é considerado uma ferramenta de amostragem estatística poderosa que pode ser utilizada para se quantificar a incerteza associada a um dado modelo estatístico. O *bootstrap* é usado, por exemplo, para se estimar o erro padrão associado aos coeficientes de uma regressão linear.

Suponha que se deseje estimar o erro associado a um suposto coeficiente α estipulado para uma função $f(x) = y$. Para se realizar essa tarefa, poderia-se calcular α para vários pares x,y e então, realizar uma média de todos os valores obtidos para esse parâmetro e se realizar uma simples média aritmética:

$$\bar{\alpha} = \frac{1}{1,000} \sum_{r=1}^{1,000} \hat{\alpha}_r \tag{2.7}$$

A partir do resultado o desvio padrão também poderia ser calculado:

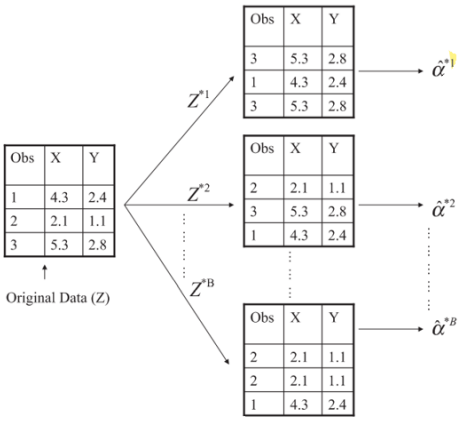
$$\sqrt{\frac{1}{1,000-1} \sum_{r=1}^{1,000} (\hat{\alpha}_r - \bar{\alpha})^2} \tag{2.8}$$

O valor de encontrado significaria que, caso selecionada uma amostra aleatória da população, seria esperada uma diferença entre o α médio e o α dessa amostragem. Na prática,

porém, essa abordagem não é utilizada para dados reais, pois não se pode gerar amostras novas da população original (JAMES et al.,2013). O *bootstrap* serve, portanto, como uma alternativa para se gerar "novos" conjuntos de dados a partir de uma população pré-definida ao invés de ter que se obter novos dados independentes.

JAMES et al. (2013) demonstra essa abordagem com um conjunto de dados denominado *Z*, contendo, por exemplo, $n = 3$ observações. São selecionados *n* dados para se montar um conjunto bootstrap Z^{*1}. A amostragem é sempre feita com reposição, ou seja, a mesma observação pode ocorrer mais de uma vez na amostragem bootstrap (na figura, pode-se observar que Z^{*1} contém a observação 3 se repetindo duas vezes, a observação 1 uma vez e a observação 2 nenhuma vez). Pode-se utilizar Z^{*1} para então e estimar um α^{*1}. Esse procedimento é repetido *B* vezes para um valor elevado de *B* para se produzir *B* amostragens bootstraps como mostrado na figura 2.11.

Figura 2.11. Esquema de amostragem bootstrap



Fonte: Adaptado de JAMES et al. (2013)

Matematicamente, o erro padrão do procedimento bootstrap pode ser escrito como:

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}. \qquad (2.9)$$

2.4 REDES NEURAIS ARTIFICIAIS

O que são as Redes Neurais Artificiais?

Os estudos sobre Redes Neurais Artificiais (ANNs) são um campo que muito se desenvolveu desde 1980 e, desde então, tem tido aplicações nas mais diversas áreas (DING et al., 2011). As ANNs constituem um subconjunto do ML e estão no núcleo do desenvolvimento dos algoritmos de deep learning. Seu funcionamento é inspirado pelo próprio cérebro humano e na forma como os neurônios biológicos transmitem informações por meio das conexões sinápticas(IBM CLOUD EDUCATION, 2020).

Em essência, essas estruturas se constituem de nós (equivalentes aos neurônios) que recebem variáveis de entradas e então fornecem uma resposta de saída. Um conjunto de nós constitui uma camada, sendo as redes neurais compostas de várias delas, tendo normalmente uma camada de entrada, uma camada de saída e várias camadas ocultas entre ambas. A ideia principal por trás do funcionamento das redes neurais é que cada nó se conecta ao outro, possuindo um peso e limite associado. Quando a saída de um nó ultrapassa um limite especificado, o nó será ativado e enviará a informação para o nó subsequente, e assim por diante (IBM CLOUD EDUCATION, 2020).

As redes neurais podem ser entendidas como sistemas não lineares e adaptativos no processamento de informações, tendo como características a capacidade adaptação e organização . Para tal, elas contam com a necessidade de um conjunto de dados para treinamento para que refinem sua precisão (IBM CLOUD EDUCATION, 2020).

Um aspecto importante das redes neurais diz respeito a sua arquitetura, ou seja, a seleção apropriada de uma estrutura, dos parâmetros , das amostras de treinamento, dos valores de entrada iniciais, dos valores de convergência e do algoritmo de ativação, por exemplo (DING et al., 2011). Essa determinação inicial é essencial pois a performance de uma rede neural é bastante sensível a tais fatores. Uma rede neural com poucos neurônios pode resultar em problemas de baixa aproximação ou alta generalização, enquanto, por outro lado, muitos neurônios resultam em problemas de overfitting ou baixa generalização (DING et al., 2011). Um dos principais desafios do machine learning é justamente saber definir o balanço ideal entre o viés e a variância de um modelo.

Ainda sobre a arquitetura, ALMEIDA (2016) destaca três tipos principais: Redes alimentadas adiante com uma única camada (feedforward), redes alimentadas adiante com múltiplas camadas (MLP) e redes recorrentes (RNN). Sobre o processo de aprendizado, as redes neurais podem ser treinadas de duas formas: aprendizagem supervisionada e não-supervisionada, tendo como diferença principal a presença de um supervisor externo no primeiro tipo (ALMEIDA, 2016). Essas técnicas, bem como os tipos de arquitetura, serão abordados mais adiante.

Funcionamento das Redes Neurais Artificiais (ANNs)

As redes neurais são compostas por nós. Para cada nó, é definida uma quantidade de entradas, pesos sinápticos (*wi*) que indicam a relevância de cada variável de entrada (*xi*) para determinado processo e um viés, que tem a função de ponderar a entrada da função de ativação (IBM CLOUD EDUCATION, 2020). A equação 2.10 também pode ser chamada de termo integrador:

$$\sum_{i=1}^{m} wixi + bias = w_1x_1 + w_2x_2 + w_3x_3 + \text{viés} \quad (2.10)$$

Uma vez determinada a camada de entrada, com suas variáveis e pesos, todas as entradas são multiplicadas e somadas. A saída é emitida por meio de uma função de ativação. Se essa saída excede um determinado valor, ela irá ativar o próximo nó, transmitindo a informação para a próxima camada da rede. Esse processo de transmissão de dados de uma camada anterior para outra posterior é chamado rede feedforward (IBM CLOUD EDUCATION, 2020).

Existem vários tipos de funções de ativação que podem ser empregadas. Sua forma geral pode ser representada pela expressão abaixo, no qual x representa a entrada de um neurônio, w representa o peso sináptico associado, b representa o viés e a somatória desses termos (integrador) corresponde a entrada da função de ativação AF, que por fim nos retorna uma resposta de saída y (ALMEIDA, 2016).

$$yi = AF(\sum_{i=1}^{n} (xiwi) + viés) \quad (2.11)$$

As funções de ativação mais comumente utilizadas, como levantado por ALMEIDA (2016), são as seguintes:

●Função limiar: É utilizada quando se quer restringir a saída em valores binário como 0 e 1. No caso, a função assume valor 1 quando o integrador é positivo ou nulo e valor 0 quando o integrador é negativo.

$$f(u) = \begin{cases} 1 & se \quad u \geq 0 \\ 0 & se \quad u < 0 \end{cases} \tag{2.12}$$

●Função linear por partes: Segue uma lógica parecida com a função limiar, porém com mais opções de saída. Quanto mais "degraus" a função linear por partes assume, pode-se dizer que a função mais se aproxima de modelos não-lineares.

$$f(u) = \begin{cases} 1 & se \quad u \geq +1/2 \\ u & se +1/2 > u > -1/2 \\ 0 & se \quad u \leq -1/2 \end{cases} \tag{2.13}$$

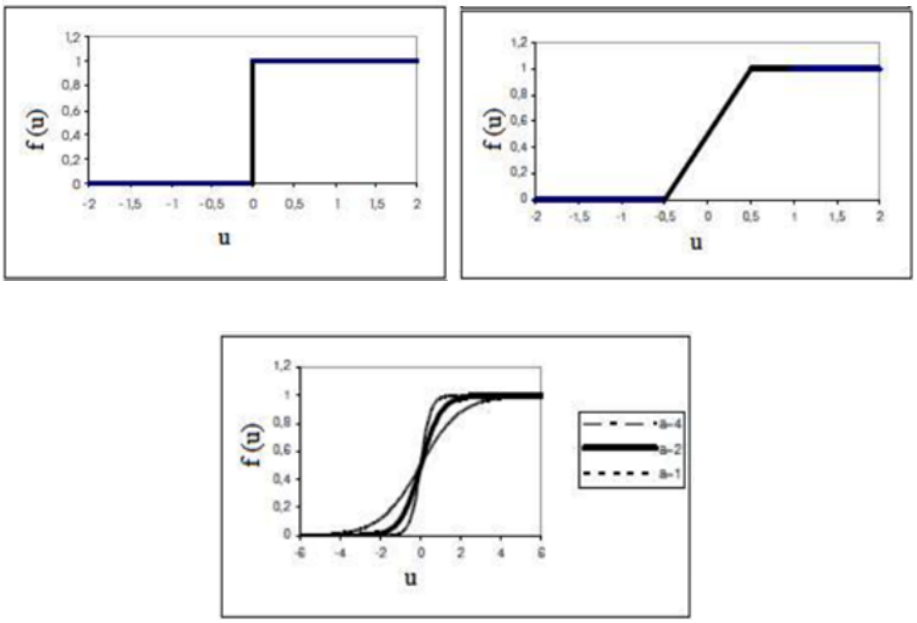
●Função sigmoidal: Trata-se da função mais comum devido ao seu balanceamento adequado entre o comportamento linear e não-linear, além de assumir valores no intervalo entre 0 e 1, sendo adequada quando se deseja restringir a saída a esse intervalo. Um exemplo da função sigmóide é a função logística representada abaixo. No caso abaixo, o parâmetro a determina a inclinação da função sigmóide (quanto mais elevado o valor de a, mais inclinada a curva se torna):

$$f(\mu) = \frac{1}{1 + \exp(-a\mu)} \tag{2.14}$$

Há vários motivos que tornam a função logística uma aplicação comum em redes neurais. Vários fenômenos naturais seguem o padrão de uma curva logística, exibindo uma progressão lenta no início e se tornando "acelerada" ao decorrer do processo, determinando seu típico formato em S. Quando não se tem um modelo específico para servir de base na modelagem de um problema para o machine learning, normalmente se adota a função logística (WIKIPEDIA, 2022). Além do mais, as curvas sigmóides são comuns em estatística tais como em distribuições cumulativas e a própria distribuição normal (WIKIPEDIA, 2022).

●Função tangente hiperbólica: Essa função caracteriza-se por ser outro exemplo de curva sigmóide. Segundo ALMEIDA (2016), é utilizada como alternativa da função logística quando não se quer restringir os valores de saída ao intervalo 0 e 1. A função tangente hiperbólica preserva o formato sigmoidal e pode assumir valores positivos e negativos.

Figura 2.12. Representação de função de ativação limiar, linear por partes e sigmóide com inclinação variável *a*.



Fonte: Adaptado de ALMEIDA (2016)

O aprendizado de uma rede neural é realizado através do treinamento, ou seja, uma série de processos iterativos aplicados aos pesos sinápticos. O aprendizado é atingido quando a rede neural atinge uma solução para determinado problema. Na prática, isso significa que a rede ajustou sua matriz de pesos sinápticos de forma que o conjunto de variáveis de saída coincida com um valor esperado para cada variável de entrada (ALMEIDA, 2016). Pode-se destacar dois tipos de aprendizados para as redes neurais:

● Aprendizado não-supervisionado: Não há a presença de um supervisor externo. Logo, a rede neural deve procurar algum tipo de correlação ou redundância nos dados de entrada.

● Aprendizado supervisionado: Um supervisor externo fornece à rede neural a saída esperada para um dado valor de entrada. Com isso, é possível comparar a saída dada pela rede com a saída esperada e obter um valor de erro. Dessa forma, os pesos sinápticos se ajustam de forma a minimizar esse erro.

Adentrando mais no aprendizado supervisionado, tem-se o conceito de erro destacado por ALMEIDA (2016), ou seja, a diferença entre a resposta esperada (*yi*) e a obtida pela rede f(*xi*), dada pela simples expressão:

$ei(n) = yi - f(xi)$ (2.15)

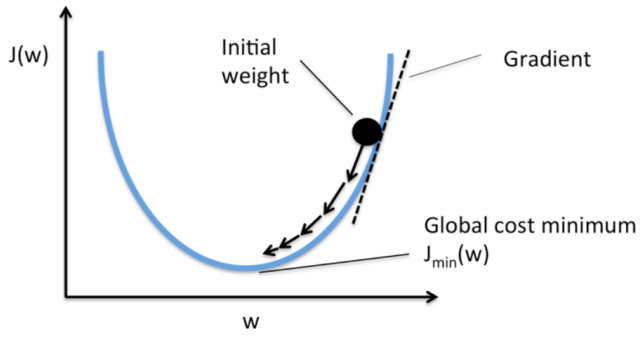
À medida que o modelo é treinado, é necessário avaliar sua precisão com base em algum cálculo que prescreve quando o erro é suficientemente pequeno para que os parâmetros (ou pesos) se tornem adequados ao processo estudado (IBM CLOUD EDUCATION,2020). Para tal, utiliza-se comumente a função de custo ou função de mínimo erro quadrado, dado pela seguinte expressão:

$$\text{Cost Function} = \text{MSE} = \frac{1}{2m} \sum_{i=1}^{m} (yi - y)^2 \quad (2.16)$$

No qual i representa o índice da amostra, *yi* o resultado previsto, *y* o resultado real e *m* o número de amostras.

A meta do treinamento da rede neural é, por fim, minimizar a função de custo para garantir a correção dos pesos com relação à observação fornecida (IBM CLOUD EDUCATION, 2020). A medida que o modelo ajusta seus pesos e vieses, ele utiliza a função de custo para atingir o mínimo local ou ponto de convergência. A função matemática mais utilizada para esse caso é a Descida do Gradiente (IBM CLOUD EDUCATION, 2020). Essa função consiste em um algoritmo de otimização usado para encontrar valores de parâmetros (coeficientes de peso e vieses) que minimizam a função de custo como mostrado na figura 2.13 (DEEP LEARNING, 2022).

Figura 2.13. Esquema de descida do gradiente



Fonte: Adaptado de DEEP LEARNING (2022)

O processo de aprendizado da rede neural em um ambiente supervisionado consiste, portanto, em um procedimento padrão. Os dados de entrada são divididos em amostras e cada amostra passa pelas entradas da rede, multiplicados pelos pesos, soma-se tudo com o viés e no final tem-se os dados de saída, que é a previsão da rede (DEEP LEARNING, 2022). Esses então são comparados com a resposta esperada, o erro é calculado e então retroagido ajustando os pesos de cada neurônio da camada. Quando os pesos acabam de ser atualizados, uma nova amostra é introduzida e multiplicada pelos pesos ajustados. E assim por diante, esse processo de atualização dos pesos até que a função de custo atinja seu mínimo é chamado aprendizado (DEEP LEARNING, 2022).

A partir do conhecimento do funcionamento de redes neurais, algumas vantagens podem ser citadas sobre o uso dessa metodologia. ALMEIDA (2016) destaca:

●Não-linearidade: Um neurônio pode ou não ser linear a depender da função de ativação empregada. Caso seja não-linear, essa característica é distribuída por toda rede e a não-linearidade torna-se uma propriedade importante do método.

●Adaptabilidade: As redes neurais possuem a capacidade de adaptar seus pesos sinápticos a modificações do ambiente. Logo, uma rede treinada para operar em determinado ambiente pode ser reorientada para lidar com pequenas alterações.

●Informação contextual: Cada neurônio da rede é potencialmente afetado pelas atividades de todos os demais, portanto, a informação da rede é sempre tratada dentro de um contexto geral e não individual.

●Resposta a evidência: Quando utilizada para classificação de padrões, a rede neural pode ser projetada para mostrar informações sobre a confiança do resultado. Logo, padrões ambíguos podem ser rejeitados, melhorando o desempenho da classificação.

As redes neurais, assim como todos os métodos estatísticos, também apresentam limitações e desvantagens. ALMEIDA (2016) destaca como desvantagens:

●Treinamento demorado: Dependendo da rede neural projetada, o treinamento pode levar horas ou mesmo dias.

●Caixa-preta: Por vezes não é possível saber o motivo que levou a rede a uma determinada conclusão.

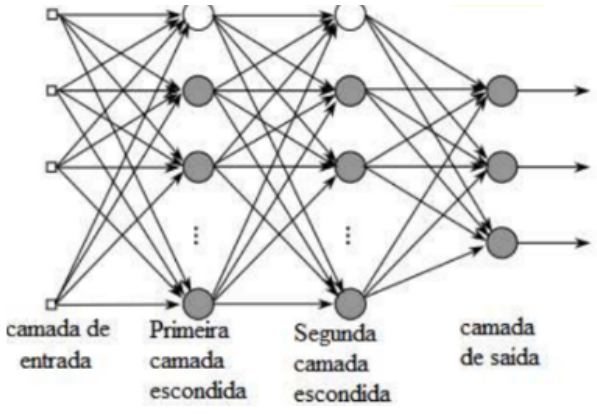
●Grande volume de dados: para o correto treinamento da rede, é necessário por vezes um grande volume de dados históricos.

●Resultados falhos: as redes podem chegar a resultados contraditórios às teorias aceitas, cabendo ao engenheiro possuir o bom senso de mensurá-las.

Perceptrons de Múltiplas Camadas (MLPs)

O percéptron consiste em uma forma simples de rede neural utilizada principalmente na classificação de padrões. As arquiteturas do tipo múltiplas camadas constituem a forma mais comum de rede neural empregada atualmente (ALMEIDA, 2016). Uma MLP consiste em uma camada de entrada, camadas intermediárias e uma camada de saída. A camada de entrada distribuí as informações para as camadas intermediárias ocultas na rede e na camada de saída, a solução do problema é obtida. Numa arquitetura MLP, os neurônios de uma camada estão conectados apenas aos neurônios da camada imediatamente posterior, não havendo realimentação nem conexão entre neurônios da mesma camada (ALMEIDA, 2016). Os sinais são propagados da esquerda para a direita como na figura 2.14, sendo essa arquitetura um tipo *feedforward* ou de alimentação para frente.Essas características estão presentes em arquiteturas neurais recorrentes, que serão explicadas mais à frente.

Figura 2.14. Esquema de MPL com duas camadas intermediárias



Adaptado de ALMEIDA (2016)

Na figura 2.14, é possível identificar uma rede do tipo MLP com duas camadas intermediárias. Na rede apresentada, observa-se que cada neurônio da rede está conectado a todas as saídas da camada anterior e o fluxo de sinais se dá da esquerda para a direita.

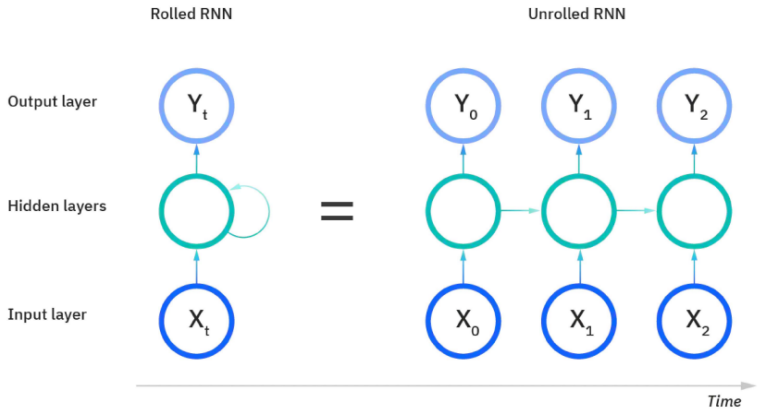
Redes Neurais Recorrentes (RNNs)

O último tipo de arquitetura aqui apresentado são as redes neurais recorrentes. O que difere essa estrutura das redes neurais apresentadas anteriormente diz respeito ao tipo de dado, pois as RNNs utilizam dados sequenciais ou séries temporais (IBM CLOUD EDUCATION, 2020). Por esse motivo, as RNNs costumam ser utilizadas em problemas envolvendo processamento de linguagem natural, dado que em sentenças gramaticais, a ordem das palavras importa para que haja atribuição correta de significado.

Da mesma forma que as redes neurais *feedforward*, as redes neurais recorrentes utilizam-se de dados de treinamento para o processo de aprendizado, porém, possuem uma "memória" no qual as informações das variáveis de entrada influenciam as variáveis de saída e de reentrada. Enquanto nas redes feedforward as entradas e saídas são assumidas como sendo independentes, nas RNNs a saída depende do elemento anterior na sequência de dados (IBM CLOUD EDUCATION, 2020).

A figura 2.15 representa um esquema de rede neural recorrente simplificada e explícita. Pode-se observar que, diferentemente das redes feedforward, os neurônios de uma mesma camada são conectados entre si, definindo assim uma "memória" ao neurônio que utiliza a informação anterior para atribuir uma saída adequada. Um exemplo de dados sequenciais poderia ser dado em uma frase como, por exemplo, "Olá, bom dia!". Nesse caso, cada saída *y0*, *y1* e *y2* representam, respectivamente, cada palavra a sentença na ordem estabelecida.

Figura 2.15. Esquema de RNN



Fonte: Adaptado de IBM CLOUD EDUCATION (2020)

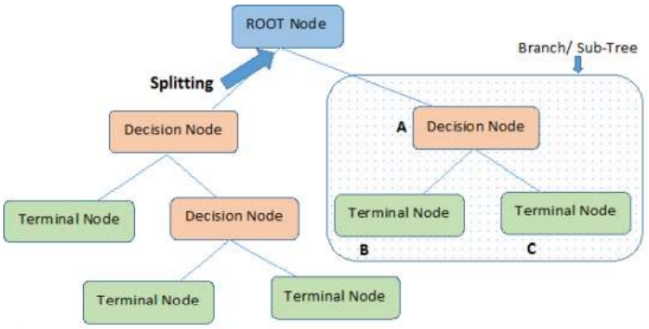
Uma outra diferença das redes neurais recorrentes diz respeito a sua forma de aprendizagem. Da mesma forma que as redes feedforward, elas se utilizam de dados de treinamento e calculam erros das variáveis de entrada e de saída, ajustando os parâmetros a cada iteração do processo. Porém, para as RNNs, os erros se somam a cada etapa do ajuste, enquanto que nas redes feedforward não necessitam somar os erros, dado que os neurônios não compartilham parâmetros em cada camada. Enquanto as redes feedforward possuem diferentes pesos associados a cada neurônio, nas RNNs esses pesos são partilhados entre os neurônios e possuem, portanto, os mesmos valores (IBM CLOUD EDUCATION, 2020).

2.5 MÉTODOS DE ÁRVORE

O que são árvores de decisão?

As árvores de decisão são um popular algoritmo de machine learning que se destaca por sua facilidade visual quando comparado a outros métodos. As etapas desse algoritmo muito lembram um fluxograma e, portanto, possuem certa facilidade de entendimento. Uma árvore de decisão como mostrada na figura 2.16 é um algoritmo de ML de aprendizagem supervisionada que pode ser utilizado tanto para fins de classificação (prever o resultado de variáveis discretas) quanto de regressão (prever o resultado de variáveis numéricas) (SACRAMENTO, 2021).

Figura 2.16. Estrutura de uma árvore de decisão



Fonte: Adaptado de MARTINS (2016)

Uma típica árvore de decisão possui em sua estrutura nós que determinam um ponto de tomada de decisão baseado na lógica "se-então", uma raiz que representa o nó de origem da árvore (ponto de input das variáveis de entrada) e suas folhas que representam as variáveis de saída ou resposta (SACRAMENTO, 2021). Ao chegar em um determinado nó, o algoritmo se pergunta acerca de uma regra de condição qual caminho deve seguir. Se a pergunta for por exemplo "O valor de dada variável é maior ou igual a X?" Se a resposta for "sim", a árvore segue para um lado do ramo, caso contrário, irá para o outro, e assim sucessivamente.

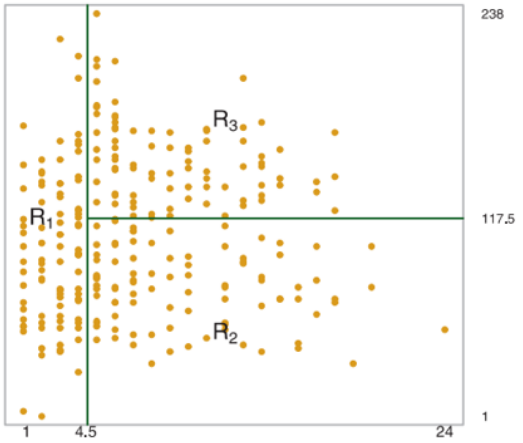
SACRAMENTO (2021) afirma que o algoritmo de árvore é chamado recursivo pois ele repete o mesmo padrão na medida em que adentra os níveis com maior profundidade. Outra característica que pode resultar em problemas de overfitting diz respeito a sua característica "gananciosa", ou seja, a tomada de decisão foca mais na tarefa que está sendo realizada do que no resultado final como um todo. Esse problema e suas formas de contorná-lo serão abordados mais à frente.

Árvores de regressão e árvores de classificação

As árvores de decisão para classificação e regressão possuem algumas diferenças. A primeira particularidade de cada uma diz respeito a sua finalidade. Enquanto árvores de regressão são utilizadas quando a variável de saída é numérica, as árvores de classificação são utilizadas quando a variável de saída é categórica (MARTINS, 2016). No caso das árvores de regressão, o valor obtido pelos nós terminais dos dados de treinamento é o valor

médio das observações, enquanto para as árvores de classificação, esse valor é a moda das observações (MARTINS, 2016). Ambas as árvores dividem o chamado espaço preditor em regiões distintas e não sobrepostas. Essas regiões podem ser pensadas como caixinhas com as devidas observações (MARTINS, 2016).

Figura 2.17. Esquema de um espaço preditor



Fonte: Adaptado de JAMES et al.(2013)

No caso do espaço preditor da figura 2.17 , as regiões R1, R2 e R3 são equivalentes aos nós terminais ou folhas, as linhas dividindo o espaço seriam os nós internos ou pontos de tomada de decisão do algoritmo (por exemplo, maior que 4.5 ou menor que 117.5) e os pontos seriam as observações (JAMES et. al., 2013).

Ambas as árvores possuem as características de serem *top-down*, gananciosas e recursivas. O *top-down* refere-se ao fato de que toda a amostragem de observações se inicia no topo da árvore e vai dividindo o espaço preditor em novos ramos conforme avança para a base da árvore (MARTINS, 2016). Gananciosas pois, como já mencionado, o algoritmo se preocupa em encontrar a melhor variável disponível na divisão em que se encontra, não levando em conta divisões futuras que poderiam levar a construção de uma árvore "melhor". Por fim, recursivas pois o algoritmo repete a mesma função seguindo com divisões binárias até que um critério de parada seja definido por um supervisor externo (MARTINS, 2016).

Em ambos os casos, árvores totalmente crescidas são sujeitas ao overfitting, dado que a divisão mais avançada alcançada seria aquela com uma só observação em cada folha. Para lidar com tal situação, pode-se utilizar a poda ou técnicas de ensembles.

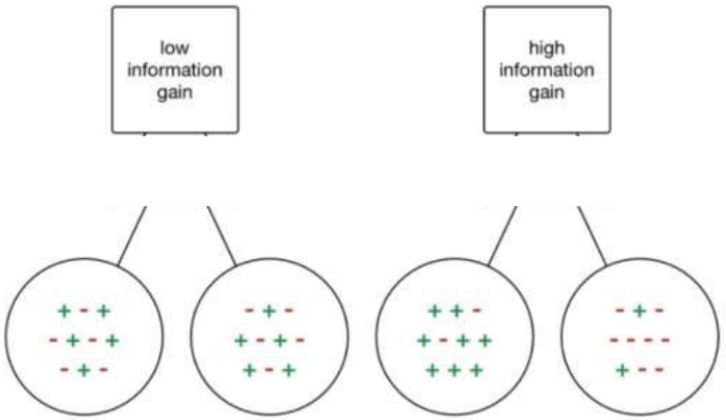
Construção de uma árvore de decisão

Uma das principais tarefas ao se construir uma árvore de decisão diz respeito à posição dos nós. Quem será o nó raiz, qual será o nó da esquerda e da direita são questões que devem ser levadas em consideração.

Existem vários algoritmos para decidir a forma como os nós são divididos. A natureza da variável resposta (categórica ou numérica) também influenciam na escolha do algoritmo, porém, de forma geral, todos os métodos buscam a criação de nós puros, ou seja, que a homogeneidade dos dados de cada ramo seja a maior possível (MARTINS, 2016).

Um algoritmo popularmente usado, tanto para árvores de regressão quanto de classificação, é o ganho de informação e entropia exemplificado pela figura 2.18. Essas duas grandezas dizem respeito ao nível de desordem e uniformidade dos dados (SACRAMENTO, 2021). Quanto mais alta a entropia, mais caóticos e heterogêneos encontram-se os dados, portanto, quanto menor a entropia, mais organizados e homogêneos estão os dados (SACRAMENTO, 2021).

Figura 2.18. Dados com alto e baixo ganho de informação.



Fonte: Adaptado de SACRAMENTO (2021)

Para definir as posições dos nós, deve-se calcular a entropia das classes de saída e o ganho de informação da base de entrada. O conjunto que possuir o maior ganho de informação dentre os atributos será o nó-raiz. Para se calcular os nós da esquerda e da direita, o mesmo processo deve ser refeito para cada um com o conjunto de dados que referente a condição (SACRAMENTO, 2021).

A divisão da base de dados depende da condição estabelecida para o nó. Se uma condição for, por exemplo, valores maiores que X para um certo lado, esse possuirá somente valores maiores que X. O ganho de informação é baseado a partir dessa lógica, dado que quando se analisa um atributo, se os dados forem o mais homogêneos possível, tem-se um ganho de informação maior, afinal, sabe-se que optando por um certo lado da condição, a probabilidade de se saber o valor dado na saída será maior (SACRAMENTO, 2021).

Matematicamente, pode-se dizer que se a amostra for completamente homogênea, a entropia será igual a 0 e se a amostra for dividida em partes iguais (50% e 50%) então a entropia será 1. A fórmula para o cálculo de entropia, segundo MITCHELL (2022), no caso de uma divisão binária é dada abaixo.

$$Entropia(S) = - p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus} \quad (2.17)$$

No qual p positivo indica a proporção de exemplos positivos na amostra, enquanto p negativo indica a proporção de exemplos negativos. Os passos para se definir os nós podem então ser simplificado da seguinte forma segundo MARTINS (2016):

- Calcula-se a entropia do nó-pai;
- Calcula-se a entropia de cada nó individual e a média ponderada de todos os sub-nós disponíveis na divisão.

Vantagens e desvantagens das árvores de decisão

Um questionamento pertinente quando se trabalha com a modelagem de dados diz respeito a qual metodologia utilizar. Se há a possibilidade de se utilizar modelos de regressão (linear ou logística), por que optar por modelos de árvores? A resposta para isso é que, em teoria, todos os modelos são possíveis de serem empregados, cabendo ao engenheiro avaliar o contexto do problema (MARTINS, 2016). Quando se sabe que a variável de entrada e a variável resposta possuem um comportamento linear, um modelo de regressão linear terá uma performance melhor do que modelos que se utilizam de nós de decisão como árvores ou redes

neurais (MARTINS, 2016). Por outro lado, se as variáveis se correlacionam de maneira mais complexa e tendendo a modelos não-lineares, as árvores de decisão poderão melhor suprir as demandas. JAMES et. al. (2013) destaca algumas das vantagens e desvantagens dos modelos de árvores. Como vantagens, pode-se destacar:

- Os modelos de árvores lidam muito bem com situações de classificação, não sendo necessário atribuir valores numéricos binários como 0 e 1.
- As árvores possuem uma visualização gráfica mais amigável e mais didática quando se pretende demonstrar os dados.
- Os modelos de árvore representam, de modo geral, um modo mais próximo ao como o cérebro humano interpreta e toma decisões, diferentemente dos modelos de regressão.

Como desvantagens, tem-se:

- As árvores não possuem o mesmo nível de acurácia para realizar previsões quanto a alguns modelos de regressão.
- As árvores não são muito robustas, ou seja, pequenas mudanças nos dados podem gerar resultados bastante diferentes na configuração da árvore final.

Lidando com *overfitting*: métodos de poda e *ensembles*

Métodos de *ensemble* consistem essencialmente em realizar agrupamentos de modelos preditivos de forma a melhorar sua precisão (MARTINS, 2016). Essa tática é necessária pois à medida que se aumenta a complexidade de um modelo, seu viés também diminui, mas em compensação, sua variância aumenta. O modelo ideal deve buscar o equilíbrio entre viés e variância, sendo os ensembles uma boa forma de se analisar essa troca (MARTINS, 2016).

Poda

MARTINS (2016) destaca que um dos problemas das árvores de decisão, como já mencionado, é o fato de seu algoritmo ser ganancioso, ou seja, ele não "enxerga" o resultado final como um todo e sim a classificação momentânea em cada nó. Dessa forma, a poda (ou pruning) se torna uma alternativa ao problema de overfitting ao fazer com que se observe alguns passos a frente. A ideia do método consiste em deixar a árvore realizar suas decisões até atingir um alto grau de profundidade e então remover as folhas que dão resultados

negativos quando comparados aos ramos anteriores. Um exemplo simples seria pensar que uma divisão fornece uma perda de 10 porém, logo em seguida, um ganho de 20. O resultado da folha final seria um ganho líquido de 10, porém, uma árvore de decisão pararia sua iteração logo no primeiro valor negativo que encontrasse. Com a poda, seria possível visualizar o ganho positivo final e manter ambas as folhas.

Florestas aleatórias

É um dos métodos mais versáteis de ML e pode ser aplicado tanto para problemas de classificação quanto de regressão, além de contornar algumas falhas na etapa do tratamento de dados como dados faltantes ou *outliers* (MARTINS, 2016). É um dos modelos de ensembles aqui apresentados pois combinam modelos mais fracos para se criar um mais forte como mostrado na figura 2.19.

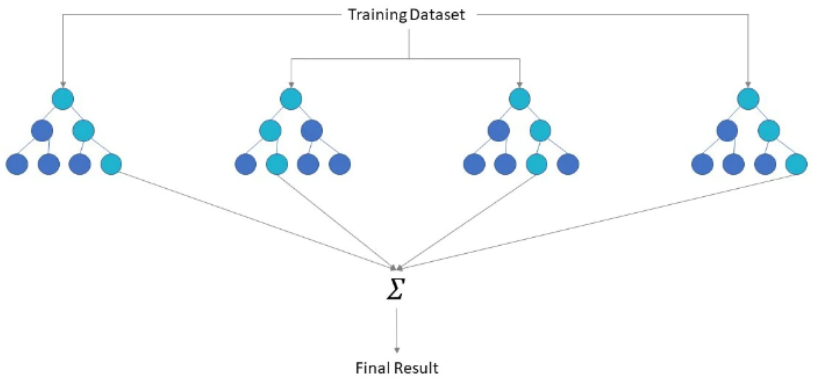
O algoritmo é formado por um conjunto de árvores e em cada uma é injetada uma amostra retirada de uma amostragem maior de treinamento com substituição, sendo também chamada de amostragem bootstrap. Da amostragem, um terço é deixado de lado como dados de teste e será utilizado posteriormente. Uma instância aleatória é então injetada no conjunto de dados de treino atual de forma a adicionar mais diversidade aos dados e diminuindo a correlação entre as árvores. Para problemas de regressão, a decisão final será a média das decisões individuais de cada árvore e para problemas de classificação, será a variável categórica de maior frequência dentre as árvores. Por fim, a amostra que foi deixada de lado é submetida a uma validação-cruzada, finalizando a predição (IBM CLOUD EDUCATION, 2020). O motivo por trás dessa validação final é mensurar erros, mostrando que a estimativa dos dados de teste são tão acurados quando um conjunto de dados de mesmo tamanho para treinamento, removendo a necessidade de um conjunto extra para testes.

Em síntese, MARTINS (2016) destaca o seguinte passo a passo para a performance do algoritmo:

- Assume-se que N casos do conjunto de treinamento e uma amostra é selecionada aleatoriamente, mas com substituição.
- Se houver M variáveis de entrada, um número $m < M$ é escolhido de modo que, em cada nó, m variáveis sejam selecionadas. O valor de m é sempre mantido constante enquanto a floresta cresce.

- Cada árvore é cultivada na maior extensão possível sem poda.
- Os dados de resposta são obtidos agregando-se os valores de maior preponderância no caso da árvore de classificação ou por meio da média das respostas no caso da árvore de regressão.

Figura 2.19. Esquema de uma floresta aleatória



Fonte: Adaptado de MARTINS (2016)

2.6 REGRESSÃO LINEAR

Segundo JAMES et al.(2013), a regressão linear é uma abordagem conhecida e simples utilizada para se predizer uma resposta quantitativa *Y* a partir de uma variável de entrada *X*. Ela assume que há uma relação linear entre as duas variáveis, sendo matematicamente expressa como:

$$Y \approx \beta_0 + \beta_1 X \tag{2.18}$$

Como já esperado, na prática, os coeficientes *B0* e *B1* são desconhecidos e para uso da função estatística, é necessário utilizar o conjunto de dados das observações para obter uma estimativa de seus valores. Em outras palavras, quer-se descobrir um *B0* e *B1* tal que a a linha resultante seja o mais próxima possível do conjunto de n observações.

A abordagem do Ajuste dos Mínimos Quadrados é uma das técnicas mais conhecidas para realizar essa tarefa. Suponha-se que para a função linear, o erro associado seja dado por:

$$e_i = y_i - \hat{y}_i \tag{2.19}$$

Esse erro representa o resíduo da função, ou seja, a diferença entre a observação e a resposta de ordem prevista pelo modelo linear desenhado. A Soma Residual dos Quadrados (RSS) é dada por:

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 \tag{2.20}$$

Na forma expandida, pode-se considerar:

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \tag{2.21}$$

A abordagem dos mínimos quadrados visa escolher um *B0* e *B1* tal que o RSS seja reduzido ao seu valor mínimo. Os cálculos mostram que, para tal, os minimizadores dos parâmetros podem ser dados pelas correlações:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.22}$$

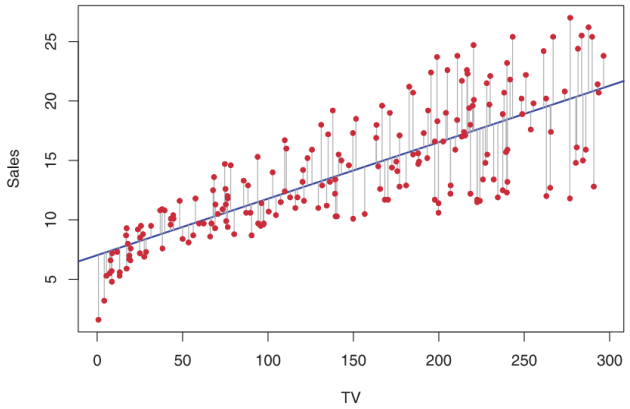
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.23}$$

No qual as médias das amostras para y e x dos dados observados são dadas por:

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.24}$$

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.25}$$

Figura 2.20. Regressão linear utilizando-se a técnica dos mínimos quadrados.



Fonte: Adaptado de JAMES et al.(2013)

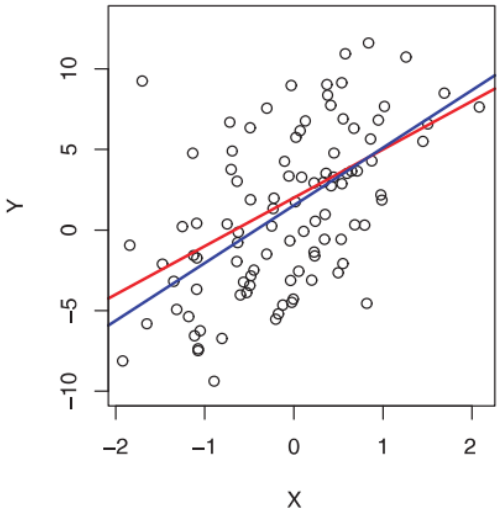
Acurácia dos coeficientes

JAMES et al.(2013) elucida que, partindo-se da premissa que um modelo possui forma linear, pode-se escrever a função na forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.26}$$

No qual ε representa o termo de erro aleatório com média zero. *B0* é o termo que intercepta o eixo Y quando $X = 0$ e *B1* é termo associado à inclinação da reta, ou seja, o aumento médio de Y associado ao aumento de uma unidade em X. O termo de erro, nesse caso, inclui em teoria o desvio assumido com o modelo linear dado que a verdadeira função pode ser não-linear, podendo haver outras variáveis que causam variações em Y. O termo de erro ε é tipicamente assumido independente de X (JAMES et al., 2013). A fórmula acima representa a regressão linear associada à população, ou seja, a melhor aproximação da verdadeira relação entre X e Y.

Figura 2.21. Esquema da curva de regressão linear da população e da curva obtida pelos mínimos quadrados



Fonte: Adaptado de JAMES et al.(2013)

Segundo JAMES et al.(2013), as linhas ilustradas na figura 2.21 (em vermelho referente a população e em azul referente a amostra) são uma extensão estatística que se utiliza de informação de uma amostra para extrapolar as características de uma população maior. Por exemplo, suponha que que se tem interesse em descobrir a média μ de uma variável qualquer *Y* referente a uma população. Como não é possível saber diretamente o valor de μ, pode-se dizer que uma estimativa razoável seria supor que μ*média*= *ymédia* (O primeiro referindo-se a média da amostra e o segundo referindo-se a média real da população). Ambas são diferentes, porém de forma geral, a média da amostra provê uma boa estimativa da média real.

Para se medir essa discrepância entre a média real e amostral, utiliza-se a formulação do desvio padrão (SE):

$$\mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{2.27}$$

De forma geral, o desvio padrão é capaz de fornecer a quantidade média pela qual as estimativas de ambas as médias (real e do modelo) se distanciam. A equação também demonstra que esse desvio diminui conforme *n* cresce pois, quanto mais amostras tem-se,

menor será o erro de estimativa (JAMES et al., 2013). Da mesma forma, pode-se inferir quão perto *B0amostra* e *B1amostra* estão próximos de *B0real* e *B1real* utilizando-se:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{2.28}$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.29}$$

A estimativa para σ é chamada erro residual médio e é dado por:

$$\sqrt{\mathrm{RSS}/(n-2)} \tag{2.30}$$

No qual RSS representa, como já visto, a soma residual dos quadrados.

Acurácia do modelo

JAMES et al.(2013) ressalta que o desvio padrão, em essência, é um auxiliador para o teste estatístico de hipóteses dos coeficientes ou parâmetros do modelo. A hipótese nula H0 seria não haver relação entre *X* e *Y*, enquanto a hipótese alternativa H1 seria que existe alguma relação entre as variáveis.

Ao rejeitar a hipótese nula em favor da alternativa, é importante realizar a quantificação de quanto a regressão linear se encaixa nos dados observados. A qualidade da regressão é medida, tipicamente, utilizando-se o erro residual padrão (RSS) e R^2, um outro elemento estatístico.

Quando se constrói um modelo, para cada informação há um erro associado. O RSE, como já pontuado, nada mais é do que do desvio padrão desse erro ε (JAMES et al.,2013). Sua forma estendida considerando-se os desvios das observações associadas aos resultados observados dados n pontos é dada por:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} \tag{2.31}$$

No entanto, o RSE provê uma medida absoluta da falta de encaixe entre o modelo e os dados. Por ser medido em unidades de *Y*, nem sempre é possível se ter certeza de quão bom foi o valor encontrado para RSE (JAMES et al.,2013). Para se contornar esse problema, mais comumente se utiliza R^2, como uma análise alternativa, o R^2 assume a forma de uma proporção da variância e, portanto, possui escala entre 0 e 1, independente da escala associada a *Y*, sendo descrito por:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{2.32}$$

Sendo TSS a soma total dos quadrados, dada por:

$$\text{TSS} = \sum (y_i - \bar{y})^2 \tag{2.33}$$

Pode-se observar que R^2 consiste em nada mais que uma proporção entre o erro residual padrão e a soma total dos quadrados. Na prática, R^2 indica a proporção da variabilidade em Y que pode ser explicada utilizando-se X (JAMES et al.,2013).

Para além de R^2, há a relação de correlação entre duas variáveis, definida como:

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2.34}$$

A correlação pode ser utilizada também para explicar a variação de *Y* dado *X* em uma regressão linear simples ($R^2 = \text{Cor}^2$). Porém, para o modelo de regressão linear múltipla, nem sempre esse é o caso, tendo a fórmula de correlação certas limitações (JAMES et al.,2013) a serem exemplificadas a seguir.

Regressão linear múltipla

A regressão linear simples é uma boa abordagem para se predizer uma resposta quando há apenas uma variável de entrada, porém, com dados reais, isso não é o mais comum. Para várias variáveis preditoras, o modelo de regressão linear múltipla se torna mais adequado (JAMES et al.,2013).

Ao invés de se modelar várias regressões lineares simples separadamente para cada variável preditora, uma melhor abordagem é estender o modelo simples para um alternativo para que se possa acomodar, mutuamente, várias variáveis preditoras que se digam influenciar o resultado. Isso pode ser feito atribuindo-se a cada preditor um coeficiente angular separado, porém, em um único modelo (JAMES et al.,2013). De forma geral, para p variáveis de entrada ou preditoras, a regressão linear múltipla possui a forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{2.35}$$

Em uma configuração multivariável, a regressão linear com 2 variáveis de entrada assume a forma de um plano. O método dos mínimos quadrados, em analogia com a regressão simples, visa geometricamente encontrar um plano que mais minimize a soma ao quadrado das distâncias verticais de cada observação ao plano (JAMES et al.,2013).

JAMES et al. (2013) exemplifica um caso de estudo de impacto do alcance da propaganda por meio de diversos veículos de comunicação. Em um método de regressão linear simples performado separadamente para os meios A, B e C, nota-se que, para as variáveis preditoras A e B, os coeficientes associados mostram-se próximos tanto na regressão simples quanto múltipla. No entanto, para C, o parâmetro indicativo de sua influência na variável resposta é praticamente nulo quando aplicados ambos os modelos de regressão.

Tabela 2.1. Coeficientes de regressão linear simples para os meios A e B separadamente

Meio	***B1***	***B0***
A	0.203	9.312
B	0.055	12.351

Fonte: Adaptado de JAMES et al.(2013)

Tabela 2.2. Coeficientes para regressão linear múltipla aplicados para os meios A, B e C simultaneamente

Meio	*B1*	*B0*
A	0.046	2.939
B	0.189	2.939
C	0.001	2.939

Fonte: Adaptado de JAMES et al.(2013)

Para JAMES et al.(2013), essa comparação ilustrada entre as tabelas 2.1 e 2.2 mostra que os resultados podem ser bem diferentes quando aplicados diferentes métodos de ajuste estatístico. Para a abordagem utilizando a regressão simples, o coeficiente encontrado indica o efeito para um aumento unitário no peso do meio C, sem levar em conta os fatores A e B. Porém, na regressão múltipla, as variáveis A e B são mantidas fixas enquanto tenta-se uma variação unitária em C. Por esse motivo, em uma análise simples que examina apenas uma variável de preditora e sua resposta associada, observa-se também que o aumento de C ocorre simultaneamente com o aumento das variáveis A e B, indicando um efeito de correlação entre C e a resposta, porém não de causalidade. O efeito de C ganha "crédito" pelo efeito de outras variáveis.Por esse motivo, uma análise mais detalhada, tal como um olhar crítico sobre o peso das variáveis no modelo a ser desenhado representa uma etapa importante na análise dos dados. Um fenômeno similar a esse ocorre também para a regressão logística múltipla, como descrito mais adiante.

Ajuste do modelo linear: técnicas de *shrinkage*

JAMES et al.(2013) em seu livro ressalta que o Ajuste dos Mínimos Quadrados, apesar de muito empregado, possui algumas limitações que podem ser diminuídas ao se empregar alternativas de modelagem linear. Os 2 fatores preponderantes que esses ajustes introduzem dizem respeito a:

● Acuracidade: Dado que a relação entre a variável resposta e suas preditoras é de fato linear, o Ajuste dos Mínimos Quadrados produz uma estimativa de baixo viés. Se o tamanho da amostra for consideravelmente maior que o número de observações ($n>>p$) então o Ajuste dos Mínimos Quadrados tenderá a ter baixa variância também e, portanto, performará bem

em observações de teste. Entretanto, se *n* não for suficientemente grande, haverá grande variabilidade e alto viés no modelo de Ajuste dos Mínimos Quadrados, resultando em performances pobres na etapa de teste. Os ajuste de *shrinkage* ou "diminuição" aqui apresentados possuem a função de reduzir a variância ao custo de uma baixa variação no viés ao se diminuir os efeitos de alguns coeficientes, melhorando significativamente a acuracidade e levando a melhores predições ao se replicar o modelo em dados de teste.

●Interpretabilidade: É relativamente comum que diversas variáveis preditoras utilizadas no modelo de regressão linear múltipla não estejam diretamente associadas com a variável resposta. Ao se incluir variáveis irrelevantes no processo, o modelo pode se tornar demasiado complexo e por consequência, com menor interpretabilidade. Ao diminuir ou remover essas variáveis, o modelo torna-se de mais fácil compreensão.

As técnicas de *shrinkage* aqui apresentadas envolvem imputar todas as *p* variáveis preditoras ao modelo (outras técnicas envolvem uma pré-seleção das variáveis que devem ser adicionadas ou simplesmente excluídas). Porém, alguns coeficientes relativos aos Mínimos Quadrados são diminuídos consideravelmente e, portanto, diminuem a variância do modelo. Dependendo da técnica utilizada, os coeficientes podem ser diminuídos para o valor exato de zero, anulando seu efeito, realizando, dessa forma, uma espécie de seleção de variáveis. As duas mais conhecidas, segundo técnicas de shrinkage segundo JAMES et al.(2013) são a Regressão Ridge e a Regressão Lasso, mostradas a seguir.

Regressão Ridge

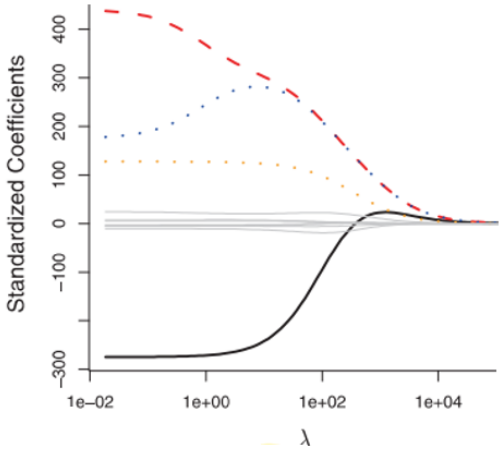
A Regressão Ridge é bastante similar ao método dos mínimos quadrados, exceto que os coeficientes são estimados de forma a se minimizar uma parte dos efeitos desses. Ela possui a seguinte forma:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2, \tag{2.36}$$

No qual λ é um parâmetro de refinamento ou tuning como visto mais adiante, sendo determinado separadamente. A equação promove um balanço de dois critérios. Segundo JAMES et al.(2013), a Regressão Ridge busca por coeficientes capazes de minimizar o efeito de RSS e, portanto, busca um ajuste mais afinado para os Mínimos Quadrados. O segundo

termo da equação, chamado de penalidade de shrinkage, é pequeno quando os coeficientes *B0, B1… Bp* são próximos de zero e logo, fazem o valor do somatório inteiro tender a zero. O parâmetro de tuning λ auxilia como um controlador do impacto entre esses dois termos. Quando λ = 0, a penalidade possui efeito nulo e a Regressão Ridge produz a própria estimativa dos Mínimos Quadrados, porém, quando λ tende ao infinito, o impacto da penalidade aumenta, fazendo com que os coeficientes da regressão tendam a zero. Para cada valor de λ, a Regressão Ridge produz um conjunto de coeficientes, diferentemente dos Mínimos Quadrados, que produz apenas um conjunto de coeficientes.

Figura 2.22. Aplicação da Regressão Ridge a um conjunto de dados

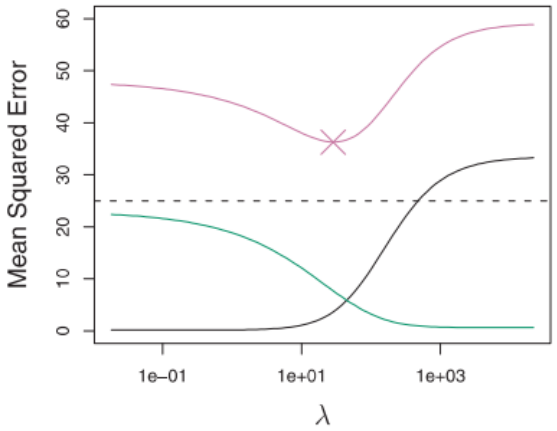


Fonte: Adaptado de JAMES et al.(2013)

Como observado na figura 2.22, cada curva corresponde a um conjunto de coeficientes de uma regressão Ridge estimados para dez variáveis (as curvas vermelha, azul, amarela e preta representando as variáveis de maior relevância, enquanto as cinzas, de menor relevância), plotados para cada valor de λ. Quando λ está próximo de zero, os coeficientes de Ridge correspondentes são os mesmo que dos Mínimos Quadrados, porém, quando λ tende a valores maiores, os coeficientes "encolhem" para valores próximos a zero e quando λ é suficientemente grande, os coeficiente são praticamente zerados. Esse último caso seria o modelo nulo, no qual não há variáveis preditoras (JAMES et al., 2013).

A Regressão Ridge possui a vantagem sobre os mínimos quadrados quanto ao balanço viés-variância. Enquanto λ sobe, a flexibilidade de regressão Ridge diminui, levando um modelo de menor variância porém maior viés. No mínimos quadrados, que corresponde a uma regressão Ridge com λ = 0, a variância é alta, porém não há viés. Porém, como já pontuado, com a diminuição dos coeficientes Ridge ocorre uma acentuada diminuição da variância e um aumento relativamente pequeno do viés (JAMES et al., 2013).

Figura 2.23. Troca viés-variância para a regressão Ridge



Fonte: Adaptado de JAMES et al.(2013)

Como ilustrado na figura 2.23, conforme o viés quadrado curva preta) cresce com o aumento de λ, a variância (curva verde) decresce e o ponto de equilíbrio representa o menor MSE para a Regressão Ridge (curva lilás).

Regressão Lasso

Para JAMES et al.(2013), a regressão Ridge possui a desvantagem de não eliminar nenhum coeficiente, mesmo que seu fator de penalidade amenize essa alta variabilidade. Embora não afete diretamente a acuracidade da predição, dependendo da quantidade de variáveis envolvidas, perde-se em termos de interpretabilidade do modelo.

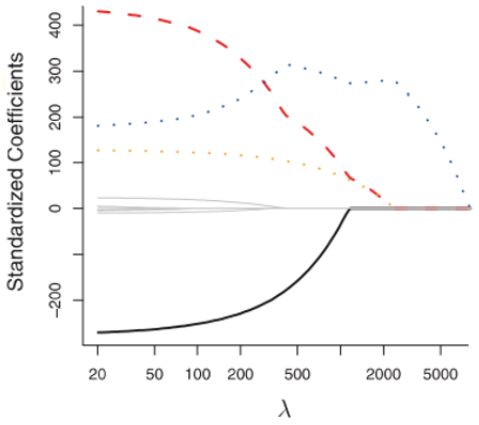
A Regressão Lasso representa uma alternativa recente à Regressão Ridge que sobrepõe-se a essa característica ao trocar o termo de penalidade quadrático por um termo em

módulo (JAMES et al., 2013). Dessa forma, obtêm-se os coeficientes da Regressão Lasso, BL, minimizando-se a relação:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j| \tag{2.37}$$

A Regressão Lasso acaba por performar uma seleção de variáveis quando λ é suficientemente grande, pois diferente da regressão Ridge, ela possui a capacidade de zerar os coeficientes e, consequentemente, as variáveis que possuam menos peso na variável resultante, como mostrado na figura 2.24.. Dessa forma, obtém-se um modelo mais "limpo" e de melhor compreensão. Para ambos os métodos de *shrinkage*, a escolha de λ torna-se um fator crítico na obtenção do modelo desejado (JAMES et al., 2013).

Figura 2.24. Aplicação da regressão Lasso a um conjunto de dados.



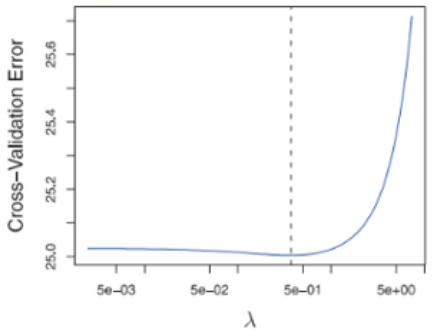
Fonte: Adaptado de JAMES et al.(2013)

Na Regressão Lasso, quando λ = 0, assim como na Regressão Ridge, o modelo torna-se equivalente aos mínimos quadrados e quando λ torna-se suficientemente grande, o modelo torna-se nulo. Entretanto, é possível observar que entre os dois extremos, há uma significativa diferença entre ambos os tipos de regressão. Dependendo do valor de λ escolhido, a regressão Lasso consegue produzir um modelo com uma quantidade de entradas p variável, enquanto na Regressão Ridge, sempre haverá, na saída, o mesmo número de coeficientes de entrada, independente de λ (JAMES et al., 2013).

Seleção de parâmetros de tuning

JAMES et al. (2013) explica que, para estimar-se os valores de λ (parâmetro de refinamento ou tuning), escolhe-se uma gama de valores e computa-se uma validação cruzada e seus erros para cada valor de λ. Dessa forma, o valor a ser escolhido é aquele que representa o menor valor de erro. Finalmente, o modelo é remodelado utilizando-se todas as variáveis observadas e um parâmetro de tuning. A figura 2.25 representa a escolha e λ por meio de uma validação cruzada LOOCV na regressão Ridge. Nesse caso, o valor de λ é relativamente baixo, indicando que o termo de *shrinkage* também será pequeno. Além do mais, a curva não é muito acentuada para uma gama grande de valores, indicando vários valores de λ que retornariam um valor igual de erro. Nesses casos, a aplicação direta dos mínimos quadrados pode ser mais simples e eficiente.

Figura 2.25. Aplicação de LOOCV para seleção de λ

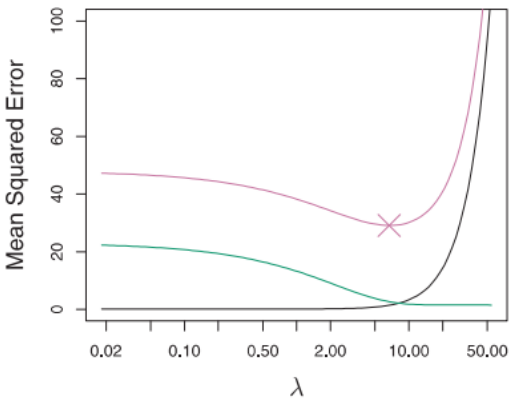


Fonte: Adaptado de JAMES et al.(2013)

Comparação: Ridge e Lasso

Segundo (JAMES et al., 2013) a Regressão Lasso possui uma maior vantagem em relação a Ridge ao produzir um modelo mais simples e interpretável, que envolve uma porção menor de variáveis preditoras a serem analisadas. A figura 2.26 mostra que o erro de MSE, assim como para a Regressão Ridge, torna-se mínimo no equilíbrio viés-variância (a curva lilás representa o MSE, a curva verde a variância e a preta o viés).

Figura 2.26. Troca viés-variância para a regressão Lasso



Fonte: Adaptado de JAMES et al.(2013)

De forma geral, pode-se esperar que a Regressão Lasso tenha melhor performance quando tem-se uma quantidade menor de variáveis preditoras com coeficientes significativos e variáveis restantes com coeficientes bem pequenos ou tendendo a zero (JAMES et al., 2013). A Regressão Ridge tende a performar melhor quando uma variável resposta é função de muitas preditoras, todas com coeficientes de proporções similares. A escolha do modelo de regressão irá depender de uma análise prévia se necessária a eliminação de parte da variância para ter-se um modelo mais interpretável e acurado, além de se definir se pode ser benéfico, para os resultados, uma quantia menor de variáveis preditoras (JAMES et al., 2013).

2.7 REGRESSÃO LOGÍSTICA

Predizer uma resposta qualitativa para uma dada observação é chamada de classificação pois envolve atribuir a ela uma certa categoria ou classe. Os métodos de classificação no geral funcionam predizendo a probabilidade de cada categoria ocorrer dada uma variável qualitativa, sendo essa a lógica básica por trás desse tipo de problema (JAMES et al., 2013).

Diferença entre regressão linear e logística

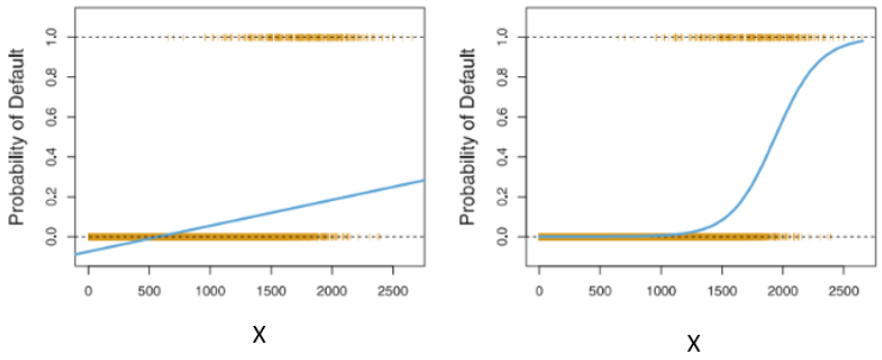
Embora haja meios de se atribuir valores numéricos às classes de saída, é importante ressaltar que nem sempre a regressão linear será apropriada para problemas de classificação ou respostas qualitativas. JAMES et al. (2013) exemplifica supondo-se que uma variável Y será classificada de acordo com uma dada numeração A-1, B-2, C-3.

Supondo-se que A, B e C sejam classes qualitativas. Ao utilizar-se, por exemplo, um modelo derivado de uma regressão linear por mínimos quadrados para predizer *Y*, estaria-se assumindo uma ordem entre os fatores, no qual A estaria na primeira posição, B na segunda e C na terceira, além de se assumir que as diferenças entre o fator A e B ou B e C é a mesma. Qualquer mudança na ordem dessas classes afetaria o modelo linear de modo a produzir outro com parâmetros completamente diferentes e por consequência, diferentes previsões dado um conjunto de dados teste (JAMES et al., 2013).

Fundamentos da regressão logística

A regressão logística é um método que considera a probabilidade de que uma variável de saída *Y* pertença a uma determinada categoria. Ao invés, portanto, de se modelar a resposta *Y* diretamente, tem-se um conjunto de dados *default*, ou seja, variáveis com valores pré-determinados porém não definidos (por exemplo, sim e não) e o cálculo de probabilidades do mesmos (JAMES et al., 2013).

Figura 2.27. Diferença de uma classificação linear e logística



Fonte: Adaptado de JAMES et al. (2013)

A figura 2.27 mostra as diferenças de performance ao se utilizar uma regressão linear e uma regressão logística. Pode-se observar que o modelo linear inclui valores negativos, o que não condiz com a realidade da probabilidade. As retas horizontais indicam as variáveis default de saída variando num intervalo de 0 a 1 (trata-se da codificação das variáveis, podendo 0 significar "sim" e 1 "não"). No modelo de regressão logística representada pela curva em S, as respostas oscilam apenas dentro do intervalo delimitado, gerando mais sentido matemático para a probabilidade.

Uma probabilidade para um conjunto de dados default dado uma variável de entrada X, por exemplo, pode ser escrita da seguinte forma (JAMES et al., 2013).:

$$\Pr(X) = \Pr(Y = \text{default}|X) \tag{2.38}$$

Os valores de Pr, como já discutido, podem variar no intervalo de 0 a 1, logo, para qualquer valor de X, uma predição de saída *default* será feita. Pode-se dizer, por exemplo, que uma saída pode resultar em default = "sim" dado que P(X) > 0,5.

A questão maior seria, portanto, como realizar a modelagem da função (42) da melhor maneira possível. Como observado previamente, o modelo linear aplicado a uma probabilidade se apresentaria na forma:

$$p(X) = \beta_0 + \beta_1 X. \tag{2.39}$$

Porém, por motivos citados, ela não se adequa muito bem à tarefa. Para contornar os problemas relacionados ao range e a distribuição dos pontos, comumente se utiliza a função logística adaptada com os coeficientes de um modelo linear:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{2.40}$$

Como observado, a função logística aplicada a uma regressão probabilística possui a característica de possuir um balanço mais acentuado, ou seja, para variáveis de entrada de menor valor, por exemplo, a função irá retornar probabilidades próximas (porém nunca abaixo) de 0. O mesmo se aplica a valores elevados de entrada, no qual a função pende a valores próximos de 1, porém sem extrapolar seus limites. A função logística possui um formato S que permite, para qualquer valor de X, uma saída mais sensível, tornando mais acurado o objetivo final do método, que é a classificação (JAMES et al., 2013).

A função logística também possibilita capturar melhor a distribuição de probabilidades do que uma regressão linear. No exemplo ilustrado, fazendo-se uma média das respostas obtidas pelos dados de treino, obtêm-se valores similares para todas as entradas (JAMES et al., 2013).

Com uma manipulação apropriada, chega-se à forma mais usual da regressão logística. O termo da esquerda é comumente referido como probabilidade logarítmica ou, abreviadamente, logit.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{2.41}$$

Em uma regressão linear, o termo *B1* fornece a variação em *Y* dado um incremento unitário de *X*. No caso da regressão logística, um incremento unitário em *X* altera o logit em *B1*, ou seja, multiplica a probabilidade por e^{B1}.

Os coeficientes *B0* e *B1* devem ser determinados com base nos dados de treinamento. Para o modelo linear, o método dos mínimos quadrados é bastante utilizado para esse fim. No caso da regressão logística, utiliza-se o método de máxima verossimilhança devido às suas propriedades estatísticas mais acuradas. O procedimento padrão para se utilizar essa abordagem para que se modele uma regressão logística é estimar *B0* e *B1* de modo que a probabilidade calculada *p(xi*) de default para cada ponto observado, utilizando-se a equação de logit, corresponda o mais próximo possível ao status *default* esperado para a observação (JAMES et al., 2013).

Em outras palavras, tenta-se encontrar um *B0* e *B1* de forma que, encaixando-os na equação de regressão logit retorne um valor próximo a 1 para indivíduos que foram classificados em um default (por exemplo, default = "sim") ou próximo a 0 no caso do default alternativo (por exemplo, default = "não").

A função de máxima verossimilhança é expressa por:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{2.42}$$

Os parâmetros *B0* e *B1* são escolhidos, portanto, de forma a se maximizar a função.

Regressão logística múltipla

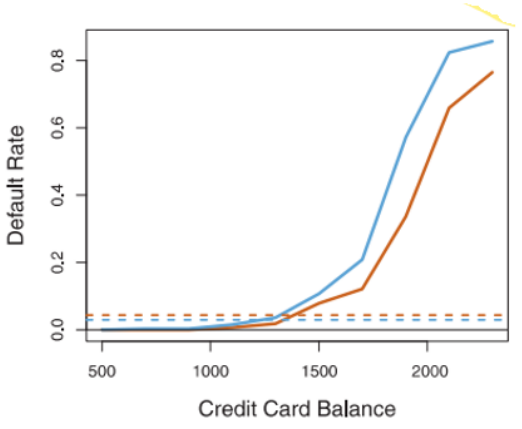
Uma outra abordagem dos problemas de regressão logística é quanto tem-se, da mesma forma que a regressão simples, uma saída de classificação binária, porém agora, considerando-se várias variáveis de entrada. A função para regressão logística múltipla pode ser representada pela seguinte função:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2.43}$$

Quando se analisa um problema de várias entradas, é importante se analisar o possível nível de correlação entre elas. De maneira geral, a análise do p-value para as variáveis de entrada com relação a saída indica se há probabilidade de algum nível de correlação. Para valores de p-value pequenos, há um forte indicativo de existência de correlação, mostrando que todas influenciam, de alguma maneira, a variável de saída (JAMES et al., 2013).

JAMES et al. (2013) elucida um exemplo de estudo entre o nível de crédito de estudantes (em laranja) e não-estudantes (em azul) com relação a possibilidade de calote. No caso, as variáveis de entrada seriam o valor do crédito, a renda e o status (estudante ou não).

Figura 2.28. Regressão logística múltipla com evidência de *coufounding*



Fonte: Adaptado de JAMES et al. (2013)

A princípio, os resultados parecem contraditórios dado que as probabilidades de calote são sempre superiores para os não-estudantes, porém, na média geral de calote, os estudantes

possuem maior valor como mostrado na figura 2.28. A resposta para isso é que as variáveis crédito e status estão correlacionadas. Estudantes tendem a possuir maiores níveis de débito e maiores níveis de débito estão associados a maiores chances de calote. No entanto, para um valor fixo de crédito, um estudante possui menos chances de calote do que um não-estudante.

Esse exemplo mostra claramente a importância de se avaliar, em uma regressão com múltiplas entradas, a necessidade de se avaliar as possíveis correlações entre as variáveis em si e com o resultado, visando melhor compreensão da metodologia. Além do mais, uma regressão com uma única entrada pode demonstrar resultados parciais e diferentes quando analisadas mais entradas, especialmente quando estas encontram-se correlacionadas (JAMES et al., 2013).

3. METODOLOGIA

A metodologia empregada baseou-se no levantamento bibliográfico sobre o uso de *Machine Learning* na indústria química. Para a compreensão dos conceitos estatísticos por trás de cada método, utilizou-se o livro base An Introduction to Statistical Learning de Gareth James e demais autores. Para melhor exemplificar a teoria e constatar as vantagens e limitações dessas aplicações, buscou-se artigos de apoio relacionando o uso do ML aplicado a estudos experimentais de engenharia química.

No primeiro estudo, TSO e YAU (2007) destacam que o consumo de energia em Hong Kong cresceu consideravelmente nas últimas décadas devido ao crescimento populacional e desenvolvimento econômico da cidade. Particularmente, houve um crescimento substancial no consumo total de eletricidade no setor doméstico, indo de 1059 GWh em 1971 para 9111 GWh em 2001. É destacado também que, atualmente, companhias do setor de energia têm se utilizado de técnicas estatísticas para previsão e análises de regressão para estimar demandas de energia futuras. Com o rápido desenvolvimento da mineração de dados, abordagens alternativas para essa função como árvores de decisão e redes neurais têm se tornado populares e conhecidas por serem mais fáceis de operar. O estudo teve por objetivo, comparar a acuracidade na previsão da demanda energética em Hong Kong utilizando-se 3 abordagens de ML: regressão linear múltipla, árvores de decisão e redes neurais.

A coleta de dados consistiu em um levantamento em 2 fases durante o verão e o inverno entre 1999-2000. A pesquisa considerou residências com consumo de pelo menos 100 kWh, tendo sido submetido um questionário aos moradores coletando dados sobre os níveis de uso dos aparelhos elétricos e taxa de consumo durante o verão e o inverno.

Após a coleta de respostas, uma base de dados foi desenvolvida contendo taxas de consumo dos aplicantes e tempo de uso para cada aparelho. A amostragem baseou-se na estratificação do tipo de habitação: locação publica, residências subsidiadas pelo governo, empreendimentos privados e coabitações. O percentual de dados alocados em cada grupo foi feito proporcionalmente com o nível de consumo energético e tamanho da população.

Com base nos dados coletados, utilizou-se os 3 algoritmos de ML citados visando analisar o consumo de energia. Para tal, utilizou-se o software estatístico SAS Enterprise Miner.

No segundo estudo apresentado por KRISHNAN et al. (2018), os autores explicam que vidros de sílica, quando expostos a água, sofrem com o processo de corrosão devido a dissolução. A durabilidade desses vidros em ambientes aquosos tem um papel crítico em várias aplicações e processos (seu uso varia de utensílios de laboratório a vidros de residência). Dependendo do tipo de aplicação, a dissolução do vidro pode ser desejável ou não. Dessa forma, o desenvolvimento de vidros com perfil dissolução conhecidos requerem uma predição acurada da taxa cinética de dissolução do componente de sílica. KRISHNAN et al. (2018) ressalta que essa tarefa, entretanto, não é muito fácil, dada a complexidade estrutural e o fato de que vários mecanismos de dissolução podem ser observados agindo individualmente ou em combinação. Nesse sentido, rotas alternativas envolvendo modelos a base de dados e guiados por ML torna-se uma ferramenta promissora para a compreensão das relações de composição-propriedades em vidros com base na ampla quantidade de dados experimentais já disponíveis.

O objetivo do estudo consistiu, portanto, em investigar habilidade de 3 algoritmos de ML (regressão linear com ajuste lasso, florestas aleatórias e redes neurais) na predição da cinética de dissolução de uma seleção de vidros compostos por aluminosilicato de sódio. Dado o conhecimento prévio da relação não-linear da relação composição-dissolução do silicato, o estudo demonstra que o uso de ANNs aponta para a predição mais confiável dentre as técnicas utilizadas. Para testar a capacidade de predição dos algoritmos, o estudo utiliza como base de dados taxas de dissolução reportadas em literatura e com experimentos

conduzidos em 8 diferentes vidros de aluminossilicato de sódio, com variações em composição, pH do meio e concentração de SiO2 no meio de dissolução, tendo sido os experimentos todos conduzidos a 25°C.

A variável resposta escolhida foi a taxa de lixiviação do SiO2 (em unidades de log[mol SiO2/cm²/s]) dado que essa grandeza captura a dissolução da estrutura de silicato presente no vidro. As variáveis de entrada foram definidas como composição do vidro, pH inicial da solução e o pH no momento da medição. Para as amostras empregadas, escolheu-se dados de treino e de teste de modo que as características da base de dados geral fosse preservada em ambos os grupos. Em outras palavras, um conjunto de dados de teste é tipicamente independente do conjunto de dados de treino, porém, seguindo a mesma probabilidade de distribuição. Para o estudo, utilizou-se, respectivamente, 70% e 30% dos dados para a finalidade de treino e teste.

Por fim, no terceiro estudo aqui apresentado, NAFOUANTI et al. (2021) elucida que, na região norte da China, águas subterrâneas são a principal fonte de água para uso doméstico, agricultura e processos industriais. No entanto, há uma crescente ameaça de contaminação desse meio devido a presença de vários componentes químicos como o flúor. Logo, o entendimento da qualidade da água do solo se faz necessária para seu manuseio correto e visando a sustentabilidade. A concentração de flúor na água é influenciada por diversos fatores, desde processos hidrogeoquímicos naturais até fatores humanos como uso de fertilizantes e irrigação. Devido a essa junção de condições, torna-se difícil determinar o destino do flúor de águas subterrâneas. Nesse contexto, a aplicação de algoritmos de ML podem ajudar a prover uma alternativa eficiente para a predição da contaminação dos lençóis freáticos. As Redes Neurais, como já visto, possuem uma ampla capacidade de detectar relações complexas entre variáveis de entrada e saída, além de conseguir lidar bem com dados contendo erros e com alto volume. As florestas aleatórias conseguem lidar com dados multidimensionais, variáveis numéricas e categóricas, dados faltantes e dados binários. A regressão logística, por fim, é um algoritmo eficiente para treinamento de dados e muito útil para classificação binária. O objetivo do estudo consistiu em identificar a abordagem que melhor previu a contaminação das águas subterrâneas por flúor, empregando-se os algoritmos citados em conjunto com parâmetros físico-químicos da área em avaliação. A determinação das variáveis que mais influenciaram nessa contaminação também foram consideradas como parte do estudo.

O pré-processamento da base de dados consistiu em se imputar as 16 variáveis que supostamente estariam ligadas ao fenômeno em estudo. Os dados foram convertidos em em 2 classes, alocando-se 0 para contaminações de flúor menores que 1mg/L e 1 para contaminações maiores que 1 mg/L. As variáveis independentes foram alocadas em valores entre 0 e 1 para os 3 algoritmos visando realçar a acuracidade do modelo. Os dados foram aleatoriamente divididos em dois grupos, sendo 80% para treino e 20% para teste.

NAFOUANTI et al. (2021) cita que, para a seleção das variáveis mais relevantes na entrada, utilizou-se o teste do chi-quadrado. De forma geral, valores baixos do teste indicam que a variável de entrada possui uma relação mais independente com relação a variável resposta, enquanto valores maiores do teste indicam uma relação de dependência maior e, portanto, devem ser escolhidas como entrada do método.

4. RESULTADOS E DISCUSSÕES

4.1 PREDIÇÃO DO CONSUMO DE ELETRICIDADE

Após a coleta de respostas, uma base de dados foi desenvolvida contendo taxas de consumo dos aplicantes e tempo de uso para cada aparelho. A amostragem baseou-se na estratificação do tipo de habitação: locação publica, residências subsidiadas pelo governo, empreendimentos privados e coabitações. O percentual de dados alocados em cada grupo foi feito proporcionalmente com o nível de consumo energético e tamanho da população.

Com base nos dados coletados, utilizou-se os 3 algoritmos de ML citados visando analisar o consumo de energia. Para tal, utilizou-se o software estatístico SAS Enterprise Miner.

As características de cada algoritmo foram descritas por TSO e YAU (2007) da seguinte forma:

●Regressão linear múltipla: sendo umas das técnicas mais populares e tendo sido previamente descrita, é uma função com múltiplas entradas e parâmetros. Utilizou-se no estudo a técnica dos mínimos quadrados para estimativa dos coeficientes, aplicando-se previamente uma seleção de variáveis de entrada.

●Redes Neurais Artificiais: As ANNs performam bem em aplicações de funções não-lineares. São consideradas uma boa alternativa quando a formulação matemática de um

dado problema é desconhecida ou se tem poucos conhecimentos da relação entre inputs e outputs. Utilizou-se no estudo a arquitetura de perceptron multicamadas (MLP) para se predizer o consumo de eletricidade. A função de ativação empregada foi a sigmóide. A quantidade de camadas ocultas é determinada através da estimativa da generalização de erro de cada rede. Para o modelo em estudo, definiu-se que uma camada oculta era adequada.

● Árvores de decisão: Para o estudo, utilizou-se o modelo CART que inclui árvores tanto de classificação quanto de regressão. Para esse modelo, a variável resposta deve ser categórica, porém as variáveis preditoras podem ser contínuas ou categóricas também. O critério de divisão leva em conta o melhor ganho em redução de entropia. As árvores de decisão tendem a não performar tão bem quanto as redes neurais para dados não-lineares e são mais susceptíveis a ruídos (erros aleatórios). De forma geral, elas são mais recomendadas para a predição de variáveis categóricas e, a menos que dados sequenciais históricos estejam disponíveis, também não são muito boas com dados em série. Para evitar um crescimento exacerbado, o número de observações empregado nas divisões foi de 50. Destaca-se também que, em modelos de árvores, valores faltantes não são considerados aceitáveis.

Para critério de seleção do melhor modelo, adotou-se a definição do quadrado dos erros médios quadráticos ou RASE, dado por:

$$\mathrm{RASE} = \sqrt{\mathrm{ASE}} = \sqrt{\mathrm{SSE}/n} \tag{4.1}$$

No qual SSE representa a soma dos erros ao quadrado e n o número de observações.

A variável resposta de saída esperado para cada um dos métodos empregados definiu-se como o consumo de eletricidade total em uma semana em kWh. No contexto, o tipo de residência, suas características e tipo de utensílio elétrico foram considerados fatores de influência no consumo de eletricidade. Os resultados para cada algoritmo são mostrados na tabela 4.1 e os fatores mais significativos para o consumo de eletricidade por método na tabela 4.2:

Tabela 4.1. Comparação dos erros RASE (kWh) associados a cada método

	Decision tree	**Neural Network**	**Regression (full)**	**Regression (intercept)**
Summer	39.363	39.527	39.627	46.300
Winter	44.397	44.142	44.973	52.096

Fonte: Adaptado de TSO e YAU (2007)

Tabela 4.2. Fatores significativos na demanda energética por método

	Summer			Winter		
	Decision tree	Neural network	Regression (stepwise)	Decision tree	Neural network	Regression (stepwise)
HOS					*	
PD				*		*
VH				*		*
AGE						
SIZE	*	*	*	*		
RENT						
INCOME				*		
MEMBER	*	*	*	*	*	*
AC	*	*	*			
FAN						
CD		*	*			
CW		*				
DEH						
EWH				*	*	*
EK						
RH		*	*	*	*	*
VFAN			*		*	

HOS	Government subsidized home ownership scheme
PD	Private development
VH	Village house
Household characteristics:	
AGE	Age of the flat in years
SIZE	Size of the flat in ft^2
RENT	A dummy variable to indicate rental flat
INCOME	Monthly household income in HK$1000
MEMBER	Number of household members
Appliance ownership:	
AC	Ownership of air-conditioner
FAN	Ownership of fan
CD	Ownership of clothes dryer
CW	Ownership of clothes washing machine
DEH	Ownership of dehumidifier
EWH	Ownership of electric water heater
EK	Ownership of electric kettle
RH	Ownership of rangehood
VFAN	Ownership of ventilation fan

Fonte: Adaptado de TSO e YAU (2007)

A conclusão do estudo, segundo TSO e YAU (2007) foi que, durante o período de verão, as árvores de decisão mostraram alguns fatores importantes que, em essência, relacionavam-se ao uso do ar-condicionado. Quando a influência desse aparelho desaparecia no inverno, o modelo de árvore mostrava outros fatores importantes. De forma geral, os 3 modelos conseguiram demonstrar os fatores mais importantes no consumo de energia em cada estação do ano. Com a análise dos valores de RASE, percebeu-se que os 3 modelos são compatíveis para a predição, sendo que no verão, as árvores de decisão conseguiram uma performance um pouco superior que os outros 2 modelos, enquanto no inverno, as redes neurais mostraram uma performance um pouco superior do que os 2 métodos concorrentes.

4.2 PREDIÇÃO DA CINÉTICA DE DISSOLUÇÃO DE VIDRO DE SÍLICA

Para melhor compreender os resultados do estudo, KRISHNAN et al. (2018) utilizou 3 algoritmos de ML cujas características e forma de uso são descritas da seguinte forma:

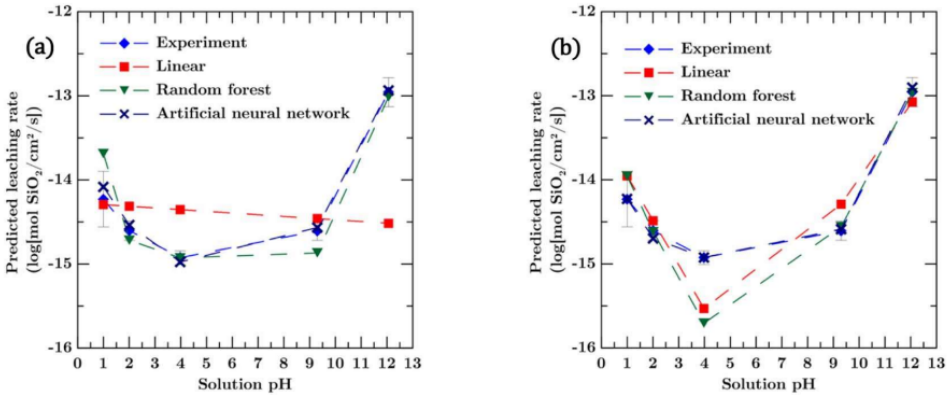
●Regressão Linear Simples e Regressão Lasso: tendo a técnica já sendo descrita, trata-se do ajuste do modelo de reta, sendo seus parâmetros determinados pelo ajuste dos mínimos quadrados, ou seja, pela minimização da soma residual dos erros quadrados. No caso da regressão Lasso, sua aplicação ocorre com o intuito de se diminuir a variância ao se introduzir um pequeno viés, diminuindo-se o peso de variáveis preditivas pouco significativas para o resultado do problema.

●Floresta Aleatória: trata-se, como já elucidado, de um conjunto de árvores preditoras. O algoritmo foi empregado no estudo utilizando-se da técnica de amostragem bootstrap, gerando-se *n* amostras bootstrap a partir dos dados de treino e criando-se uma árvore de decisão de cada amostra gerada para, posteriormente, selecionar a melhor divisão entre os dados. A predição final se dá pela média dos dados agregados das *n* árvores geradas.

●Redes Neurais: Para o estudo, utilizou-se como função de ativação a função sigmóide , dado o conhecimento de não-linearidade da relação composição-dissolução. Os dados foram divididos em um conjunto de treino (55%), de validação (15%) e de teste (30%). A rede foi primeiramente treinada com os dados de treino para ajuste dos pesos sinápticos. Os dados de validação foram utilizados para otimizar os pesos e evitar o *overfitting* e finalmente, a eficiência do modelo foi medida utilizando-se os dados de teste.

KRISHNAN et al. (2018) menciona que, a princípio, todas as predições foram realizadas com base em modelos movidos puramente por dados, ou seja, não contendo nenhuma informação referente ao mecanismo físico-químico de dissolução. Para fins de comparação e, tendo-se conhecimento de que a taxa de dissolução do silicato possui uma relação levemente linear para com a composição do solvente (tendência de decrescimento para meios ácidos e crescimento em meios alcalinos), dividiu-se o conjunto de dados em duas situações. Em um primeiro conjunto, o estudo é conduzido puramente a base dos dados e, no segundo, fez-se a introdução de informações experimentais do comportamento da dissolução em meio ácido e básico, e então, os procedimentos com os algoritmos foram realizados da mesma forma que para o primeiro conjunto.

Figura 4.1. Predição da taxa de dissolução do silicato em modelo totalmente a base de dados e com introdução de informação físico-química.



Fonte: Adaptado de KRISHNAN et al. (2018)

As conclusões do estudo elucidas por KRISHNAN et al. (2018) foi que, dos 3 algoritmos, nota-se que todos obtêm melhor performance quando apresentados os dados sobre a acidez e alcalinidade do meio, como mostrado pela figura 4.1. Ainda assim, as redes neurais foram as que apresentaram melhor desempenho geral. Isso ocorre dado o conhecimento prévio de que a relação entre as variáveis de entrada e saída possuíam uma relação não-linear. Embora a floresta aleatória tenha apresentado resultado satisfatório, seu poder de extrapolação se torna baixo quando os dados introduzidos no teste fogem dos limites dos dados introduzidos para treino. As árvores empregadas possuíam como variáveis de saída intervalos valores numéricos discretos, logo, para se obter um maior poder de acuracidade na

extrapolação, os dados teriam que ser exaustivamente medidos, de forma a incluir uma range muito maior de intervalos de valores numéricos. Por outro lado, as redes neurais apresentam um alto poder de extrapolação quando as variáveis de entrada e saída pertencem à mesma classe de valores.

4.3 PREDIÇÃO DA CONTAMINAÇÃO POR FLÚOR EM LENÇÓIS FREÁTICOS

Para seleção prévia das variáveis de entrada, utilizou-se o teste do qui-quadrado descrito como:

$$X_c^2 = \sum \frac{(Oi - Ei)^2}{Ei} \tag{4.2}$$

No qual *O* representa as os valores observados, *E* representa os valores esperados e *C* representa ao número máximo de variáveis independentes com liberdade para variar no conjunto de amostragem. Os resultados obtidos estão descritos na tebela 4.3:

Tabela 4.3. Seleção das entradas utilizando-se o teste do chi-quadrado.

Variáveis	TDS	Na^+	HCO_3^-	NO_3^-	SO_4^{2-}	Cl^-	Ca^{2+}	Mg^{2+}	K^+	Zn
x_c^2	20668.6	8967.3	8515.7	5226.4	2132.2	1583.5	1001.9	459.9	59.2	7.3

Fonte: Adaptado de NAFOUANTI et al. (2021)

As técnicas de ML empregadas são descritas por NAFOUANTI et al. (2021) a seguir:

●Floresta aleatória: Para o estudo, utilizou-se o modelo de floresta de classificação para se determinar a contaminação de flúor na água. O modelo combina várias árvores de decisão para se criar um modelo mais robusto e de maior acuracidade, evitando-se o overfitting. No estudo, 100 árvores foram geradas para o modelo.

●Redes neurais: A arquitetura escolhida para o estudo foi o MLP, tendo-se 10 neurônios na camada de entrada (número de variáveis selecionadas pelo teste do

chi-quadrado), 2 camadas ocultas e uma camada de saída. A função de ativação utilizada nas camadas ocultas foi definida como:

$$f(x) = \max(0, x =)f(x) = \begin{cases} xi\ if\ xi > 0 \\ 0, if\ xi < 0 \end{cases} \tag{4.3}$$

Na camada de saída, a função de ativação depende da predição do modelo. Para a análise, utilizou-se a função sigmóide.

●Regressão logística: esse método é empregado principalmente para classificações binárias. É uma conversão da regressão linear adaptada à função sigmóide.

Para se avaliar os modelos, utilizou-se como critérios a acuracidade, sensibilidade , especificidade e uma taxa de erro. Definiu-se a sensitividade como o percentual correto de classificação da concentração de flúor, enquanto a não-contaminação de flúor classificada corretamente foi denominada especificidade. Essas relações são explicitadas a seguir:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.6}$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN} \tag{4.7}$$

Sendo TP os positivos verdadeiros, TN os negativos verdadeiros, FP os falsos positivos e FN os falsos negativos. Os resultados obtidos estão expressos na tabela 4.4.

Tabela 4.4. Acuracidade, sensibilidade e erro para os modelos de ML

	RF	ANN	LR
Accuracy	0.89	0.85	0.76
Sensitivity	0.98	0.89	0.82
Error rate	0.10	0.14	0.23

Adaptado de NAFOUANTI et al. (2021)

Segundo NAFOUANTI et al. (2021), pelos resultados obtidos, percebe-se que o modelo de floresta aleatória para se prever a contaminação de flúor trouxe resultados satisfatórios com alta sensibilidade e especificidade, ou seja, com menor erro na predição, superando as redes neurais e a regressão logística. A melhor acuracidade pode ser justificada pela pré-seleção de variáveis empregando-se o teste do chi-quadrado. Outro fator que justifica a boa performance do método de floresta é que essa abordagem não considera com facilidade um único fator de entrada, sendo necessárias várias amostragens aleatórias de várias variáveis de entrada utilizadas para a construção de cada árvore, evitando-se problemas de overfitting e melhorando a acurácia. A performance intermediária as ANNs pode ser explicada por um poder menor de extrapolação, nesse caso, para além dos dados de treino que foram utilizados, o que, possivelmente, contribuiu para um overfitting. Por fim, a regressão logística demonstrou a pior performance, não tendo conseguido lidar bem com dados multidimensionais na fase de treinamento e, portanto, perdendo em acurácia na fase de teste.

5. CONCLUSÃO

Os estudos sobre ML aplicados a campos altamente dependentes de experimentação como a engenharia química mostraram-se, a partir do levantamento bibliográfico realizado, promissores, porém com certas limitações. Percebe-se, por exemplo, que a junção de informações acerca de parâmetros físico-químicos muito contribuem para a acurácia das respostas e diminuição dos erros associados quando comparados a modelos puramente movido a dados. Por outro lado, modelos com embasamento principal em ML mostram-se, por vezes, como única alternativa viável em problemas no qual as propriedades de um determinado fenômeno são de difícil compreensão. Com o tempo, a perspectiva é que dados melhor estruturados e disponíveis para uso possam aperfeiçoar as técnicas de modelagem usando ML. Constatou-se também que o conhecimento acerca da natureza da função estatística procurada muito influenciam a performance dos algoritmos, sendo, de forma geral, as Redes Neurais e os Métodos de Árvores mais apropriados para modelagens não lineares, a Regressão Logística mais apropriada para classificações binárias e a Regressão Linear, embora bastante tradicional, ainda muito válida e com alta acurácia quando se tem conhecimento sobre a relação entre as variáveis preditoras e respostas. Assim como em todo problema estatístico, a junção do conhecimento do problema, o uso crítico das técnicas de amostragem e uma boa seleção de modelo que encontre um bom equilíbrio viés-variância muito contribuem para a qualidades dos resultados buscados, sendo, portanto, indispensável que esses três fatores caminhem juntos.

6. REFERÊNCIAS

ALMEIDA, G. **Redes neurais artificiais: Princípios básicos artificial.** Revista Eletrônica Científica Inovação e Tecnologia, v. 1, n. 13, p. 47-57, 2016.

Aprendizado com a descida do gradiente. Deep learning, 2022. Disponível em: https://drive.google.com/file/d/1jXnqGRuXxN5rAPrlCbzIsMCCvi3s_pPB/view. Acesso em: 10 agost. 2022.

DING, S.,LI, H., SU, C., YU, J., JIN, F. **Evolutionary artificial neural networks: a review.** Artif Intell Rev, China, n. 39, p. 251-260, 2013.

JAMES, G., WITTEN D., HASTIE T., TIBSHIRANI, R. **An Introduction to Statistical Learning.** London: Springer, 2013.

KRISHNAN, N.M.A., MANGALATHU, S., SMEDSKJAER, M. M., TANDIA, A., BURTON, H., BAUCHY, M. **Predicting the dissolution kinetics of silicate glasses using machine learning**. Journal of Non-Crystalline Solids, n. 487, p. 37-45, 2018.

MARTINS, D. Um **tutorial completo sobre modelagem baseada em árvores de decisão (códigos R e Ryton**. Vooo-Insights, 2016. Disponível em: https://drive.google.com/file/d/1-YzyojAA7P6O8qvGWnLAZ4XB56mUHTBr/view. Acesso em 2 agosto. 2022.

MITCHELL, T. **Árvores de decisão. Sistemas inteligentes**. Disponível em: https://drive.google.com/file/d/1Ibnz4bq_4j0n30CszcuCKb3d4FfLfOs9/view. Acesso em 5 agost. 2022.

NAFOUANTI, M. B., LI, J., MUSTAPHA, N., A., UWAMUNGU, P., AL-ALIMI, D. **Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: Comparison of random forest, logistic regression and artificial neural network**. Applied Geochemistry, n. 132, 2021.

Random forest. IBM cloud, 2020. Disponível em: https://drive.google.com/file/d/1t8CUAB_DB_39UbQI8ViCGlOk35tGniWe/view. Acesso em agost 9. 2022.

Recurrent neural networks. IBM cloud, 2020. Disponível em: https://drive.google.com/file/d/1plzpUnzkGWfP8v8Be3sLvRp1Sog7hVkl/view. Acesso em: 8 agost. 2022.

Redes neurais. IBM cloud, 2020. Disponível em: https://drive.google.com/file/d/1VHH8mzdCaZxfD4dHZLGoGLEsNL_tfFgZ/view. Acesso em 9 agost. 2022.

SACRAMENTO, DANIEL. **Árvore de decisão: entenda este algoritmo de machine learning.** Carreiras em dados, 2021**. Disponível em:** https://drive.google.com/file/d/1RZ49nndv7T1A6GJBRhWv3syqhdogiaqX/view. Acesso em: 01 set. 2022.

SCHWEIDTMANN, A. M., ESCHE E., FISCHER, A., KLOFT, M., REPKE, S. S., MITSOS, A. **Machine Learning in Chemical Engineering: A perspective**. Chemie Ingenieur Technik, n. 12, p. 2029-2039, 2021.

Sigmoid function. Wikpedia, 2022. Disponível em: https://drive.google.com/file/d/1kfUa9jyNhg3hTjrKZ2FKyMJESVq_z6Nh/view. Acesso em: 10 agost. 2022.

THEBELT, A., WIEBE, J., KRONQVIST, J., TSAY, C., MISENER, R. **Maximizing information from chemical engineering data sets: Applications to machine learning**. Chemical Engineering Science, London, 252, 2022.

TSO, G.K.F., YAU, K.K.W. **Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks**. Energy, n. 32, p. 1761-1768, 2007.