UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
DEPARTAMENTO DE COMPUTAÇÃO CIENTÍFICA

Danilo Santos Vieira

UM PROBLEMA DE CLASSIFICAÇÃO SUPERVISIONADA PARA DETERMINAÇÃO DA PRESENÇA DE LIGAÇÃO QUÍMICA NA RELAÇÃO EURÓPIO-NITROGÊNIO E EURÓPIO-OXIGÊNIO

JOÃO PESSOA
2020

DANILO SANTOS VIEIRA

UM PROBLEMA DE CLASSIFICAÇÃO SUPERVISIONADA PARA DETERMINAÇÃO DA PRESENÇA DE LIGAÇÃO QUÍMICA NA RELAÇÃO EURÓPIO-NITROGÊNIO E EURÓPIO-OXIGÊNIO

Trabalho de Conclusão de Curso submetido à Universidade Federal da Paraíba, como requisito necessário para obtenção do grau de Bacharel em Matemática Computacional.

João Pessoa, 14 de Dezembro de 2020

UNIVERSIDADE FEDERAL DA PARAÍBA

DANILO SANTOS VIEIRA

Esta Monografia foi julgada adequada para a obtenção do título de Bacharel em Matemática Computacional, sendo aprovada em sua forma final pela banca examinadora:

Prof. Dra. Andréa Vanessa Rocha
Universidade Federal da Paraíba - UFPB
(Orientadora)

Prof. Dr. Jose Miguel Aroztegui Massera
Universidade Federal da Paraíba - UFPB
(Membro da banca)

Prof. Dr. Sérgio de Carvalho Bezerra
Universidade Federal da Paraíba - UFPB
(Membro da banca)

João Pessoa, 14 de Dezembro de 2020

Agradecimentos

Aos amigos e familiares, por todo incentivo e ajuda para não desanimar nessa jornada.

Agradeço a minha orientadora Andrea Vanessa Rocha por aceitar conduzir o meu trabalho de pesquisa.

Aos professores, por suas correções e ensinamentos que contribuiram para moldar a minha formação.

Resumo

Este trabalho descreve um estudo realizado em relação a um problema de classificação, mais especificamente, a partir de um conjunto de dados, tem-se a tarefa de visualizar através da estatística se um elemento químico denominado Európio possui uma ligação química ou não com o elemento químico oxigênio, e posteriormente o estudo é reaplicado para determinar se o Európio possui uma ligação com o Nitrogênio ou não. Para esse estudo, buscou-se compreender diversos conhecimentos correlacionados para compreender todo o desenvolvimento matemático computacional do problema como, por exemplo, à análise de dados e a técnica de classificação de Regressão Logística presente no contexto da aprendizagem de máquina supervisionada. Vale ressaltar, que o seu desenvolvimento deve-se a junção da aplicações de técnicas de tratamento ou pré-processamento de dados, dados de treinamento, criação de um modelo preditivo de regressão logística em relação ao conjunto de dados, e por fim a aplicação de técnicas para avaliar a precisão do modelo gerado.

Palavras-chave: Aprendizado de Máquina Supervisionado. Análise de dados. Regressão Logística.

Abstract

This work describes a study carried out in relation to a classification problem, more specifically, from a set of data, it has the task of visualizing through the statistics if a chemical element called Europium has a chemical bond or not with the element chemical oxygen, and afterwards the study is reapplied to determine if the Europium has a bond with Nitrogen or not. For this study, we sought to understand several correlated knowledge to understand all the computational mathematical development of the problem, such as, for example, data analysis and the Logistic Regression classification technique present in the context of supervised machine learning. It is worth mentioning that its development is due to the combination of applications of data treatment or preprocessing techniques, training data, creation of a predictive model of logistic regression in relation to the data set, and finally the application of techniques to assess the accuracy of the generated model.

Keywords: Supervised Machine Learning. Data analysis. Logistic Regression.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM - Aprendizado de Máquina

AMS - Aprendizado de Máquina Supervisionado

DG - Developers Google

Eu-N - Európio-Nitrogênio

Eu-O - Európio-Oxigênio

FEUP - Faculdade de Engenharia da Universidade de Porto

RL - Regressão Logística

Sumário

1 Introdução

O problema de classificação supervisionada consiste em utilizar um banco de dados onde as observações deste banco de dados estão distribuídos em diferentes classes, cada observação podendo pertencer a apenas uma delas. O problema de classificação possui inúmeras aplicações no mundo real. Neste trabalho estaremos focados no problema de classificação supervisionada binária, isto é, onde temos apenas duas classes disponíveis. Para mais detalhes e exemplos ver James et al. (2013, cap.4).

1.1 Contextualização

A partir do estudo de (Zhang et.al, 2013), que trata das características das ligações químicas de complexos de lantanídeos em um caso de estudo de valência, foi observado que existe uma correlação linear entre os parâmetros de valência de ligação para ligações entre um determinado cátion e dois ânions diferentes da mesma maneira. Em outras palavras, essa correlação linear indica que as variáveis independentes (cátion-ânion) estão associadas uma com a outra. Vale ressaltar que foi observado a relação $Ln - N$ (relação latanídeos - Nitrogênio) e o parâmetro de valência da ligação R_{Ln-N} em termos de correlação. O cátion é um íon com carga positiva, ou seja, é um átomo que perde elétrons. Enquanto o ânion é um ametal que pode possuir uma ligação iônica, ou seja, ter ligação com metais por ter alta eletronegatividade e é um átomo que ganha elétrons.

De acordo com o pesquisador, os parâmetros de valência da ligação diminuem com o aumento do número atómico de Ln e aumentam na mesma taxa com que os números de coordenação de Ln, o que está correto tendo em vista a diminuição dos raios iônicos e raios atômicos. Além disso, essa correlação indicou que os parâmetros de valência indisponíveis poderiam ser obtidos por interpolação linear para os parâmetros de ligação de valência de complexos orgânicos metálicos em química de coordenação. Uma vez que o parâmetro é conhecido para ligações com um ânion, isso pode ser reaplicado para encontrar esse parâmetro para uma relação com outro ânion, consequentemente conhecendo os parâmetros a_{jk} e b_{jk} da equação linear (1.1) apresentada seguir.

$$R_{ij} = a_{jk} + b_{jk}R_{ik} \tag{1.1}$$

$$d_{Ln-N} = R_{Ln-N} - b\ln\left(\frac{V_{Ln}}{CN_{Ln}}\right) \tag{1.2}$$

- R_{Ln-N} origina-se da equação (1.1), onde R_{ik} é o parâmetro de ligação de valência para um cátion, enquanto que R_{jk} é para diferentes ânions a_{jk} (superior) e $b_{jk}(inferior)$.

- d_{Ln-N} indica a distância entre lantanídeo e o Nitrogênio.
- V_i é a valência total do átomo ou estado de oxidação de i.
- CN_{Ln} é o número de coordenação para complexos de metal orgânico.

A partir da equação (1.1) é possível estudar outras relações possíveis entre latanídeos e algum ânion como, por exemplo, Ln-P e Ln-Br.

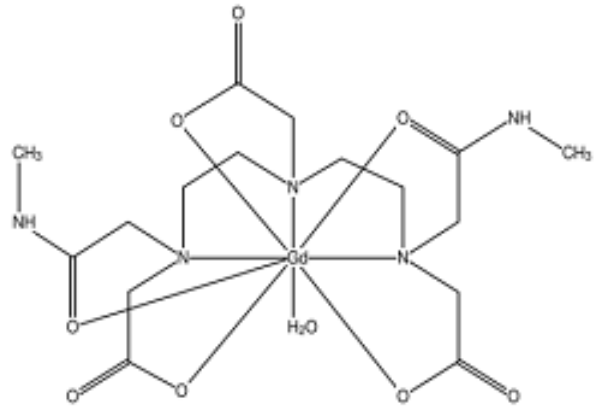
Sobre a importância da aplicação do estudo, existe uma classe de compostos químicos muito usada em lasers, análises clínicas tipo imunoensaios por fluorescência, em telas de celulares, em telas de televisores QLEDs, em dispositivos luminescentes (que trabalham por emissão de luz por matéria fria), ou como agentes de contraste em exames diagnósticos por ressonância magnética, dentre várias de suas inúmeras aplicações (CLAUDE et al, 2015). Alguns destes compostos são até mesmo usados como qubits em computação quântica (MACALUS et al, 2020). Estes compostos são os compostos de coordenação de ligantes orgânicos com íons de metais lantanóides, também conhecidos como complexos de íons lantanóides.

Lantanóides são quinze elementos da tabela periódica que vão do lantânio ao lutécio. Por exemplo, íons de lantânio se ligam a um compostos ligante. Para que aconteça a ligação deve existir oxigênio ou nitrogênio no composto ligante. O tipo de ligação presente no composto ligante é chamado de ligação de coordenação, um tipo de ligação química bem distinto das mais comuns e conhecidas, como as ligações do tipo covalente e iônica. A Figura 1 apresenta a representação de um composto químico inorgânico onde pode-se ver o metal lantanóide central, no caso o gadolínio, representado pelo seu símbolo químico, Gd. As ligações de coordenação são representadas nesta figura pelas linhas que conectam o Gd diretamente aos átomos de O ou N dos ligantes orgânicos. Quando o átomo de O, ou de N, não está coordenado, não há uma linha conectando diretamente o mesmo ao átomo central de gadolínio.

Para tornar a aplicação dos compostos de coordenação dos íons lantanóides eficaz, é de fundamental importância conhecer em detalhes a sua estrutura química, em particular a posição exata de cada átomo no espaço. A principal técnica para isto é a difração de raios-X de cristais destes compostos, a partir da qual pode-se obter as coordenadas de cada átomo pertencente à estrutura. De posse destas coordenadas, obtém-se então todas as distâncias interatômicas possíveis do composto.

É relativamente fácil identificar quais são as ligações de coordenação que estão presentes com base apenas na estrutura química e conhecimento sobre valências. Além disso, como pode-se depreender da Figura 1, distâncias interatômicas correspondentes a ligações de coordenação são geralmente mais curtas que as outras; porém nem sempre é assim.

Figura 1 – Representação esquemática da estrutura do composto aqua de gadolínio.



O grupo de pesquisas liderado pelo Prof. Alfredo Simas, da Universidade Federal de Pernambuco, tem publicado vários artigos sobre a modelagem destes complexos de lantanídeos por Química Quântica, em especial no desenvolvimento dos modelos Sparkle (DUTRA et al, 2013) e RM1 (DUTRA et al, 2014) . Estes modelos conseguem prever as geometrias destes complexos com exatidão semelhante à obtida por cristalografia de Raios-X. Para diversas aplicações, como por exemplo, a previsão da luminescência pelo modelo implementado no software LUMPAC do Prof. Ricardo Freire, da Universidade Federal de Sergipe, que utiliza intensamente os modelos Sparkle e RM1, seria muito útil e importante conseguir determinar se uma determinada distância interatômica corresponde de fato a uma ligação de coordenação ou não é muito útil para diversas aplicações, como por exemplo, a previsão da luminescência (FILHO et al, 2013; SILVA et al, 2018).

Em função disto, buscamos um modelo estatístico que possa identificar quando uma distância interatômica Európio-Oxigênio (Eu-O) ou uma distância interatômica Európio-Nitrogênio (Eu-N) corresponde a uma ligação de coordenação, ou quando não corresponde. Todos os quinze elementos lantanóides se comportam de maneira muito semelhante no que diz respeito às distâncias, podendo as mesmas serem tratadas como um grupo homogêneo. Foi então preparada uma lista destas distâncias presentes em dezenas de compostos de vários metais lantanóides, com indicação, em cada uma, da presença ou ausência de uma ligação de coordenação. Todas as distâncias da lista estão apresentadas na unidade de distância Angstrom, Å, onde 1 Å é igual a 10^{-10}m. O objetivo do trabalho é desenvolver um modelo que possa fazer esta identificação com elevado grau de acerto.

1.2 Definição do Problema

Nesse trabalho, busca-se encontrar um modelo de classificação que consiga inferir, em um conjunto de dados, a existência de ligação de coordenação entre o íon de Európio e

o Oxigênio ou o Nitrogênio. O conjunto de dados contem as distâncias entre o Európio e o Oxigênio ou Nitrogênio assim como uma indicação da existência desta ligação. A indicação "0"significa a ausência de ligação, caso contrário, se a indicação for "1"isso significa que existe ligação.

1.3 Objetivos Gerais

Este trabalho tem como objetivo geral pesquisar e desenvolver a parte matemática e computacional dos métodos de pré-processamento de dados, aprendizado de máquina supervisionado e Regressão Logística com o intuito de determinar um modelo capaz de decidir se há uma ligação química entre átomos de oxigênio ou nitrogênio e um íon metálico (Európio) a partir de suas respectivas distâncias.

1.4 Objetivos Específicos

Como objetivos específicos, pretende-se:

- Buscar informações correlacionadas ao problema na literatura.
- Estudo matemático e estatístico de uma das técnicas de classificação e de Aprendizado de Máquina Supervisionada.
- Determinar o modelo de classificação a ser utilizado e verificar a eficiência do modelo escolhido a partir da medida de acurácia.

2 *Referencial Teórico*

A análise de dados é o processo de examinar dados com o auxílio da matemática computacional, da estatística e da Inteligência Artificial. Além disso, da sua aplicação, busca-se extrair informações e padrões que auxiliem na tomada de decisões.

A análise de dados na ciência moderna envolve a aplicação de várias técnicas estatísticas, como correlação, regressão, análise de variância e testes de qui-quadrado. Essas técnicas fornecem uma maneira de fazer inferências indutivas de dados e distinguir quaisquer fenômenos ou efeitos reais de flutuações aleatórias presentes nos dados. Vale ressaltar que a inferência sobre esses dados deve ser de natureza imparcial (SHAMOO e RESNIK, 2009). Ainda sobre os dados, é de suma importância a organização e preparação dos dados antes de executar uma técnica de análise dados. Dessa forma, no próximo tópico aborda-se essa questão.

2.1 Organização e Preparação dos dados

Como os dados não necessariamente terão a natureza adequada para a aplicação das técnicas de análise de dados, é necessário realizar um tratamento de dados. A seguir, apresenta-se três técnicas utilizadas para preparar os dados, são elas: normalização, categorização e amostragem.

2.1.1 Normalização

As técnicas de normalização podem ser usadas para alterar uma variável contínua para cair dentro de um intervalo especificado, mantendo as diferenças relativas entre os valores para a variável. A abordagem mais simples para a normalização é a normalização de intervalo, que realiza uma escala linear dos valores originais da variável contínua em um determinado intervalo (KELLEHER et al, 2015). Segundo a Developers Google (2020), o objetivo da normalização é transformar variáveis em uma escala semelhante, consequentemente obtém-se um aperfeiçoamento do desempenho e estabilidade de treinamento do modelo. Além disso, a Developers Google (DG) mostra quatro técnicas de normalização comuns, a seguir apresenta-se uma visão geral sobre cada uma delas:

- Dimensionamento Linear: é utilizada quando os dados são distribuídos de maneira uniforme em um intervalo. Para utilizar essa técnica é necessário conhecer os limites superior e inferior dos dados e não haver outlier (DG, 2020), ou seja, espera-se não encontrar valores que sejam considerados diferenciados das outras observações. Vale ressaltar que outlier pode ser caracterizado como uma observação que desvia tanto

das outras observações que levanta suspeita de que foi gerada por um mecanismo diferenciado (HAWKINS, 1980). A formulação matemática para essa técnica é dada pela expressão abaixo:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

onde: x é um vetor de atributos dado.

- Recorte: é utilizada quando há outliers extremos no conjunto de dados. A formulação matemática para essa técnica é dada pela expressão abaixo:

- se x > max, então x' = max(x).
- se x < min, então x' = min(x).

- Escala de Log: é utilizada quando os dados estar em conformidade com a lei de potência, ou seja, a escala de registro é útil quando alguns dos seus valores têm muitos pontos, enquanto a maioria dos outros valores têm poucos pontos. Além disso, o escalonamento de log calcula o log de seus valores para compactar uma ampla faixa em uma faixa estreita (DG, 2020).

$$x' = log(x)$$

- Z-Score: é utilizada quando a distribuição das variáveis não contém outliers, onde há uma variação da escala que representa o número de desvios padrão da média (DG, 2020). A formulação matemática para essa técnica é dada pela expressão abaixo:

$$x' = \frac{x - \mu}{\sigma},$$

onde μ e σ são, respectivamente, a média (valor 0) e o desvio padrão (valor 1) amostrais.

2.1.2 Categorização

A categorização envolve a conversão de uma variável discreta ou contínua em uma variável categórica, isto é, uma variável que assume uma quantidade finita de valores. Um exemplo para essa definição seria o número de acertos de um aluno em uma determinada prova por categorias da seguinte forma: categoria 1: entre 0 e 5 acertos; categoria 2: entre 6 e 10 acertos e, categoria 3: acima de 10 acertos.

2.1.3 Amostragem

Uma vez que o conjunto de dados é consideravelmente grande, pega-se uma amostra com porcentagem pequena (tipicamente menor do que 10%) desse conjunto de

dados. Entretanto, é necessário que essa amostra seja representativa e que nenhum viés não intencional seja introduzido, ou seja, as distribuições de recursos no conjunto de dados amostrados não deve ser muito diferentes das distribuições de recursos do conjunto de dados original (KELLEHER et al., 2015).

Em outras palavras, a ideia intuitiva diretamente relacionada a técnica de amostragem é de pegar uma amostra pequena, mas que seja considerada representativa para um conjunto de dados original maior. E posteriormente, aprender alguma coisa por parte da amostra sobre o conjunto de dados original (MAYER, 2016).

2.1.4 Balanceamento dos dados

Essa técnica é utilizada quando há uma diferença considerável entre o número de objetos para diferentes classes, pois vários algoritmos de aprendizado de máquina (AM) tem o desempenho prejudicado pelos dados desbalanceados. Para a amostra ser considerada aceitável, a acurácia preditiva de um classificador para um conjunto de dados desbalanceados deve ser maior que a acurácia obtida atribuindo todo novo objeto à classe majoritária (FACELLI et al, 2011). Ainda segundo esse autor, as principais técnicas para solucionar esse problema são:

- Redefinir o tamanho do conjunto de dados reduzindo nos dados de classes maioritárias para a classe de menos dados;
- Utilizar diferentes custos de classificação para as classes;
- Induzir um modelo para uma classe, ou seja, calcular exemplos de classes minoritárias para igualar a de classes maioritárias.

Essas são algumas das principais técnicas utilizadas para a etapa de pré-processamento de dados. Abaixo apresenta-se outras duas técnicas de balanceamento de dados: undersampling e oversampling.

2.1.4.1 Undersampling

Essa técnica consiste em realizar uma "subamostragem"para uma classe maior do conjunto de dados original.

Exemplo 1: "Subamostragem"de um conjunto de dados desbalanceado

- Classe A tem 200 valores
- Classe B tem 90 valores

O algoritmo consiste em pegar uma amostra aleatória (normalmente sem reposição) com 90 observações classe A e compor um novo conjunto de dados com B, também com 90 valores, totalizando 180 observações.

2.1.4.2 Oversampling

Essa técnica consiste em realizar uma superamostragem para uma classe menor do conjunto de dados original.

Exemplo 2: "Superamostragem"de um conjunto de dados desbalanceado

- Classe A tem 200 valores
- Classe B tem 90 valores

O algoritmo consiste em pegar uma amostra aleatória (obrigatoriamente com reposição) com 200 observações aleatórias obtidas dos 90 valores da classe B e compor um novo conjunto de dados com A, também com 200 valores, totalizando 400 observações.

2.1.5 Visão Geral

Uma vez visto todas as técnicas de tratamento de dados, entra-se na etapa de utilizar esses dados ajustados e aplicar o método de classificação. Para isso, nos próximos tópicos são abordados o Aprendizado de Máquina Supervisionado (AMS) com ênfase na classificação.

2.2 Aprendizado de Máquina (AM)

Segundo Mitchell (1997), um programa de computador aprende com a experiência E com relação a alguma tarefa T e alguma medida de desempenho P, se seu desempenho em T, conforme medido por P, melhora com a experiência E e com o aumento da taxa P. Em outras palavras, um algoritmo de AM aprende a partir de um conjunto de dados de treinamento, onde busca-se uma hipótese dentro de um universo de possíveis hipóteses, capaz de descrever as relações entre os objetos e que melhor se ajusta aos dados de treinamento (FACELLI et al., 2011).

Exemplo 3: Resolução de questões de uma lista de exercícios sobre um determinado tema

- E = experiência com a resolução de questões sobre o tema

- T = a tarefa de resolver questões sobre o tema
- P = a probabilidade de acertar a resolução da questão

O exemplo anterior apresenta noções preliminares relacionadas ao algoritmo de AM. Uma outra visão para esse algoritmo é a busca por determinar padrões ou percepções a partir dos comportamentos observados a partir de um determinado conjunto de dados de treinamento Y de um conjunto original X. Nesse conjunto de dados observado, há um conjunto de rótulos associados que correspondem a classes ou valores obtidos por alguma função desconhecida. Desse modo, um algoritmo de classificação buscará produzir um classificador capaz de generalizar as informações contidas no conjunto de treinamento, com a finalidade de classificar, posteriormente, objetos cujo rótulo seja desconhecido (CARVALHO et al., 2017). Esse processo pode ser visualizado na figura abaixo:

Figura 2 – Processo de tomada de decisão a partir dos dados. Machine Learning for Predictive Data Analytics

Fonte: KELLEHER et al., 2015

A figura 2 descreve uma situação da tomada de decisão no qual tem como etapas preliminares, o conjunto de dados e consequentemente a análise desses dados para retornar uma ideia. Usando a representação computacional dos dados, as técnicas de AM podem gerar modelos capazes de reconhecer ou ainda imitar o comportamento de um especialista humano de modo automático, ou seja, essas técnicas podem ser abordada de forma tradicional (FILHO, 2017).

Primeiramente, as técnicas de AMS aprendem automaticamente um modelo da relação entre um conjunto de variáveis descritivas (também conhecidas como covariáveis) e uma variável de destino (também conhecida como variável resposta) com base em um conjunto de exemplos históricos ou instâncias. De acordo com Kelleher, pode-se construir os modelos para análise de dados preditivos, usando o AMS e consequentemente predizendo novas instâncias. Para cada observação i se tem uma resposta y_i. A observação é representada por um vetor $x^{(i)}$ com n entradas. Cada entrada representa uma covariável ou variável descritiva. Queremos ajustar um modelo que relacione a resposta aos preditores, com o objetivo de predizer com precisão a resposta para observações futuras (predição) ou melhor compreender a relação entre a resposta e os preditores (inferência). Muitos

métodos clássicos de aprendizagem estatística, como regressão linear, regressão logística (RL), Modelos Lineares Generalizados, Modelos Aditivos Generalizados, *gradient boosting* (de uma coleção de modelos, tipicamente árvores), máquinas de vetores de suporte, florestas aleatórias (random forests), entre outros., operam no domínio de aprendizagem supervisionada (JAMES et al., 2013). Em Aprendizado de Máquina supervisionada (AMS), os algoritmos são utilizados para induzir modelos preditivos por meio da observação de um conjunto de objetos rotulados (Facelli et al, 2011).

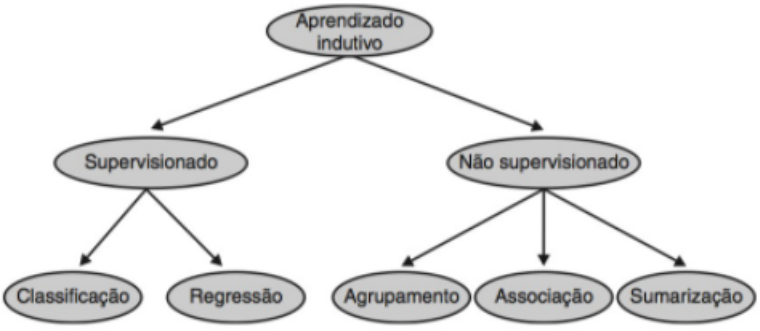
Em outra abordagem, segundo Filho (2017), a técnica de aprendizado de máquina não-supervisionado (AMNS) é feita com dados não rotulados, ou seja, não se sabe quantas classes ou quais classes existem, ou seja, é uma técnica usada para encontrar subgrupos de objetos. Além disso, a AMNS descreve uma situação um pouco mais desafiadora em que para cada observação $i = 1, ..., n$, observamos um vetor de medidas x_i, mas nenhuma resposta associada y_i. Nesse cenário, estamos de alguma forma trabalhando cegos, a situação é chamada de não supervisionada porque não temos uma variável de resposta que possa supervisionar nossa análise. (JAMES et al., 2013).

Essas duas abordagens retratam o paradigma de aprendizado a ser adotado para lidar com a tarefa, ou seja, podem ser de natureza preditiva (AMS) ou descritiva - AMNS (FACELLI et al., 2011).

> O Aprendizado Indutivo é a inferência de conhecimento a partir dos dados. A aprendizagem indutiva é o processo de construção de um modelo em que o ambiente é analisado, ou seja a base de dados. Na procura de tendências e padrões. Por exemplo, objetos com características similares são agrupados em classes e formuladas regras em que é possível prever a classe dos objetos que venham a ser analisados futuramente. Há que ter em atenção que o ambiente é dinâmico, logo o modelo deve ser adaptativo, isto é, deve ter a capacidade de aprender (FEUP, 2000).

Essas abordagens são ilustradas na figura a seguir:

Figura 3 – Hierarquia de aprendizado.



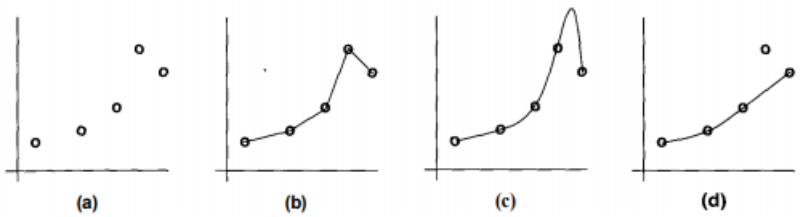
Fonte: FACELLI et al., 2011.

A figura 3 apresenta ambas abordagens e vale ressaltar que elas possuem áreas de estudo com objetivos específicos. Nos próximos tópicos desse trabalho serão explicados o AMS e posteriormente a classificação.

2.2.1 Aprendizado Supervisionado

O aprendizado supervisionado é o conjunto de algoritmos que buscam obter características presentes em um determinado conjunto de dados e, consequentemente obter um modelo que consiga estimar novos rótulos sobre novos dados (Bertozzo, 2019). Essas características podem ser chamadas de padrões, tendências ou até mesmo alguma informação que auxilie na tarefa de inferir e construir modelos preditivos em relação a esses dados. Do ponto de vista matemático, na AM, o elemento de aprendizagem recebe o valor correto (ou aproximadamente correto) da função para entradas específicas e altera sua representação da função para tentar corresponder às informações fornecidas pelo feedback. Mais formalmente, dizemos que um exemplo é um par $(x, f(x))$, onde x é a entrada e $f(x)$ é a saída da função aplicada a x. A tarefa da inferência indutiva pura é esta: dada uma coleção de exemplos de f, retorne uma função h que se aproxime de f. A função h é chamada de hipótese (RUSSEL e NORVIG, 1995).

Figura 4 – Em (a), temos alguns pares de exemplo (entrada, saída). Em (b), (c) e (d), temos três hipóteses para funções das quais esses exemplos podem ser extraídos.



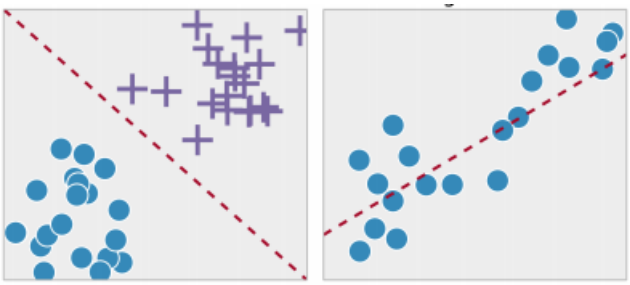
Fonte: RUSSEL e NORVIG, 1995.

A figura 4, trata de um exemplo envolvendo a geometria plana, de forma que a figura 4 (a) envolve pontos no plano (x, y), onde $y = f(x)$, e a tarefa é encontrar uma função $h(x)$ que se ajuste bem aos pontos (RUSSEL E NORVIG, 1995). Na Figura 4 (b), temos uma função h linear por partes, enquanto na Figura (c) temos uma função h mais complicada, mas difere da figura 4 (b) nos valores y que atribuem a outras entradas x, já 3 (d) é uma função que aparentemente ignora um dos pontos do exemplo, mas se ajusta aos outros com uma função simples (RUSSEL E NORVIG, 1995). Assim, x e $f(x)$ indicam a conexão entre a entrada e saída previstas. Além disso, a AM é composta por duas categorias: Classificação e Regressão.

2.2.2 Diferenças entre Classificação e Regressão

Essas categorias representam a generalidade de problemas no qual a mineração de dados é aplicada atualmente, nomeadamente através da criação de modelos, à custa de um conjunto de exemplos classificados, para prever para cada registo que faça parte de um outro conjunto de dados, ou seja, se pertencer a uma determinada classe, trata-se de um método de classificação, caso pertença a um valor, trata-se de regressão (FEUP - FACULDADE DE ENGENHARIA DA UNIVERSIDADE DE PORTO, 2000). A figura 5 abaixo ilustra as técnicas de classificação e regressão respectivamente.

Figura 5 – Classificação x Regressão.



Fonte: USP, 2020.

As variáveis presentes no modelo podem ser caracterizadas como quantitativas ou qualitativas. Variáveis quantitativas assumem valores numéricos. Em contraste, as variáveis qualitativas assumem valores em uma das K diferentes classes ou categorias. Tendemos a nos referir aos problemas com uma resposta quantitativa como problemas de regressão, enquanto aqueles que envolvem uma resposta qualitativa são freqüentemente chamados de problemas de classificação (JAMES et al., 2013).

Os métodos de classificação tratam de um abordagem ampla de aprendizado supervisionado que treina um programa para categorizar informações novas e não rotuladas com base em sua relevância para dados rotulados conhecidos (DEEPAI, 2020). A seguir, aborda-se uma das técnicas de classificação: a regressão logística.

É importante observar que na teoria da estatística, todos os modelos estatísticos aplicados em aprendizagem supervisionada são chamados de modelos de regressão, ou seja, na teoria de estatística não costuma-se utilizar o vocabulário de modelos de classificação. Na teoria estatística, a diferença ocorre apenas na distribuição da variável resposta. Assim, se a distribuição da variável resposta tiver suporte em uma quantidade finita de valores, pela ótica da ciência de dados, estamos lidando com um problema de classificação, caso contrário, pela ciência de dados estaremos lidando com um problema de regressão. Por fim, é importante observar que isso justifica o nome “regressão logística” para um modelo que é utilizado na ciência de dados como um modelo de classificação.

2.3 Regressão Logística (RL)

De acordo com (GONZALEZ, 2018), a RL é uma técnica estatística que tem como objetivo produzir a partir de um conjunto de observações um modelo que permita a predição de valores tomados por um variável categória (binária), em função de uma ou mais variáveis independentes binárias.

Esse método tem como objetivo realizar predições ou explicar a ocorrência de determinados fenômenos quando a variável dependente é de natureza binária. Às variáveis independentes podem ser quantitativas ou qualitativas. Além disso, mesmo quando a resposta de interesse não é originalmente da natureza esperada, busca-se a resposta de modo que a probabilidade de sucesso possa ser ajustada através da RL. Esse modelo expandiu-se por outros campos para modelar relacionamentos que envolvessem uma variável dependente dicotômica e um conjunto de variáveis dependentes. O Modelo de RL é adequado para estudar situações em que existe um conjunto de variáveis explicativas que se correlacionam com uma variável resposta dicotômica (SOUZA, 2013).

Uma das vantagens do método é a sua alta capacidade de processamento por computadores e pelo desenvolvimento de pacotes estatísticos. Essas vantagens ampliam a sua área de atuação em áreas como: economia, mineração, transportes, sensoriamento remoto, medicina e nas Ciências Sociais (Bezerra, 2012). Um dos princípios fundamentais para o método é converter as observações dessas variáveis em chance e submetê-las a uma transformação logarítmica. O modelo apresentado nessa seção será evidenciado a seguir com a Distribuição de Bernoulli.

2.3.1 Formulação Matemática

De acordo com Rigollet (2015), temos um problema de classificação no qual observamos pares ordenados $(X_1, Y_1), ...(X_n, Y_n)$ que são n cópias aleatórias independentes de $(X, Y) \in X \times \{0, 1\}$. Denote por $P_{X,Y}$ a distribuição conjunta de (X, Y). A variável X vive em algum espaço abstrato X (no nosso caso iremos considerar $X = \mathbb{R}^d$) e $Y \in \{0, 1\}$ é chamado de rótulo ou variável resposta. O objetivo da classificação binária é construir uma regra para prever Y dado X usando apenas os dados. Essa regra é uma função $h : X \to \{0, 1\}$ chamada de classificador. Vale ressaltar que, $Y \in \{0, 1\}$, então Y tem uma distribuição Bernoulli.

As variáveis do nosso modelo podem ser classificadas como ligado ao oxigênio ou não ligado ao oxigênio e, ou ligado ao nitrogênio ou não ligado ao nitrogênio, o que caracteriza a natureza dicotômica da variável aleatória Y e da distribuição de Bernoulli. As probabilidades podem ser descritas como:

$$P(Y = 1|X = x) = \pi(x) \tag{2.1}$$

e

$$P(Y = 0|X = x) = 1 - \pi(x) \tag{2.2}$$

Nesse estudo, $Y = 0$ designa que o európio está ligado ao oxigênio , já $Y=1$ indica que o európio não está ligado ao oxigênio. Segundo Bezerra (2012), esse modelo é baseado na função logística, portanto assume-se que $g(\mu)$ é a função logit, podemos representá-la da seguinte forma:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r = \eta \tag{2.3}$$

Onde:

x_i é o vetor de covariáveis explicativas, $\forall$ i = 1,..,r.

β_i é o coeficiente de estimação das variáveis, $\forall$ i = 0,..,r.

$\eta_i = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$ é a resposta média logit para i-ésima observação.

$\mu = P(Y = 1|X = x) = \frac{e^\eta}{1+e^\eta}$ é a probabilidade

De acordo com Bezerra (2012), esse modelo irá exprimir a relação entre a variável dependente e a as variáveis independentes. Para se atingir os valores referentes às probabilidades do evento temos

$$\mu = P(Y = 1|X = x) = \frac{e^\eta}{1 + e^\eta} \tag{2.4}$$

onde,

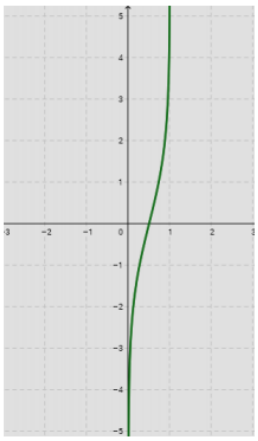
$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

O modelo apresentado permite que o pesquisador modele a probabilidade de um evento ocorrer dependendendo das variáveis categóricas, estime a probabilidade de um evento ocorrer ou não para uma observação aletória, prever o efeito do conjunto de variáveis sobre a variável dependente binária, classificar observações, estimando a probabilidade de uma observação estar em uma categoria determinada (GONZALEZ, 2018).

2.3.1.1 Função Logit

A função logit (figura 6) é uma função que mapeia uma combinação linear de variáveis (equação 2.6) e retornam um número no intervalo $(0, 1)$ (equação 2.4). Na realidade, a própria equação 2.4 nos indica que a função logit nos retorna a probabilidade de $Y = 1$ dado o valor observado de X.

Figura 6 – Gráfico da função logit(p).



Fonte: Cabral, 2020.

Essa transformação define-se como:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \tag{2.5}$$

Seja g(x) $= \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$, então iguala-se g(x) a equação logit.

$$\beta_0 + \beta_1 x_1 + ... + \beta_r x_r = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \tag{2.6}$$

logo,

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1+...+\beta_r x_r} \tag{2.7}$$

Isola-se p, assim obtemos a probabilidade estimada $\hat{p}$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_r x_r}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_r x_r+...+\beta_r x_r}} \tag{2.8}$$

Outro resultado importante sobre a função logit é a sua inversa, que é conhecida na ciência de dados como sigmóide (de forma genéreica), ou como logística (de forma de

mais precisa). Em resumo, graficamente é 90 graus invertido em relação ao gráfico da figura 5 e consequentemente a forma da figura fica semelhante a de um "S"(ou sigmóide), e o domínio da função passa de x para y.

Por fim, vale a pena mencionar que, na estatística, uma função g satisfazendo uma relação como a da equação (2.3) é conhecida como uma *função de ligação*. Já na ciência de dados, utiliza-se a inversa da função g, e esta inversa é conhecida na ciência de dados como *função de ativação*.

2.3.1.2 Estimação de Parâmetros

Segundo Gonzalez (2018), a estimação dos parâmetros, ou seja, $\beta = [\beta_0, \beta_1]$ e das probabilidades condicionais (2.1) e (2.2) será obtida através do Método de Estimação da Máxima Verossimilhança (equação 2.9). Vale ressaltar que a dedução da estimação dos parâmetros foi retirada de FIGUEIRA (2006).

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{2.9}$$

Aplica-se o logaritmo natural a ambos membros

$$l(\beta) = ln[L(\beta)] = \sum_{i=1}^{n}[y_i ln\pi(x_i) + (1 - y_i)ln(1 - \pi(x_i))] \tag{2.10}$$

Deriva-se $l(\beta)$ em relação a β_0 e β_1,

$$\hat{\beta_0} = \frac{\partial ln[L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \pi(x_i)] \tag{2.11}$$

$$\hat{\beta_1} = \frac{\partial ln[L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n}[x_i - \pi(x_i)] \tag{2.12}$$

Os estimadores $\hat{\beta_0}$ e $\hat{\beta_1}$ medem a taxa de variação do logit para uma unidade de variação na VI, ou seja, a inclinação da linha de regressão entre a VD y_i e a sua VI x_i (GONZALEZ, 2018).

2.3.1.3 Avaliação do modelo

Após a definição do modelo, é necessário testar a sua validade e também a sua precisão para um possível diagnóstico ou tomada de decisão. Esse procedimento permite

identificar as variáveis que não se ajustam ao modelo proposto ou que têm forte influência sobre a estimação dos parâmetros. Em RL há uma série de gráficos, testes de ajuste e outras medidas para assegurar a validade do modelo (GAMBARRA, 2012). No próximo, apresenta-se em detalhes o método que tem os fundamentos necessários de avaliar ou verificar a precisão do modelo gerado para o conjunto de dados fornecidos.

2.4 Precisão do modelo: Sensitividade, Especificidade e Acurácia

Segundo Borges (2016), um teste de diagnóstico (classificação) considerado perfeito tem o potencial de discriminar adequadamente o conjunto de dados. No contexto desse trabalho, os valores de um teste de classificação que são maiores ou iguais ao ponto de corte indicam a presença da ligação química, enquanto que valores abaixo do ponto de corte indicam a ausência de ligação química.

A sensitividade, especificidade e acurácia são apresentadas a seguir, vale ressaltar antes o significado dos parâmetros de interesse, ou seja,

- Verdadeiro Positivo (VP) - há a ligação química e o valor de um parâmetro de interesse maior ou igual ao ponto de corte.
- Verdadeiro Negativo (VN) - ausência de ligação química e o valor de um parâmetro de interesse maior ou igual ao ponto de corte.
- Falso Positivo (FP) - ausência de ligação química e o valor de um parâmetro de interesse abaixo do ponto de corte.
- Falso Negativo (FN) - há a ligação química e o valor de um parâmetro de interesse abaixo do ponto de corte.

Uma vez que conhecidos os parâmetros de interesse, na próxima relação apresenta-se os testes que medirão a precisão do método.

a) Sensitividade: É a probabilidade do teste sob investigação fornecer um resultado positivo dado que o elemento químico európio é ligado ao oxigênio e/ou nitrogênio:

$S_E = \frac{VP}{VP+FN}$

b) Especificidade: É a probabilidade do teste fornecer um resultado negativo, dado que o elemento químico európio é ligado ao oxigênio e/ou nitrogênio:

$E_S = \frac{VN}{FP+VN}$

c) Acurácia: É a proporção de previsões corretas (tanto para verdadeiro positivo quanto para verdadeiro negativo) entre todos os casos examinados:

$A_C = \frac{VP+VN}{VP+VN+FP+FN}$

3 Metodologia

Neste capítulo serão descritos as etapas, os materiais e os métodos utilizados para solucionar o problema de classificação apresentado. Em resumo as etapas consistem em: pré-processamento dos dados, aplicação do método de classsificação de Regressão Logística e avaliação do modelo gerado na etapa de classificação.

3.1 Material

Inicialmente, a base de dados encontra-se em uma planilha de extensão '.xlsx'. Em seguida, o pré-processamento de dados, método de Regressão Logística e a avaliação pela Sensibilidade, Especificidade e Acurácia foram implementados na linguagem R 4.0.2, utilizando as bibliotecas readxl, pROC, ROCR, dplyr. Todos as implementações foram realizadas e testadas em um computador com CPU AMD Ryzen^{TM} 5 2600 3.4 GHz, CPU AMD Radeon^{TM} RX 550 2GB VRAM e memória RAM de 8GB DDR4.

3.1.1 Bibliotecas

A seguir, apresenta-se cada biblioteca utilizada para o desenvolvimento do trabalho, vale ressaltar que dois comandos são essenciais, o primeiro consiste em baixar e instalar a biblioteca solicitada, enquanto que o segundo carrega a biblioteca em questão. Os comandos são respectivamente: install.packages("nome da biblioteca") e library("nome da biblioteca).

- readxl: Essa biblioteca é utilizada no R para a leitura (importação) dos dados contidos em uma planilha com a extensão 'xlsx'. De acordo com (WICKHAM, 2016), trata-se de uma biblioteca de fácil instalação, pois é possível utilizá-la em qualquer sistema operacional. Além disso, essa biblioteca é projetada para trabalhar com dados tabulares armazenados em uma única folha.
- pROC: É uma biblioteca que contém ferramentas que permitem a visualização e análise das curvas ROC geradas a partir do modelo. A área parcial sob a curva (AUC) pode ser comparada com testes estatísticos e os seus intervalos de confiança podem ser calculados para p curvas AUC ou ROC (ROBIN, 2020).
- ROCR: É uma biblioteca que também contém ferramentas que permitem a visualização e análise das curvas ROC geradas a partir do modelo. Nessa biblioteca é possível visualizar medidas de desempenho como, por exemplo, taxa de erro, AUC, erro de calibração, recall, entre outros (ERNST, 2020).
- dplyr: é uma biblioteca voltada para a manipulação de dados.

3.1.2 Conjunto de Dados

Nesse trabalho foi utilizada uma base de dados da área de química, onde trata-se de um problema de classificação que envolve um elemento químico denominado európio. O európio pode estar ligado ao oxigênio ou não ligado. Além disso, também extende-se a aplicação da técnica para identificar se o európio está ligado ao nitrogênio ou não ligado. Ou seja, aplica-se a técnica de classificação de regressão logística duas vezes, para a relação Európio-Oxigênio e em seguida para a relação Európio-Nitrogênio.

Essa base de dados contém um total de 1820 valores numéricos que indicam as relações citadas anteriormente. A princípio, temos a seguinte divisão:

- Európio - Oxigênio (0 - não ligado): 870 dados numéricos (undersampling utiliza 458 valores aleatórios do conjunto de dados original sem reposição).
- Európio - Oxigênio (1 - ligado): 458 dados numéricos (oversampling utiliza 458 valores aleatórios do conjunto de dados original com reposição).
- Európio - Nitrogênio (0 - não ligado): 421 dados numéricos (undersampling utiliza 71 valores aleatórios do conjunto de dados original sem reposição).
- Európio - Nitrogênio (1 - ligado): 71 dados numéricos (oversampling utiliza 421 valores aleatórios do conjunto de dados original com reposição).

Como o problema apresentado é de aprendizagen supervisionada, mais especificamente de classificação, após a visualização do conjunto de dados foram necessárias efetuar as etapas de pré-processamento de dados com o intuito de efetuar o balanceamento de dados e preparar os dados.

3.1.3 Pré-processamento dos dados

Como os dados fornecidos estão desbalanceados em relação ao tamanho, foi necessário realizar um tratamento de pré-processamento dos dados. Esse tratamento consistiu em aplicar as técnicas de undersampling e oversampling, com isso, tinha-se um conjunto de dados novo e balanceado, e posteriormente realizar a seguinte tarefa: utiliza-se uma amostra aleatória do conjunto de dados novo de 80% separadamente dos valores gerados por cada umas técnicas como conjunto de treinamento e os outros 20% fica como conjunto de teste que posteriormente é utilizado para validação do modelo através da sensitividade, especificidade e acurácia do modelo gerado.

3.1.4 Construção do modelo e sua avaliação

Após a etapa do pré-processamento dos dados, aplica-se o método de Regressão Logística no novo conjunto de dados organizado e os dados de treinamento. Posteriormente,

aplica-se o teste da Acurácia para obter a precisão do modelo preditor. A seguir, apresenta-se a definição dos principais comandos utilizados nessa etapa.

I. **glm**: É uma função usada para gerar um ajuste para modelos lineares generalizados com ênfase em Regressão Logística, especificados fornecendo uma descrição simbólica do preditor linear e uma descrição da distribuição de erros. Nessa função, coloca-se a fórmula gerada a partir das variáveis de estudo, ou seja, os dados de química para a relação Eu-N e Eu-O. Além desse argumento, a binomial foi escolhida como família ou distribuição de probabilidade.

II. **summary**: Com o ajuste obtido, usa-se essa função com a finalidade de produzir resumos de resultados de várias funções de ajuste de modelo. São exemplos desses resultados: *Deviance Residuals*, os coeficientes das variáveis do modelo com o valor 'p', o desvio nulo, o desvio residual, o valor do critério de informação AIC e o número de iterações de pontuação de Fisher.

III. O parâmetro "response"(resposta) que equivale ao vetor que corresponde ao grupo das ligações químicas, ou seja, 0 caso não haja ligação química e 1 caso exista.

IV. O **fitted.values** (valor ajustado) para o ajuste que foi gerado, com o objetivo de fazer uma previsão de um modelo estatístico do valor médio da resposta quando você insere os valores dos preditores, níveis de fator ou componentes no modelo.

4 Resultados

Nessa seção, apresenta-se os principais resultados obtidos. O modelo considerado para analisar a ligação Európio-Oxigênio será:

$$\ln\left(\frac{\pi_{Eu-O}(d_{Eu-O})}{1-\pi_{Eu-O}(d_{Eu-O})}\right) = \beta_0 + \beta_1 d_{Eu-O}, \tag{4.1}$$

onde d_{Eu-O} representa a distância atômica entre o Európio e o Oxigênio em questão e π_{Eu-O} representa a probabilidade deles estarem ligados. Analogamente, o modelo considerado para analisar a ligação Európio-Nitrogênio será:

$$\ln\left(\frac{\pi_{Eu-N}(d_{Eu-N})}{1-\pi_{Eu-N}(d_{Eu-N})}\right) = \beta_0 + \beta_1 d_{Eu-N}, \tag{4.2}$$

onde d_{Eu-N} representa a distância atômica entre o Európio e o Nitrogênio em questão e π_{Eu-N} representa a probabilidade deles estarem ligados.

Para ajuste e seleções dos modelos faremos a partição do conjunto de dados em conjuntos de treino e de teste na proporção 80% treino e 20% teste.

Os ajustes e seleções dos modelos serão realizados nos conjuntos de treino e as avaliações (via sensitividade, especificidade e acurácia) dos modelos finais serão realizadas nos conjuntos de testes.

4.1 Dados undersampling: Európio - Oxigênio

4.1.1 Ajuste do modelo - Conjunto de Treino

O modelo de Regressão Logística construído a partir dos dados balanceados através da técnica de pré-processamento de dados undersampling é apresentado na tabela a seguir:

Tabela 1 – Resultado da Aplicação do Modelo Eu-O - Dados "undersampling"

	Parâmetro	Estimativa	Erro Padrão	Z-Valor	P-Valor
Intercepto	β_0	35.694	3.955	9.026	<2e-16
Eu-O	β_1	-14.178	1.609	-8.812	2e-16

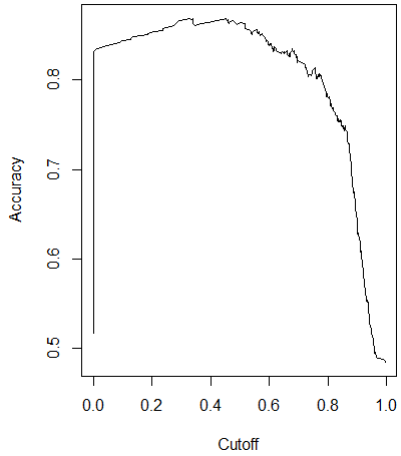
Fonte: Autoria própria.

A Tabela 1 apresenta as estimativas dos parâmetros do modelo (4.1) aplicadas no conjunto de dados, sob o método de reamostragem *undersampling*. Conforme esperado, a distância tem peso negativo na probabilidade, isto é, quanto maior a distância, menor a probabilidade e estarem ligados.

4.1.2 Acurácia dos dados para seleção do corte para classificação - Conjunto de Treino

Na figura 7, observe que é plotado o gráfico da acurácia x corte fora.

Figura 7 – Acurácia Eu-O - Conjunto de treino "undersampling"



Fonte: Autoria própria.

Da figura 7, considera-se que o corte pra o teste preditivo está no ponto 0.4.

4.1.3 Sensitividade, Especificidade e Acurácia (Precisão) - Conjunto de Teste

Na Tabela 2 a seguir, apresenta-se os valores obtidos nos testes de avaliação do modelo de regressão logística Európio - Oxigênio.

Tabela 2 – Avaliação do modelo Eu-O - Dados "undersampling"

Sensitividade	0.7142857
Especificidade	0.979798
Acurácia	0.8579235

Fonte: Autoria própria.

Observamos na Tabela 2 que o modelo possui especificidade bastante alta, com sensitividade não tão alta. Ou seja, quando o modelo afirma que não há ligação, esta afirmação muito provavelmente será verdadeira, já quando o modelo afirma que há ligação, temos ainda aproximadamente 30% de chance de que não haja ligação. No geral, o modelo é satisfatório fornecendo uma acurácia de aproximadamente 86%.

4.2 Dados oversampling: Európio - Oxigênio

4.2.1 Ajuste do modelo - Conjunto de Treino

O modelo de Regressão Logística construído a partir dos dados balanceados através da técnica de pré-processamento de dados oversampling é apresentado na tabela a seguir:

Tabela 3 – Resultado da Aplicação do Modelo Eu-O - Dados "oversampling"

	Parâmetro	Estimativa	Erro Padrão	Z-Valor	P-Valor
Intercepto	β_0	35.426	2.819	12.57	<2e-16
Eu-O	β_1	-14.084	1.149	-12.26	2e-16

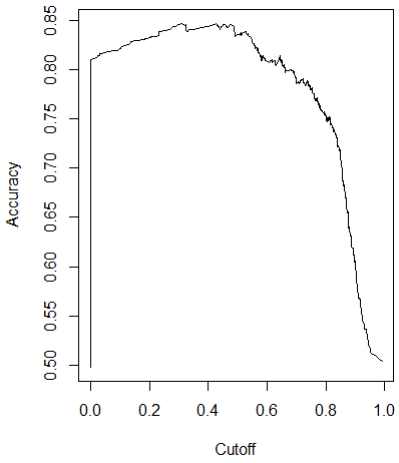
Fonte: Autoria própria.

A Tabela 3 apresenta as estimativas dos parâmetros do modelo (4.1) aplicadas no conjunto de dados, sob o método de reamostragem *oversampling*. Assim como no método *undersampling*, ocorreu o esperado, isto é, a distância tem peso negativo na probabilidade, isto é, quanto maior a distância, menor a probabilidade e estarem ligados.

4.2.2 Acurácia dos dados para seleção do corte para classificação - Conjunto de Treino

Na figura 8, observe que é plotado o gráfico da acurácia x corte fora.

Figura 8 – Acurácia Eu-O - Conjunto de treino "oversampling"



Fonte: Autoria própria.

De forma analoga ao modelo 1, da figura 8, considera-se que o corte pra o teste preditivo também está no ponto 0.4.

4.2.3 Sensitividade, Especificidade e Acurácia (Precisão) - Conjunto de Teste

Na tabela 4 a seguir, apresenta-se os valores obtidos nos testes de avaliação do modelo de regressão logística Európio - Oxigênio.

Tabela 4 – Avaliação do modelo Eu-O - Dados "undersampling"

Sensitividade	0.7251462
Especificidade	0.9717514
Acurácia	0.8505747

Fonte: Autoria própria.

Observamos na Tabela 4 que os resultados foram bastante próximos dos resultados para o método *undersampling*, em particular, o modelo também possui especificidade bastante alta, com sensitividade não tão alta. Ou seja, quando o modelo afirma que não há ligação, esta afirmação muito provavelmente será verdadeira, já quando o modelo afirma que há ligação, temos ainda aproximadamente 30% de chance de que não haja ligação. O modelo possui acurácia levemente menor que o modelo com *undersampling*, tendo 85% de acurácia.

Vamos escolher este modelo como o nosso "melhor" modelo, tendo em vista que acurácia é a métrica mais comum para definir escolhas de modelos. Porém, caso o interesse seja em maximizar a probabilidade da ocorrência de verdadeiros positivos, este modelo (*oversampling*) deve ser escolhido.

4.3 Modelo 1: Probabilidade Eu-O

Para a construção do modelo probabilístico que indica a ligação Eu-O através do modelo logístico apresentado na seção 4.1, na Tabela 1. Observando-se as Tabelas 2 e 4, e tomando como base o critério da acurácia, isto é, escolher o modelo com maior acurácia, vamos escolher o modelo com *undersampling* como nosso "melhor" modelo.

Ressaltamos a observação feita no final da seção anterior, isto é, caso o interesse seja em maximizar a probabilidade da ocorrência de verdadeiros positivos, o modelo com *oversampling* deve ser escolhido.

Portanto, o modelo escolhido (de maior acurácia) foi o:

$$\pi_{Eu-O}(d_{Eu-O}) = \frac{e^{35.694-14.178d_{Eu-O}}}{1+e^{35.694-14.178d_{Eu-O}}}$$

4.4 Dados undersampling: Európio - Nitrogênio

4.4.1 Ajuste do modelo - Conjunto de Treino

O modelo de Regressão Logística construído a partir dos dados balanceados através da técnica de pré-processamento de dados undersampling é apresentado na tabela a seguir:

Tabela 5 – Resultado da Aplicação do Modelo Eu-N - Dados "undersampling"

	Parâmetro	Estimativa	Erro Padrão	Z-Valor	P-Valor
Intercepto	β_0	21.148	5.291	3.997	6.4e-05
Eu-N	β_1	-7.765	2.009	-3.866	0.000111

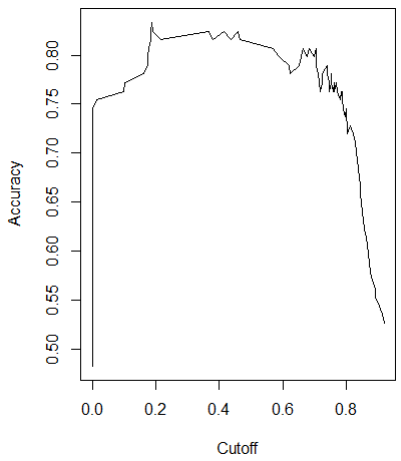
Fonte: Autoria própria.

A Tabela 5 apresenta as estimativas dos parâmetros do modelo (4.2) aplicadas no conjunto de dados, sob o método de reamostragem *undersampling*. Assim como nos modelos para ligações Európio-Oxigênio, ocorreu o esperado, isto é, a distância tem peso negativo na probabilidade, isto é, quanto maior a distância, menor a probabilidade e estarem ligados.

4.4.2 Acurácia dos dados para seleção do corte para classificação - Conjunto de Treino

Na figura 9, observe que é plotado o gráfico da acurácia x corte fora.

Figura 9 – Acurácia Eu-N - Conjunto de treino "undersampling"



Fonte: Autoria própria.

Da figura 9, considera-se que o corte pra o teste preditivo está no ponto 0.4.

4.4.3 Sensitividade, Especificidade e Acurácia (Precisão) - Conjunto de Teste

Na Tabela 6 a seguir, apresenta-se os valores obtidos nos testes de avaliação do modelo de regressão logística Európio - Nitrogênio.

Tabela 6 – Avaliação do modelo Eu-N - Dados "undersampling"

Sensitividade	0.6666667
Especificidade	1
Acurácia	0.8571429

Fonte: Autoria própria.

Neste caso, pela Tabela 6 observa-se que a especificidade foi de 100%, isto é, se o modelo diz que não há ligação, é porque quase certamente não há. A sensitividade foi relativamente baixa, isto é, cerca de 33% dos casos positivos fornecidos pelo modelo são falsos positivos. Por fim, o modelo obteve uma acurácia aceitável de aproximadamente 86%.

4.5 Dados oversampling - Modelo 4: Európio - Nitrogênio

4.5.1 Ajuste do modelo - Conjunto de Treino

O modelo de Regressão Logística construído a partir dos dados balanceados através da técnica de pré-processamento de dados oversampling é apresentado na tabela a seguir:

Tabela 7 – Resultado da Aplicação do Modelo Eu-N - Dados "oversampling"

	Parâmetro	Estimativa	Erro Padrão	Z-Valor	P-Valor
Intercepto	β_0	25.0934	2.5445	9.862	<2e-16
$Eu - N$	β_1	-9.2564	0.9697	-9.546	<2e-16

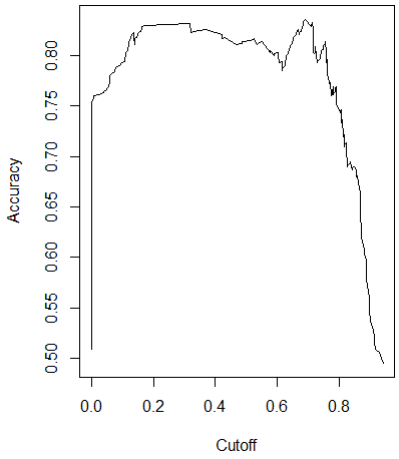
Fonte: Autoria própria.

A Tabela 7 apresenta as estimativas dos parâmetros do modelo (4.2) aplicadas no conjunto de dados, sob o método de reamostragem *oversampling*. Assim como nos demais modelos, ocorreu o esperado, isto é, a distância tem peso negativo na probabilidade, isto é, quanto maior a distância, menor a probabilidade e estarem ligados.

4.5.2 Acurácia dos dados para seleção do corte para classificação - Conjunto de Treino

Na figura 10, observe que é plotado o gráfico da acurácia x corte fora.

Figura 10 – Acurácia Eu-N - Conjunto de treino "oversampling"



Fonte: Autoria própria.

De forma analoga ao modelo 3, da figura 10, considera-se que o corte pra o teste preditivo também está no ponto 0.4.

4.5.3 Sensitividade, Especificidade e Acurácia (Precisão) - Conjunto de Teste

Na tabela 8 a seguir, apresenta-se os valores obtidos nos testes de avaliação do modelo de regressão logística Európio - Nitrogênio.

Tabela 8 – Avaliação do modelo Eu-N - Dados "undersampling"

Sensitividade	0.6666667
Especificidade	0.9615385
Acurácia	0.8035714

Fonte: Autoria própria.

Ao olharmos a Tabela 8, vemos que o modelo com *oversampling* possui a mesma sensitividade que o modelo com *undersampling* porém possui especificidade e acurácia inferiores. Neste sentido fica evidente que no estudo das ligações Európio-Nitrogênio, o modelo com *undersampling* é mais adequado.

4.6 Modelo 2: Probabilidade Eu-N

Para a construção do modelo probabilístico que indica a ligação Eu-N através do modelo logístico apresentado na seção 4.4, na Tabela 5. Observando-se as Tabelas 6 e 8, vamos escolher o modelo com *undersampling* como nosso “melhor” modelo. Neste

caso, diferentemente do caso na ligação Európio-Oxigênio, a escolha foi evidente já que ambos os modelos tiveram a mesma sensitividade, porém o modelo com *undersampling* teve maior especificidade e maior acurácia.

Portanto, o modelo escolhido (de maior acurácia, maior especificidade e mesma sensitividade) foi o:

$$\pi_{Eu-N}(d_{Eu-N}) = \frac{e^{21.148-7.765d_{Eu-N}}}{1+e^{21.148-7.765d_{Eu-N}}}$$

5 Considerações Finais

A partir dos gráficos gerados na seção anterior (figura 7 á figura 10), observou-se que a partir do desenvolvimento da pesquisa em relação a Regressão Logística e a acurácia é possível observar que os métodos de balanceamento dos dados, ou seja, oversampling e undersampling, e também da modelagem dos dados de treinamento, quando aplicados ao modelo regressão de logística retornam um bom índice de precisão. No entanto, a título de escolha, tomando como critério a acurácia, temos:

- Relação Európio - Oxigênio: O Método de undersampling. 85.79% de acurácia.
- Relação Európio - Nitrogênio: O Método de undersampling. 85.71% de acurácia.

Uma observação importante é que no caso das ligações do tipo Európio-Oxigênio, o modelo com *oversampling* possuiu sensitividade maior, ainda que acurácia levemente menor. Assim, caso o interesse seja o de maximizar os verdadeiros positivos, o modelo com *oversampling* é mais adequado.

Por fim, o desenvolvimento da pesquisa em relação a Regressão Logística e o Teste de precisão de Acurácia possibilitou um estudo matemático em relação ao problema e com o auxílio da linguagem de programação R observou-se como determinar o tipo de classificação da ligação química a partir das distâncias Eu-N e Eu-O . Além disso, esse estudo permitiu uma visão mais ampla do campo de aplicações onde a Regresão Logística pode ser aplicada.

REFERÊNCIAS

[1] BEZERRA, Giulyanna Karlla Arruda. **Modelo de Regressão Logística para Previsão do Óbito na Unidade de Terapia Intensiva**. 2012. 90 f. Dissertação (Mestrado) - Curso de Modelos de Decisão e Saúde, Universidade Federal da Paraíba, João Pessoa, 2012.

[2] BORGES, Leonardo Silva Roever. **Medidas de Acurácia Diagnóstica na Pesquisa Cardiovascular**. International Journal Of Cardiovascular Sciences: Diagnostic Accuracy Measures in Cardiovascular Research, Uberlândia, Mg, v. 29, n. 3, p.218-222, 10 jul. 2016. Disponível em: <http://www.onlineijcs.org/sumario/29/pdf/v29n3a09.pdf>. Acesso em: 11 de outubro de 2020.

[3] BERTOZZO, R. J. **Aplicação de Machine Learning em Dataset de Consultas Médicas do SUS**. 2019. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis, 2019. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/202663/TCC.pdf>. Acesso em: 19 de novembro de 2020.

[4] BRAGA, Ana Cristina da Silva. **CURVAS ROC: ASPECTOS FUNCIONAIS E APLICACÕES**. 2000. 267f. Tese (Doutorado) - Curso de Engenharia de Produção e Sistemas, Universidade do Minho Braga. 2000. Disponível em: < http://repositorium.sd um.uminho.pt/handle/1822/195 >. Acesso em: 03 de setembro de 2020.

[5] CARVALHO, A. C. P. L. F.; PADILHA, V. A. **Mineração de Dados em Python: Classificação e Regressão**. ICMC-USP, 2017.

[6] CABRAL, C. I. S. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado**. 2013. Dissertação (Mestrado em Matemática Aplicada à Economia e à Gestão) - Departamento de Estatística e Investigação Operacional, Universidade de Lisboa, 2013.

[7] CLAUDE, J.; BÜNZLI,G. **On the design of highly luminescent lanthanide complexes**. Coordination Chemistry Reviews, Volumes 293–294, 15 June 2015, Pages 19-47.

[8] DeepAI. **Statistical Classification: What is Statistical Classification?**. Disponível em: <deepai.org/machine-learning-glossary-and-terms/statistical-classification>. Acesso em: 17 de novembro de 2020.

[9] DEVELOPERS GOOGLE. **Data Preparation and Feature Engineering for Machine Learning: Normalization**. Disponível em: <https://developers.google.

com/machine-learning/data-prep/transform/normalization>. Acesso em: 20 de novembro de 2020.

[10] DUTRA, J. D. L.; FILHO, M. A. M.; ROCHA, G. B.; FREIRE, R. O.; SIMAS, A. M.; Stewart, J. J. P.; Chem, J. **Sparkle/PM7 Lanthanide Parameters for the Modeling of Complexes and Materials**. Theory Comput. 2013, 9, 8, 3333–3341.

[11] DUTRA, J. D. L.; Bispo, T. D.; FREIRE, R. O. **LUMPAC lanthanide luminescence software: Efficient and user friendly**. Journal of Computational Chemistry, 35(10), 772-775 (2014).

[12] ERNST, Felix G.M.. **ROCR: visualizing the performance of scoring classifiers**. Visualizing the Performance of Scoring Classifiers. 2020. Disponível em: <https://www.rdocumentation.org/packages/ROCR/versions/1.0-11>. Acesso em: 11 nov. 2020.

[13] FACELLI, K., LORENA, A. C., GAMA, J., e CARVALHO, A. C. P. L. F. (2011). Inteligência artificial: Uma abordagem de aprendizado de máquina. 2011. Rio de Janeiro: GEN - LTC, 2:192

[14] FACULDADE DE ENGENHARIA DA UNIVERSIDADE DE PORTO. **Tipos de Data Mining**. Disponível em: <https://paginas.fe.up.pt/~mgi99021/it/tipos.htm>. Acesso em: 13 de outubro de 2020.

[15] FACULDADE DE MEDICINA DE PORTO. **Avaliação de Testes Diagnósticos**. 2020.

[16] FIGUEIRA, C. V. **Modelos de Regressão Logística**. Dissertação (Mestrado em Matemática) Programa de Pós-Graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio Grande do Sul. Porto Alegre, 2006.

[17] FILHO, C. H. P. **Técnicas de aprendizado não supervisionado baseadas no algoritmo da caminhada do turista**. 2017. Dissertação (Mestrado em Bioengenharia) - Bioengenharia, Universidade de São Paulo, São Carlos, 2017. doi:10.11606/D.82.2018.tde-20082018-122603. Acesso em: 2020-11-20.

[18] FILHO, M. A. M.; Dutra, J. D. L.; CAVALCANTI, H. L. B.; Rocha, G. B.; SIMAS, A. M.; FREIRE, R. O.; CHEM, J. **RM1 Model for the Prediction of Geometries of Complexes of the Trications of Eu, Gd, and Tb**. Theory Comput. 2014, 10, 8, 3031–3037.

[19] GAMBARRA, Priscila Alves Nóbrega. **As repercussões do Ruído Ocupacional na Audição dos Cirurgiões Dentistas das Unidades de Saúde da Família de João Pessoa-PB**. 2012. 125 f. Dissertação (Mestrado) - Curso de Modelos de Decisão e saúde, Universidade Federal da Paraíba, João Pessoa, 2012.

[20] GONZALEZ, L. A. **Regressão Logística e suas Aplicações**. 2018. 46f. Monografia (Graduação) - Ciência da Computação, Universidade Federal do Maranhão,

São Luís, 2018. Disponível em: <https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf>. Acesso em: 13 de novembro de 2020.

[21] JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R**. 2013. Springer.

[22] Hawkins, D. M. **Identification of Outliers**. 1980. Springer.

[23] KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. **Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, And Case Studies**. Massachusetts, Cambridge: The MIT Press, 2015.

[24] MACALUS, E.; RUBÍN, M.; AGUILÀ, D.; CHIESA, A.; BARRIOS, L.A; MARTÍNEZ, J.I.; ALONSO, P.J.; ROBEAU, O.; LUIS, F.; AROMÍ, G.; CARRETA, S. **A heterometallic [LnLnLn] lanthanide complex as a qubit with embedded quantum error correction**. Chemical Science, 2020, 11, 10337-10343.

[25] MARTINEZ, E. Z; LOUZADA, F; PEREIRA, B. (2003). **A curva ROC para testes diagnósticos**. Cad Saúde Coletiva. 11. 7-31. <https://www.researchgate.net/publication/284295708_A_curva_ROC_para_testes_diagnosticos>. Acesso em: 13 de novembro de 2020

[26] MAYER, Fernanda de Pol et al. **Introdução à Estatística e conceitos de amostragem**. Curitiba: Ufpr, 2016. 48 slides, color. Disponível em: <http://leg.ufpr.br/~fernandomayer/aulas/ce001e-2016-2/01_introducao_e_amostragem/01_Introducao_a_Estatistica_e_amostragem.pdf>. Acesso em: 18 nov. 2020.

[27] MITCHELL, T. M. **Machine learning**. 1997. Burr Ridge, IL: McGraw Hill, 45(37):870–877.

[28] PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD M. C. **Evaluating Classifiers Using ROC Curves**. IEEE Latin America Transactions, vol. 6, no. 2, pp. 215-222, Junho de 2008, doi: 10.1109/TLA.2008.4609920. Disponível em: <https://ieeexplore.ieee.org/document/4609920/authors#authors>. Acesso em: 20 de outubro de 2020.

[29] RIGOLLET, P. **Mathematicals of Machine Learning**. 18.657 Mathematics of Machine Learning. Fall 2015. Massachusetts Institute of Technology: MIT OpenCourseWare. Acesso em: <https://ocw.mit.edu>. License: Creative Commons BY-NC-SA.

[30] ROBIN, X. **PROC: display and analyze roc curves. Display and Analyze ROC Curve**. 2020. Disponível em: <https://www.rdocumentation.org/packages/pROC/versions/1.16.2>. Acesso em: 13 novembro de 2020.

[31] RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. New Jessey: Prentice Hall, Englewood, 1995.

[32] SHAMOO, A. E.; RESNIK, D. B. **Responsible Conduct of Research**. Segunda Edição. New York: Oxford University Press, 2009.

[33] SILVA, F. T.; LINS, S. L. S.; SIMAS, A. M; CHEM, I. **Stereoisomerism in Lanthanide Complexes: Enumeration, Chirality, Identification, Random Coordination Ratios**. 2018, 57, 17, 10557–10567.

[34] SOUZA, L. D. A. **Uso de Regressão Logística para Identificar os Fatores de Risco associados à Ocorrência de Anomalias Congênitas em Recém-nascidos**. 2013. 38f. Monografia (Graduação) - Bacharelado em Estatística, Universidade Federal da Paraíba, João Pessoa, 2013. Disponível em: <https://repositorio.ufpb.br/jspui/bitstream/123456789/823/1/LDAS03112013.pdf>. Acesso em: 12 de novembro de 2020.

[35] WICKHAM, H. **readxl: Read Excel Files**.2020 Disponível em: <https://www.rdocumentation.org/packages/readxl/versions/0.1.1>. Acesso em: 13 de novembro de 2020.

[36] ZHANG, H.; ZHANG, X.; MA, J. **Chemical Bonding Characteristics of Lanthanide Complexes: A Case of Valence Study**. Advanced Materials Research, vol. 634–638, Trans Tech Publications, Ltd., Jan. 2013, pp. 3–6. Crossref, doi:10.4028. Disponível em: <www.scientific.net/amr>. 634-638.3.