**UNIVERSIDADE DE SÃO PAULO**
**INSTITUTO DE MATEMÁTICA E ESTATÍSTICA**

# Educação Estatística na Escola Básica: Introduzindo Software CODAP na Análise Descritiva e Árvore de Decisão

Rafael Vieira Bonangelo

Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo como exigência para obtenção do título de Mestre em Ciências.

Programa: Mestrado Profissional em Ensino de Matemática
Orientadora: Profa. Dra. Lisbeth Kaiserlian Cordani

São Paulo
Dezembro de 2023

**UNIVERSIDADE DE SÃO PAULO**
**INSTITUTO DE MATEMÁTICA E ESTATÍSTICA**

# Educação Estatística na Escola Básica: Introduzindo Software CODAP na Análise Descritiva e Árvore De Decisão

Rafael Vieira Bonangelo

Versão corrigida da Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo como parte dos requisitos para obtenção do título de Mestre em Ciências, no Programa de Mestrado Profissional em Ensino de Matemática.

São Paulo
Outubro de 2023

BONANGELO, Rafael Vieira. **Educação Estatística na Escola Básica: Introduzindo *Software* CODAP na Análise Descritiva e Árvore De Decisão**. 2023. 94p. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Aprovado em: 01/12/2023

Banca Examinadora

Prof(a). Dr(a). Lisbeth Kaiserlian Cordani (Presidente)
Instituição: Universidade de São Paulo (USP)
Julgamento: _____

Prof(a). Dr(a). Ângela Tavares Paes
Instituição: Universidade Federal de São Paulo (Unifesp)
Julgamento: _____

Prof(a). Dr(a). Cláudia Borim da Silva
Instituição: Faculdade das Américas (FAM)
Julgamento: _____

**Agradecimentos**

À minha mãe Marinalva Vieira por toda companhia, força, dedicação e encorajamento em toda a minha vida.

Ao Alexandre Rodrigues por todas as reflexões que me ajudaram a ser uma pessoa melhor.

À professora Lisbeth Cordani pela orientação para o desenvolvimento deste trabalho e o acompanhamento durante o mestrado.

Às professoras Ângela Paes, Cláudia Borim e aos professores Marcos Magalhães e Alonso Soler pelas contribuições na defesa e na qualificação do mestrado.

Aos amigos Jacqueline, Débora, Pedro, Ivo, Rubens, Elaine e Barbara pela inspiração nos estudos e na vida.

Aos diversos professores que me ensinaram muito sobre o que é ser professor, em especial à Jacqueline, Maria Silvia e Lilian.

À Júlia Otero pela correção do português na versão final deste texto e à Telma Melo pela versão da qualificação.

**RESUMO**

BONANGELO, Rafael Vieira. **Educação Estatística na Escola Básica: Introduzindo *Software* CODAP na Análise Descritiva e Árvore de Decisão**. 2023. 94p. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

A Educação Estatística é fundamental para compreender diversas informações do cotidiano. Outra possibilidade é explicar áreas e tecnologias em destaque nos meios de comunicação, como a Inteligência Artificial. Portanto, neste trabalho o objetivo foi verificar os assuntos estudados nas aulas de Estatística da escola e que possam explicar um algoritmo de Inteligência Artificial, por exemplo a Árvore de Decisão. Para isso, foi feito uma revisão da literatura para definir Estatística, Educação Estatística, Inteligência Artificial, *Big Data* e Ciência de Dados. Também foram realizadas duas Análises Exploratórias de Dados com duas bases de dados distintas, utilizando medidas de posição, de variação, gráficos uni e bivariados. Um dos dois conjuntos de dados analisados foi utilizado para ilustrar o funcionamento da Árvore de Decisão. Dessa forma, a análise de dados multivariados pode ser abordada de forma exploratória na Educação Básica.

**Palavras Chaves**: Estatística; Educação Estatística; Ciência de Dados; Inteligência Artificial; Árvore de Decisão

**ABSTRACT**

BONANGELO, Rafael Vieira. **Statistical Education in Elementary School: Introducing CODAP Software in Descriptive Analysis and Decision Trees**. 2023. 94p. Dissertation (Master's degree) – Institute of Mathematics and Statistics. University of São Paulo

Statistical education is fundamental to understanding a variety of everyday information. Another possibility is to explain areas and technologies highlighted in the media, such as Artificial Intelligence. Therefore, the aim of this work was to verify the subjects studied in Statistics classes at school that could explain an Artificial Intelligence algorithm, for example the Decision Tree. To do this, a literature review was carried out to define Statistics, Statistics Education, Artificial Intelligence, Big Data and Data Science. Two Exploratory Data Analyses were also carried out with two different databases, using measures of position, variation, uni- and bivariate graphs. One of the two sets of data analyzed was used to illustrate how the Decision Tree works. In this way, multivariate data analysis can be approached in an exploratory way in Basic Education.

**Keywords**: Statistics; Statistics Education; Data Science; Artificial Intelligence; Decision Tree

# LISTA DE ILUSTRAÇÕES

# Sumário

## INTRODUÇÃO

A Educação Estatística tem um papel fundamental no entendimento de informações propagadas por diversos meios de comunicação e na interpretação de dados gerados pelas próprias pessoas. Além disso, o Ensino de Estatística pode auxiliar no entendimento de diversas áreas baseadas em conhecimentos estatísticos, como a Inteligência Artificial.

A Inteligência Artificial apresenta potencialidades que são desconhecidas pela maioria da população, permitindo que seja tratada como uma entidade superior, quando, na verdade, seu funcionamento está embasado em diversas áreas, dentre elas a Estatística, e pode ser compreendida, de forma bastante simplificada, com conceitos aprendidos na Educação Básica.

A organização dos conhecimentos estatísticos a serem ensinados na Educação Básica é apresentada pela Base Nacional Comum Curricular (BRASIL, 2018), um documento que propõe todos os conteúdos a serem abordados na escola e em todo território brasileiro. Anteriormente, os Parâmetros Curriculares Nacionais (PCN) (BRASIL, 2000) eram os documentos vigentes, os quais foram os primeiros a formalizar o ensino de Estatística para todo o território nacional.

Atualmente, os alunos da Educação Básica têm contato com a Estatística desde o ensino fundamental até o final do Ensino Médio. Contudo, a formação inicial dos professores de matemática começou a explorar conteúdos de Estatística a partir da década de 1950 e ainda é tema de pesquisa constante.

Os professores que ensinam Estatística na escola são licenciados em Matemática, sendo que a Estatística ocupa uma pequena parte na carga horária dos cursos de formação inicial, em média, 2,5% da carga horária total do curso, segundo Viali (2008). Rodrigues e Silva (2019) corroboram com o fato de que 2% da carga horária dos cursos de licenciatura em Matemática possuem matérias relacionadas à Estatística, além de existirem seis licenciaturas em matemática de 190 cursos ofertados que não oferecem nenhuma matéria relacionada à Estatística.

Junto com a Estatística, também é interessante o uso de tecnologias digitais para diminuir a ênfase em cálculos e permitir maior tempo para a interpretação de resultados e gráficos estatísticos, além do estudo de casos reais (BARGAGLIOTTI *et al*., 2020), (BRASIL, 2018), (BRASIL, 2000), (YATES, 1968), (TUKEY, 1965).

A ênfase na tecnologia não deve ser na programação de computadores, pois isso adicionaria uma dificuldade ao ensino de conceitos diversos na Estatística abordada na Educação Básica. Portanto, também é essencial utilizar tecnologias de fácil acesso, desenvolvidas para o ensino de Estatística e com uso de diversos professores e pesquisadores pelo mundo.

Existem inúmeras opções de tecnologias utilizadas no ensino de Estatística. Alguns trabalhos (BIEHLER, 2019), (BONANGELO; CORDANI, 2022) (ENGEL, 2018), (BUDDE et al., 2020), (FRISCHEMEIER et al., 2021), (BIEHLER; FLEISCHER, 2021) enfatizam o uso das tecnologias para facilitação em algumas etapas da análise de dados, o estudo de simulações ou o estudo de tópicos de Inteligência Artificial com base na Estatística. Tais trabalhos são importantes, dado que educadores exploram conceitos tecnológicos para melhorar práticas de ensino e o aprendizado dos alunos. O contrário, ou seja, cientistas da computação desenvolverem técnicas para auxiliar no ensino pode não ser tão proveitoso, como discutido por Klutka, Ackerly e Magda (2018), que mostram a ênfase em algoritmos ou ferramentas destinadas para a administração escolar no nível da educação universitária. Sendo assim, é importante que profissionais da educação se apropriem desses conhecimentos e o utilizem de forma a melhorar sua prática diária em sala de aula.

O desenvolvimento de materiais para o ensino de Estatística é importante para que a Inteligência Artificial e outras tecnologias com base estatística sejam mais bem compreendidas pela população geral. Deste modo, explorar essas tecnologias com conhecimentos da Educação Básica é uma forma de contextualizar a Estatística escolar e de propiciar o desenvolvimento de um senso crítico às pessoas quanto ao funcionamento, uso e possibilidades de tecnologias para melhoria da sociedade.

Com isso, o questionamento a ser explorado neste trabalho é: “De que forma os conhecimentos estatísticos aprendidos na Educação Básica podem explicar o funcionamento, uso e possibilidades das tecnologias atuais, como a Inteligência Artificial?”. Assim, o objetivo geral é explorar os conhecimentos estatísticos desenvolvidos na escola básica para exemplificar o funcionamento da Inteligência Artificial e de um algoritmo chamado Árvore de Decisão.

A fim de responder à pergunta deste trabalho, foram propostos cinco capítulos, sendo o primeiro uma contextualização da Educação Estatística, sua história e abrangência, assim como as tecnologias baseadas no conhecimento estatístico, por exemplo a Inteligência Artificial, a Ciência de Dados e o Big Data. No segundo

capítulo, a Base Nacional Comum Curricular será explorada, assim como trabalhos que sugerem o uso de tecnologia no Ensino de Estatística. No terceiro capítulo serão revisados trabalhos que utilizaram algum *software* para analisar dados com alunos da Educação Básica, além do desenvolvimento de duas Análises Exploratórias de dados para exemplificar conceitos passíveis de serem abordados nas aulas de Estatística na escola. No quarto capítulo foi exemplificado o funcionamento de um algoritmo de Inteligência Artificial a partir de uma das bases analisadas no capítulo anterior. Por fim, o quinto capítulo será destinado para as considerações finais.

# CAPÍTULO 1

## 1. Estatística

Na presente pesquisa, a Estatística será adotada como uma disciplina presente no trabalho com dados, suas variações e probabilidades, como ciência em si mesma, mas também como ferramenta para as outras áreas que utilizam seus conhecimentos e formas de pensamento.

A palavra “Estatística” vem da palavra alemã *Statistik,* a qual deriva da palavra *status* em Latim e é relacionada à palavra Estado. Seu uso se deu em relação à análise de dados sobre o estado, ou a aritmética política, utilizada com bastante frequência no século XVIII nas estatísticas oficiais, além de aplicações na economia, demografia e política (FIENBERG, 2014) (JOHN 1883).

Contudo, depois de tantos anos, o que se entende por Estatística? Não há um consenso entre os autores, pois a palavra é classificada como ferramenta, ciência, ou arte e sempre possui resultados adquiridos a partir de dados, suas variabilidades, suas representações e seus processos inferenciais. Além da discussão do que significa estatística (mais detalhes em Fienberg, 2014), a seguir são apresentadas outras perspectivas de acordo com célebres estatísticos.

O renomado estatístico do século XX, Kendall (1945), define a Estatística como um ramo do método científico, com abrangência a qualquer área de conhecimento que o utilize.

> Estatística é o ramo do método científico, o qual trata com dados obtidos por contagem, ou mensuração de propriedades da população do fenômeno natural. Nesta definição “fenômeno natural” inclui todos os acontecimentos do mundo externo, sejam humanos, ou não. (KENDALL; 1945, p. 2, *tradução nossa*)

De forma similar, Moore (1998) entende a Estatística como uma das *Liberal Arts,* sendo esta definida como “(...) flexíveis e amplamente aplicáveis modos de pensar” (p. 1254, *tradução nossa*). Ainda propõe a independência dela a qualquer outra disciplina e, de maneira similar a Kendall (1945), a descreve como um método, neste caso “fundamental” e não só parte do “científico”.

> Estatística é um método intelectual *geral* que se aplica onde dados, variação e possibilidade apareçam. É um método *fundamental* porque dados, variação e possibilidades são onipresentes na vida moderna. É uma disciplina

> *independente* com suas próprias ideias básicas ao invés de, por exemplo, um ramo da matemática (p. 1254, grifo do autor, *tradução nossa*).

Similar à proposta de Kendall (1945), Kish (1978) discorre sobre a presença da Estatística em qualquer outra ciência na interação entre chance e dados empíricos.

Em estudos mais recentes, os autores

> entendem a Estatística como um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento (MAGALHÃES; LIMA; 2015, p. 1).

Desde o advento dos computadores e a possibilidade de processamento massivo de informações e automatização de processos, questiona-se, ainda hoje, a tarefa dos estatísticos, em especial os diversos métodos ensinados comumente. “Estatísticos precisam ser capazes de pensar de diversas formas: estatisticamente, matematicamente e computacionalmente” (WILD; UTTS; HORTON, 2018, p. 10, *tradução nossa*). Restando questionamentos do tipo como, por que e para que movimentar tais formas de pensar.

Ao se trabalhar com temas reais é necessário refletir sobre o que será coletado, analisado e o que auxiliará no processo. Em outras palavras, “qualquer aprendizagem substancial através dos dados envolve a extrapolação do que você consegue ver nos dados, você deve saber como isso está relacionado com um universo maior” (WILD; UTTS; HORTON, 2018, p. 11, *tradução nossa*). Além disso,

> Estatística é uma meta-disciplina que se concentra em como pensar sobre a conversão de dados que leve a uma compreensão no mundo real. Estatística como meta-disciplina avança quando lições metodológicas e princípios de uma parte particular do trabalho são abstraídos e incorporados em uma estrutura teórica que permite que ela seja utilizada em diversos outros problemas e em diversos outros lugares (WILD; UTTS; HORTON; 2018, p. 7, *tradução nossa*).

O uso da Estatística em outras áreas será apresentado a seguir, pois muito da tecnologia hoje disponível possui uma base estatística antiga e consagrada. Após tal discussão, serão abordados tópicos relacionados a Educação Estatística.

## 1.1. Ciência de Dados

Uma área a ser destacada e com imensa base estatística é a Ciência de Dados (CD). De forma simplificada, Gould (2017) diz que a CD é uma combinação de

pensamentos estatísticos, computacionais e matemáticos. Note que a CD não é entendida pelo autor como a interseção dessas três grandes áreas, ou formada por partes iguais de cada uma.

Em um editorial, MacGillivrary (2019) argumenta que Estatística e CD não são subáreas da Matemática e ainda propõe a diferença entre ambas da seguinte forma: "Estatística é a ciência da incerteza, variação e dados; ciência de dados é a ciência dos dados" (MacGillivrary, 2019, p.41, *tradução nossa*).

A fundamentação do termo CD é essencial para entender a evolução desta área e sua presença no cotidiano. Existe a necessidade de se entender o que tal área propõe e sua ligação com a área de Estatística, já que aparentemente esta cedeu seu lugar para a CD. Segundo Wild, Utts e Horton (2018), por causa da disponibilidade de dados e a disposição de *softwares* de fácil manuseio para tratá-los, a Estatística parece marginalizada com a ascensão da CD.

A CD não deveria ser um termo exclusivo para unificar Estatística, Análise de Dados e seus métodos, mas também deve agrupar seus resultados. Com o crescente volume de dados gerados por cada usuário no mundo digital, tal área ganha destaque contínuo, abordando conhecimentos matemáticos, computacionais e estatísticos (Hayashi,1998). Contudo, a separação da Análise de Dados da Estatística propõe estranhamento, pois a Estatística possui como um de seus pilares a Análise de Dados.

A imensa disponibilidade de dados e o poder computacional crescente a cada ano aumentam notoriedade da CD, ainda que essa ciência seja conduzida também pela Estatística. A percepção desses campos como subáreas se deve à larga abrangência de ambas e a utilização instrumental de cada uma em outras ciências. Alguns autores propuseram discussões acerca da abrangência do fazer estatístico e como há espaço para a Estatística e a Ciência de Dados, por exemplo, Tukey (1962), Marquardt (1987), Chambers (1993), Cleveland (2001) e Donoho (2017).

Tukey (1962) propõe o termo Análise de Dados (AD) para uma perspectiva mais prática para a Estatística numa época em que ela era abordada como um ramo da Matemática e seus pesquisadores buscavam a otimização de modelos, teorias de inferência e suas formalizações, além de outras atividades semelhantes. Ele propôs outras concepções, visando assuntos de cunho mais prático, nas quais as preocupações eram os

> (...) procedimentos para análise de dados, técnicas para a interpretação dos resultados de tais métodos, formas de planejar o agrupamento dos dados

> para fazer a análise mais fácil, mais precisa, ou com veracidade, e todo o maquinário e resultados (matemático) estatísticos aplicados na análise dos dados. (TUKEY, 1962, p. 2, *tradução nossa*).

Mesmo que "(...) o termo aqui esteja aumentado além da sua filologia" (Tukey, 1962, p. 2, *tradução nossa*), com tais alegações, o pesquisador continua sua argumentação, definindo a AD como ciência e não só uma simples segmentação da Matemática.

A evolução da AD[1] para CD foi acontecendo de modo natural, conforme ambientes de programação quantitativa surgiram, como a linguagem S em meados de 1970 e sua derivada, a linguagem R na década de 1990. Com esse desenvolvimento, os algoritmos processados podem ser divulgados em artigos e textos, possibilitando a experimentação por parte de outros cientistas (DONOHO, 2017) e aprimorando, com isso, o método científico.

Marquadt (1987) propõe uma discussão sobre a visão restrita e abrangente do trabalho com estatística. Um tópico apontado pelo autor é a discussão de outros profissionais que também trabalham com estatística e isso não cria “competidores”. Os estatísticos deveriam buscar união com esses outros especialistas de outras áreas para melhorarem o uso do método científico.

Donoho (2017) propõe a diferenciação de Estatística e da CD através de várias abordagens populares. Uma delas é a separação de ambas as áreas através do imenso volume de dados, habilidade e trabalho.

O volume de dados é comum ao trabalho do estatístico (por exemplo no Censo Demográfico), assim como do cientista de dados (dados analisados sincronamente em redes sociais). Sobre habilidades, os cientistas de dados são exaltados por seus conhecimentos computacionais, porém esses não são novidades para estatísticos devido a implantação de *software* estatísticos, por exemplo as linguagens S e R. Por fim, sobre o trabalho, talvez a única vantagem dos profissionais de CD sobre os de Estatística é a formação abordando uma maior variedade de tecnologias.

Outra visão ainda proposta por Donoho (2017) é a evolução da CD a partir de uma defasagem no método de trabalho da Estatística de meados do século XX. Enquanto esta última enfatizou a modelagem e rigor matemático se ocupando com questões de teoremas, modelos e suas demonstrações, havia uma comunidade

[1] Definida como ciência por Tukey (1962), mas sem grande impacto na comunidade acadêmica e nos estudiosos da área até o começo do século XXI

menor de estatísticos e outra exterior (que provavelmente originou o chamado “cientista de dados”) trabalhando com a proposta primordial da Estatística no trabalho com dados. Aqui, encontrava-se a diferença: na adaptabilidade em resolver problemas práticos de modo mais pragmático.

Outros autores que também abordaram essa diferença nas comunidades de estatísticos e trabalhos foram John Chambers (1993) ao abordar as possibilidades que surgiram com o computador e Cleveland (2001) que descreveu o trabalho de ensino de Estatística nas universidades, já incluindo a CD.

Chambers (1993) define *Greater Statistics* e *Lesser Statistics,* equivalente a Estatística Abrangente e Estatística Estrita (*tradução nossa*) no qual a primeira se refere, de forma simplificada, ao aprendizado através de dados, uma área que

> (...) tende a ser inclusiva, eclética com respeito a metodologia, mais associada a outras disciplinas, e praticada por muitos fora da academia e, normalmente, por estatísticos profissionais distantes dela também (p. 182, *tradução nossa*).

Ou seja, profissionais não necessariamente com formação acadêmica em Estatística trabalharam com a CD emergente na época.

Enquanto a segunda, Estatística Estrita, refere-se a profissionais envolvidos com trabalhos em artigos, dissertações e teses,

> (...) tende a ser exclusiva, orientada para técnicas matemáticas, com colaboração menos frequente em outras disciplinas, e primordialmente praticada por membros de departamentos de estatística da universidade (CHAMBERS, 1993, p. 182, *tradução nossa*).

Trata-se, portanto, de uma Estatística mais acadêmica, concentrada em métodos matemáticos.

Ainda no final do século XX, Chambers (1993) afirma que o uso do computador na Estatística "(...) tem o potencial de ampliar nossa visão. Para exercer um papel importante, *softwares* estatísticos devem ser integrados em todo o processo de aprendizado a partir de dados". (p. 182, *tradução nossa*). Em conjunto com essa tecnologia, o autor também propõe um modelo para o trabalho na Estatística Abrangente para:

- “*Preparação* dos dados, incluindo planejamento, coleta, organização e validação.
- *Análise* dos dados, através de modelos ou outras metodologias.

- *Apresentação* dos dados escrita, gráfica ou de outra forma" (CHAMBERS, 1993, p. 182, grifo do autor, *tradução nossa*).

Contudo, além das máquinas e dos estatísticos, há ainda os centros de formação. Cleveland (2001), por exemplo, já no início do século XXI, planeja uma estrutura para cursos de estatística abordarem as diversas áreas técnicas de forma sistemática, incluindo a CD, possibilitando que analistas e cientistas aprendam a partir de dados. A estrutura desses cursos é segmentada em **Investigação Multidisciplinar**, **Modelos e Métodos para Dados**, **Computação com Dados**, **Pedagogia**, **Avaliação de Ferramentas** e **Teoria**.

O motivo da **Investigação Multidisciplinar** é justificado em Cleveland (2001) com o trabalho de pesquisadores da Estatística, como Ronald Fisher (1890 – 1962) no experimento com dados da agricultura, John Tukey (1915 – 2000) e os provenientes das ciências físicas/aplicadas ou George Box (1919 – 2013) com estudos baseados em processos químicos – três estatísticos que promoveram grandes avanços em outras áreas além da Estatística.

Em relação aos **Modelos e Métodos**, o analista de dados encontra dois tópicos fundamentais: 1) a construção de um modelo para os dados e 2) a estimação e análise da distribuição de atividades que necessitam do conhecimento da área, além das áreas da Matemática e Estatística.

> Construção de modelos é complexa, porque requer a combinação de informação originada da exploração de dados e informações de fontes externas aos dados, tal como tópicos importantes da teoria e outros conjuntos de dados (CLEVELAND, 2001, p. 23, *tradução nossa*).

Sobre a **Computação com Dados** "Uma coleção de modelos e métodos para análise de dados será usada se, e só se, a coleção for implementada em um ambiente computacional no qual faça os conjuntos de hipóteses e normas suficientemente eficientes para uso" (CLEVELAND, 2001, p. 23, *tradução nossa*). Além de métodos computacionais, há a preocupação com a administração de sistemas de bancos de dados, outros sistemas e *hardware* para a análise dos mesmos (CLEVELAND, 2001). Seria então a CD uma área da computação?

Com o estudo de recursos e conhecimentos computacionais, seria razoável pressupor isso. Caso fosse, seu desenvolvimento não seria independente, mas "Sem o estímulo de pesquisadores dedicados a inovação, o progresso na computação com dados teria sido mais lenta do que poderia ser (...)" (CLEVELAND, 2001, p. 23,

*tradução nossa*) sendo possível, a partir do desenvolvimento de computadores mais potentes, internet mais rápida e menor custo para acesso de ambos.

Sobre a **Pedagogia**, Cleveland (2001) sugere que o departamento de Ciência de Dados se preocupe em ensinar suas técnicas para estatísticos, mas dirija atenção aos não estatísticos, mostrando o quão valioso é tal área para entender sobre o mundo. Pois, segundo o autor, seria uma concepção limitada a CD ser praticada somente por estatísticos.

Acerca da **Avaliação de Ferramentas**, o mesmo autor também trata da aplicação de questionários para entender possibilidades e necessidades da área, além do desenvolvimento e do aperfeiçoamento de instrumentos de trabalho.

Cleveland (2001) ressalta ainda a importância da **Teoria** através da abordagem formal da estatística matemática, bem como da não formal, pois é preciso pensar em modelos modernos, uma vez que dados novos geram a necessidade de novas ferramentas, e essas estão imersas em novas teorias que guiarão o seu desenvolvimento.

> Ainda assim, todos os estudantes precisam de bons fundamentos em probabilidade matemática, mas a abordagem da probabilidade precisa ser no sentido de variação e variáveis aleatórias, e não a probabilidade no sentido de teoria da medida e funções mensuráveis, necessárias somente para alguns estudantes. A razão se volta aos dados. Dados variam, e essa variação normalmente precisa ser pensada de forma probabilística, e uma ótima intuição para a variação probabilística vem a ser o básico para modelagem avançada de variação (CLEVELAND, 2001, p. 24, *tradução nossa*).

Essa discussão sobre a formação de estatísticos é fundamental. Ainda mais porque já existiam tecnologias com algoritmos poderosos, como a rede neural e a Árvore de Decisão no trabalho com reconhecimento de voz ou imagem, reconhecimento de escrita manual e predição do mercado financeiro. Ou seja, tais comunidades já eram formadas por jovens cientistas da computação, físicos, engenheiros e alguns estatísticos mais experientes (BREIMAN, 2001).

Complementando as reflexões de Cleveland, Breiman (2001) discorre sobre duas abordagens possíveis em se explorar os dados, ressaltando a importância de se entender o que se faz, para se avançar com outras perguntas tais quais como, por que e para que fazer?

Nesse caminho, Breiman (2001) ressalta dois tratamentos de trabalhos estatísticos. Um que pressupõe que os dados são gerados a partir de um modelo, de modo que o estudo é orientado para questões sobre a acurácia e os resíduos do

modelo. E o outro, cuja origem dos dados é desconhecida; portanto, o tratamento propõe um algoritmo para predição da natureza dos dados, um método que é validado através da medição de sua acurácia em fazer previsões.

O trabalho de Breiman (2001) é relevante por ressaltar a predominância dos acadêmicos no primeiro método de tratamento, os quais estão preocupados com a matemática abstrata, enquanto o segundo é mais desempenhado por uma comunidade externa, ressaltando a importância da Estatística se apropriar de meios da CD.

Assim, pode-se entender Estatística como uma disciplina presente no trabalho com dados, suas variações e probabilidades, menos como ciência em si mesma, e mais como ferramenta para as outras ciências que utilizam seus conhecimentos e formas de pensamento. Portanto, Ciência de Dados é a ciência *dos* dados, como dito por MacGillivrary (2019), é munida de técnicas da Estatística abrangente, como definido por Chambers (1993) e é composta por métodos de modelagem com algoritmos, segundo Breiman (2001).

Para além da CD, também há outras áreas a serem destacadas, devido aos avanços tecnológicos na criação, processamento e organização dos dados. Tais áreas ganharam destaque com os avanços tecnológicos do século XXI e também são fundamentadas na Estatística. São elas o *Big Data* e a Inteligência Artificial.

## *1.2. Big Data*

Com a imensa quantidade de dados gerados e processados atualmente, tem-se a discussão sobre o *Big Data* (BD) cuja caracterização pode ser proposta pela perspectiva técnica através dos 4Vs (volume, variação, velocidade e veracidade) (CALDAS; SILVA, 2016) (McKINSEY&COMPANY, 2011) (IBM, 2012), e basta que pelo menos um deles esteja presente para ter a classificação de BD. Entretanto, também há uma perspectiva humana através de fatores científicos, sociais e culturais, como propõe Rieder e Simon (2016), boyd e Crawford (2012). O BD traz novas concepções para o conhecimento e possibilita a exploração de novas informações de relações humanas (boyd; CRAWFORD, 2012) (PENTLAND, 2013), propondo questões sobre o que se tem a aprender com tais dados, suas limitações e expectativas.

Nas propostas do BD, os espaços de convivência humana (físicos ou digitais) se tornam mais quantificáveis, possibilitando que áreas exclusivamente humanas tenham uma abordagem quantitativa (boyd; CRAWFORD, 2012). Dados do BD são coletados a partir de perguntas feitas, métodos propostos e experimentos realizados, sendo que eles não são neutros, pois são originados de intenções, mensuram ações e reações sob diferentes contextos, o que sugere que as informações geradas a partir deles trazem interpretações diversas.

> Na era do *big data,* as pessoas não deveriam ser simplesmente receptoras passivas de relatórios baseados em dados. Elas precisam se tornar exploradoras ativas de dados que podem planejar, adquirir, administrar, analisar e inferir através dos dados. O objetivo é usar os dados para descrever o mundo e responder questões difíceis com a ajuda de ferramentas e visualizações de análise de dados. Entender *big data* e suas possibilidades e limitações é importante para incentivar a conduta cidadã e para a prosperidade de uma sociedade democrática. (BEN-ZVI, 2017, p.32, *tradução nossa*)

Um exemplo de área que pode utilizar muitos dados provindos de diversas fontes e agir a partir deles é a Inteligência Artificial.

## 1.3. Inteligência Artificial

A Inteligência Artificial (IA) é uma área cujo nome foi proposto em 1956 pelo cientista John McCarthy (1927-2011). Todavia, suas pesquisas nesse âmbito remontam a meados de 1940 (TUNES, 2019a). Em 1943, os pesquisadores Warren-McCulloch e Walter Pitts já propunham um modelo de neurônio artificial (TUNES, 2019b).

Embora amplamente estudada nas últimas décadas, a definição de IA ainda é bastante discutida. Dentre as oito possíveis definições elaboradas sob as quatro perspectivas que os pesquisadores Russel e Norvig (2016) discutem, podemos dizer que IA é uma área que estuda meios para que a máquina possa pensar e agir segundo perspectivas humanas e racionais, com o intuito de automatizar ou propor a não intervenção humana em um processo em execução.

Perguntas inquietantes na IA investigam como ocorrem estes processos de automatização e se é necessário listar todas as eventualidades, ou se o robô tratará deste problema.

A intenção da área é trabalhar com possibilidades fundamentais da racionalidade e da mente humana, para que a máquina possa funcionar da melhor forma possível dentro de uma determinada situação. Por exemplo, os carros autônomos que devem tomar diversas decisões enquanto estiverem operando, seja parar, ou desviar de algum objeto para evitar um acidente.

Esse tipo de automatização é realizado através de dados que são inseridos ou coletados por meio da máquina. Os robôs aspiradores que mapeiam a casa são um exemplo de máquinas que a partir da disposição dos móveis, mapeiam a casa e aprendem onde podem ir ou não. Os carros autônomos não têm essa possibilidade de aprender em ambiente real. Por isso, diversas imagens de semáforos e de faróis são inseridas na sua programação, e ele aprende o que fazer ao encontrar luzes vermelha, amarela, verde ou branca.

Outros exemplos também podem ilustrar tal questão, como é o caso do artista que ensina a IA a gerar desenhos de observação (CANQUERINO, 2019) ou pesquisadores que ensinam a escrever poemas a partir de imagens (LIU et al., 2018). A IA que auxilia no desenvolvimento de uma simulação para estudos de buracos negros (PEREIRA, 2020). Ou ainda o próprio Google Tradutor (GOOGLE, 2010).

Todos esses processos de automatização funcionam de uma forma muito semelhante. Por exemplo, para as traduções do Google Tradutor[2], o programa analisa muitos textos originais e suas traduções em outros idiomas, e com base nesses dados, infere as regras e faz conexões entre os dois idiomas. Depois de diversos exemplos, quando o programa é acionado para traduzir um texto, ele consulta as regras inferidas e disponibiliza uma alternativa. Esse funcionamento é similar ao atual Chat GPT[3].

Além dos casos citados, a Figura 1 ilustra alguns destaques históricos relacionados a essa área, como o modelo *Deep Blue* vencendo uma partida de xadrez do campeão mundial, o programa MYCIN auxiliando na saúde já em 1970 e o *chatbot* TAY evidenciando questões éticas que são amplamente discutidas na atualidade.

[2] https://translate.google.com.br/
[3] https://openai.com/chatgpt

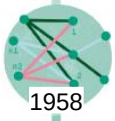
**Figura *1*** - Linha do tempo da Inteligência Artificial


Os cientistas norte-americanos Warren McCulloch e Walter Pitts publicam trabalho pioneiro de inteligência artificial (IA): um modelo de **neurônios artificiais.**


O britânico Alan Turing propõe o **teste de turing**: caso uma máquina se passe por um humano, então ela será chamada de inteligente.

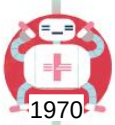

A **expressão inteligência artificial** passa a ser utilizada pelo cientista John McCarthy, da Universidade Stanford (EUA) para batizar um novo campo da ciência.


Os pesquisadores H. Simon, J. C. Shaw e A. Newell criam o **Solucionador de Problemas Gerais (GPS)**, que resolvia problemas expressos em uma representação lógica.




A linguagem **LISP** é desenvolvida e tornou-se dominante no campo de IA. No mesmo ano, o psicólogo F. Rosenblatt lança o **primeiro algoritmo de aprendizado de máquina**.


O cientista da computação J. Weizenbaum cria o **ELIZA**, um programa capaz de processar linguagem e manter um diálogo por meio de frases pré-programadas.




E. Shortliffe desenvolve o **MYCIN** pioneiro entre os sistemas especialistas. Ele codificava o conhecimento para realizar diagnóstico de algumas infecções bacterianas.


Realização do **1º simpósio brasileiro de Inteligência Artificial**, na Universidade Federal do Rio Grande do Sul (UFRGS), em Porto Alegre.


O computador **DEEP BLUE**, da IBM, vence uma partida de xadrez em um jogo com o campeão mundial russo Garry Kasparov.


Um **carro autônomo** desenvolvido na Universidade Stanford vence a DARPA Grand Challenge, competição de automóveis de direção automática.


O computador **EUGENE** passa no Teste de Turing. Criado na Rússia, ele se fez passar por um menino de 13 anos e “enganou” um terço dos juízes.


A Microsoft lança o **chatbot TAY**, um computador com capacidade de aprendizado que simula conversação. Colocado em teste nas redes sociais, em apenas um dia, o programa se torna racista, xenófobo e homofóbico.

Fonte: TUNES, 2019b, (Adaptado)

A área de Inteligência Artificial possui várias formas de trabalho, mas uma delas é o Aprendizado de Máquina (AM). O AM é uma subárea da IA e possui três formas principais de trabalho: 1) Aprendizado Supervisionado; 2) Aprendizado Não Supervisionado e 3) Aprendizado por Reforço.

No Aprendizado Supervisionado, são utilizados dados classificados previamente. A partir dessas informações, um algoritmo infere padrões para novos dados. Um exemplo já citado é o Google Tradutor, pois são apresentados dois textos, o original e o traduzido. Ambos possuem seus idiomas informados para um algoritmo e, a partir disso, o aplicativo pode tentar fazer outras traduções.

O Aprendizado Não Supervisionado possui diversos dados e não se conhece os padrões prévios deles; portanto, o foco será segmentá-los em grupos que possam ser mais bem compreendidos. Por exemplo, é possível segmentar o perfil de consumidores de uma loja ao estudar seus padrões de compras.

O Aprendizado por Reforço propõe uma tarefa para o computador e conforme ele progride positivamente, o algoritmo é recompensado e, caso contrário, é penalizado. Um exemplo clássico é o *DeepMind* citado na linha do tempo (Figura 1) e abordado em uma matéria de jornal por Duarte (2021).

As áreas da Inteligência Artificial e suas subáreas não são tão recentes, mas ganharam destaque atualmente com a grande quantidade de dados produzidos no mundo digital, pois com o aumento no volume de dados, o nosso entendimento sobre esses dados decai (WITTEN; FRANK, 2005).

Para retomar a compreensão das informações adquiridas através desse grande volume de dados a Educação Estatística é imprescindível e se faz presente a partir do início da Educação Básica.

# 2. Educação Estatística

A Educação Estatística é a área que investiga o ensino e o aprendizado da Estatística, considerando características afetivas e cognitivas (CAZORLA; KATAOKA; SILVA, 2010).

> A missão da educação estatística é providenciar uma estrutura conceitual (meios estruturados de pensamento) e habilidades práticas para melhor equipar nossos estudantes para o seu futuro em um mundo em rápida transição. (WILD; UTTS; HORTON; 2018, p. 6, *tradução nossa*)

Tal futuro traz consigo uma grande variedade de possibilidades e de incertezas. Os conhecimentos e competências estatísticas necessários para as pessoas dependerão das suas necessidades individuais ao se trabalhar com dados. Estas individualidades podem ser categorizadas como **produtores de dados** (por exemplo, pesquisadores na área financeira), **usuários da estatística** (que utilizam as elaborações dos anteriores) e **consumidores da área** (cuja tarefa é entender como foram utilizadas e interpretadas as estatísticas em situações cotidianas), como propõem Wild, Utts e Horton (2018).

Os consumidores são a maioria da sociedade moderna. Eles precisam entender a validade das conclusões expostas nas mídias, além de saberem como o pensamento estatístico pode ajudá-los a responder questões e tomar decisões minimizando riscos (WILD; UTTS; HORTON, 2018). Portanto, a população geral precisa compreender estatísticas diárias.

A história mostra que a organização da Educação Estatística é recente. Além disso, a Estatística é abrangente no cotidiano da população e se faz necessária o ensino da Estatística na Educação Básica.

## 2.1. Uma breve história do Ensino da Estatística

As universidades no mundo ofereciam matérias em estatística fundamentais para diversos estudiosos, por exemplo os cientistas sociais, psicólogos, médicos e agrônomos. No entanto, a sua inserção na Educação Básica ocorreu somente no século XX, tanto em âmbito nacional quanto no mundial (CORDANI, 2014) (SILVA; VALENTE, 2015).

Um panorama é proposto por Cordani (2014) no qual a pesquisadora relata o trabalho de instituições e grupos com o objetivo de disseminar o ensino de Estatística para todos, propondo uma postura questionadora para a população em contato com os dados e informações diárias.

Há acontecimentos que merecem destaque quanto ao ensino de Estatística discutido nessas reuniões. Por exemplo, Cordani (2014) cita o caso do estatístico Frank Yates (1902-1994) em um relatório de 1968 para a *Royal Statistical Society* no qual propõe uma redução da quantidade de cálculos e mais atenção à interpretação. Também há o exemplo da criação do periódico *Teaching Statistics*[4] em 1979, destinado a professores do ciclo básico e a promoção de eventos periódicos, como as mesas redondas (*Roundtables*) e a Conferência Internacional sobre Ensino da Estatística (ICOTS) em 1985, ambos para a discussão sobre o ensino e aprendizado de Estatística.

Tais reuniões e propostas foram fundamentais, pois "Para ser inserida no currículo escolar, uma disciplina precisa ser reconhecida como necessária à formação do indivíduo enquanto pessoa e também como profissional"(CORDANI, 2014, p. 160).

A situação brasileira estava na mesma direção que a mundial, porque a Estatística enquanto matéria no ensino universitário estava bem fundamentada para diversas áreas, como na medicina e psicologia. O Brasil no final do século XIX e começo do XX iniciava um processo de cientificidade da educação através da estatística educacional e, de acordo com Silva e Valente (2015), por meio de tais levantamentos, seriam reconhecidos elementos para o desenvolvimento do país.

De relatórios que podem ser destacados, o Censo brasileiro foi posto como obrigatório a partir da década de 1880, periodicamente em dez anos. No entanto, ele teve um início conturbado, sem atender as demandas solicitadas até 1920. Entretanto, com a intervenção do médico sanitarista Bulhões Carvalho (1866 – 1940), a estatística foi bem executada e objeto de reflexão e discussão. Posteriormente, um grande marco na estatística brasileira foi a criação do Instituto Brasileiro de Geografia e Estatística (IBGE) na década de 1930, destinado à coleta, organização e divulgação de estatísticas nacionais (CORDANI, 2014). Atualmente, o IBGE possui como missão “Retratar o Brasil com informações necessárias ao conhecimento de sua realidade e ao exercício da cidadania” (IBGE, 2023).

[4] https://onlinelibrary.wiley.com/journal/14679639

Trabalhos com o foco educacional são citados por Silva e Valente (2015) como a série “O Ensino no Brasil”, de 1939, produzido pelo Serviço de Estatística de Educação e Saúde (SEES) com dados e informações desde 1931 originados de um Convênio Estatístico. “Essa série tinha por objetivo apresentar os resultados da educação no país de forma padronizada, regular e frequente” (SILVA; VALENTE, 2015, p. 446). Na mesma década, os autores também citam o trabalho do educador Lourenço Filho (1897 – 1970), o qual publicou o livro “Tendências da educação brasileira” que em um capítulo justifica a necessidade das estatísticas educacionais, pois se obteria dados e informações para metas claras e passíveis de avaliação. Diversos outros trabalhos eram desenvolvidos e relatórios gerados, como aqueles desenvolvidos pelas delegacias regionais de ensino.

Com tantas informações geradas, a preparação de universitários em cursos de pedagogia e ciências sociais foi necessária. No caso da Universidade de São Paulo (USP), foram propostas matérias denominadas “Estatística geral e aplicada” e “Estatística educacional”. Posteriormente, seus nomes foram modificados para “Estatística I” e “Estatística II”, respectivamente (SILVA; VALENTE, 2015).

Como citado, também era uma preocupação que a sociedade entendesse a Estatística, suas possibilidades e limitações, além de interpretar o que era informado. A inserção da Estatística na escola básica tem destaque com os livros didáticos de Matemática do autor Osvaldo Sangiorgi (1921 – 2017). De acordo com Valente (2007), Sangiorgi propôs um livro chamado “Matemática e Estatística”, no qual há três capítulos: “Aritmética aplicada”, “Geometria Aplicada” e “Noções de Estatística”. Uma obra com 250 páginas, mas somente 50 delas são destinadas à Estatística, abordando tópicos de coleta de dados, representações gráficas, medidas de posição, dispersão e aplicações à Educação.

O uso consciente desse material e o ensino da Estatística é amplamente discutido em âmbito nacional na década de 1950 com a portaria nº 49 de 4 de dezembro de 1954, que propôs a abordagem de Matemática e Estatística em cursos de formação inicial de professores no estado de São Paulo. Essa prática se dissemina pelo Brasil. Porém, o intuito dos cursos de estatística consistia em instruir professores a gerar indicadores para a implantação de políticas públicas (VALENTE, 2007).

Em um artigo na revista “Atualidades Pedagógicas” o autor Sangiorgi (1957) relata a abrangência da proposta de formação para professores da Educação Básica em diferentes estados, no qual alguns detalhes merecem destaque: (I) Os cursos de

formação de professores no Estado de São Paulo, copiados por alguns outros estados, teriam as matérias de “Aritmética aplicada”, “Geometria aplicada” e “noções de Estatística aplicada à Educação”, conforme citado anteriormente sobre o ensino de estatística e a implantação de políticas públicas. (II) Não há uniformidade entre os cursos de formação inicial de professores. Por exemplo, Estatística é adotada em São Paulo e em Minas Gerais, facultativo no Paraná e, por exemplo, não aparece na descrição do curso da Paraíba e de Pernambuco.

Em diversos países do mundo, incluindo o Brasil, havia a consciência da importância do ensino de Estatística desde os anos iniciais da educação formal, porém faltava formação geral e uniforme para os docentes. Cordani (2014) relata que houve investimento em diversos projetos, porém, com a falta de obrigatoriedade em oferecer o ensino de Estatística, muitos foram interrompidos. Para os que continuaram, faltava a noção de como avaliar os trabalhos investigativos com dados. Portanto, houve um retorno para cursos com uma abordagem mais matemática, em que os docentes (que eram da matemática) se sentiam mais confortáveis.

Havia propostas de ensino de estatística elaboradas em âmbito estadual (LOPES, 1998). Carvalho (1995) discute tais documentos e analisa os currículos estaduais referentes ao ensino de Matemática no ensino fundamental nas décadas de 1980 a 1990. Em relação a Estatística há, como pontos positivos:

1) "O tratamento e análise de dados por meio de gráficos" (p. 58). Abordagem facilitada pelos computadores, como será discutida na próxima seção;

2) "A introdução de noções de estatística e de probabilidade" (p. 58);

3) "A percepção de que a função da matemática escolar é preparar o cidadão para uma atuação participativa, crítica na sociedade em que vive" (p. 58). Algo condizente com o que já foi descrito até então neste texto.

Por outro lado, Carvalho (1995) ressalta indícios preocupantes como:

1) "Grande ênfase em detalhamento de conteúdos, como se isso fosse suficiente para garantir uma boa aprendizagem" (p. 58). Uma observação já feita por Frank Yates em 1968 e comentada anteriormente nesta seção.

2) "A ênfase em algoritmos das operações, priorizando-os em relação aos conceitos" (p. 58);

3) "A ausência marcante de noções elementares de estatística e probabilidade, que podem ser apresentadas, respeitando-se o estágio de desenvolvimento dos alunos, desde as primeiras séries" (p. 58).

Além dos currículos estaduais contendo o ensino de estatística, na mesma década de 1990, há a elaboração dos Parâmetros Curriculares Nacionais (PCN) para que secretarias de estados e municípios tivessem diretrizes nacionais para a elaboração de seus próprios projetos educativos. Assim, o ensino de Estatística com o tópico “Tratamento de Informações” é oficializado (CORDANI, 2014) e com a mesma proposta em território nacional. A partir de tal documento ações políticas convergentes poderiam ser propostas, seja na formação inicial ou continuada de professores, desenvolvimento de materiais didáticos e programas de avaliação.

Dentre diversos destaques do currículo, ressaltamos aqui a inclusão dos tópicos de Estatística, Probabilidade e Combinatória desde os primeiros anos do fundamental no conjunto "Tratamento da Informação", além do uso de tecnologias da comunicação (LOPES, 1998).

Com uma breve discussão sobre a formação da Educação Estatística no Brasil para a geração de informações, implantação de políticas públicas ou seu ensino, entende-se a necessidade de compreender conceitos de incerteza, variabilidade, limites, possibilidades dos estudos estatísticos ou o entendimento da coleta, processamento e comunicação de dados e informações. Ou seja, a Estatística fundamenta a postura do cidadão crítico e da sociedade reflexiva, como discorrem Cordani (2014) e Lopes (2008). Logo, conceitos maiores referentes à leitura, ao pensamento e ao raciocínio estatístico emergem. Tópicos que serão abordados a seguir.

## 2.2. Documentos no Ensino Básico

Os Parâmetros Curriculares Nacionais (PCN) e a Base Nacional Comum Curricular (BNCC) são documentos oficiais de âmbito nacional e apresentam diretrizes para o trabalho com dados como parte da Estatística escolar. Com destaque à análise de dados para possibilitar a apropriação do mundo real, o PCN e a BNCC propõem o desenvolvimento de criticidade para com o espaço físico e o consumo consciente de informações. Ambos os documentos são fundamentados sobre a perspectiva da resolução de problemas e uso das tecnologias digitais.

Os PCNs separaram a informática como uma linguagem e ressaltam que o importante não é o aprendizado da programação em si mesma, mas sim, o uso do que é atual para entender as alterações sociais.

A Base Nacional Comum Curricular (BNCC), baseada nos PCNs, também propõe o desenvolvimento do pensamento computacional[5] e ressalta a aplicabilidade em trabalhos e vivências futuras.

> Um dos desafios para a aprendizagem da Matemática no Ensino Médio é exatamente proporcionar aos estudantes a visão de que ela não é um conjunto de regras e técnicas, mas faz parte de nossa cultura e de nossa história. (BNCC; 2018, p. 522)

Na BNCC, há propostas para desenvolver o trabalho com dados e a criticidade do aluno, a partir da interpretação e análise de problemas reais.

Nas seções 2.3 e 2.4 apresentamos o conteúdo proposto em cada um desses documentos para a área de Matemática, por vezes citados como temas de Estatística, Análise, Ciência de Dados e as tecnologias que podem favorecer tal trabalho. Todos esses são assuntos que fazem parte da área de Matemática.

## 2.3. Parâmetros Curriculares Nacionais

Os PCNs foram propostos no final de 1990 e aperfeiçoados até o começo dos anos 2000. Também foram um dos documentos utilizados para a elaboração da BNCC (MEC, 2018). Nos PCNs, há uma discussão sobre as missões da Educação e o estabelecimento de referências para que os alunos estudem matemática no ensino básico. Nesse sentido, são propostas unidades temáticas, dentre elas a análise de dados e o uso de tecnologias no ensino.

“A análise de dados tem sido essencial em problemas sociais e econômicos, como nas estatísticas relacionadas a saúde, populações, transportes, orçamentos e questões de mercado” (BRASIL, 2002, p. 126). Com a possibilidade de "(...)aproximar o aluno da realidade e fazê-lo vivenciar situações próximas que lhe permitam reconhecer a diversidade que o cerca e reconhecer-se como indivíduo capaz de ler e atuar nesta realidade" (BRASIL, 2002, p. 126). Além do reconhecimento, está também a análise para a exploração de possibilidades e tomada de decisões (BRASIL, 2002).

No que diz respeito ao processo auxiliado pelas tecnologias digitais, questiona-se se a inserção dos computadores e a geração de imensas quantidades de dados

[5] No documento falta comentar sobre essa forma de pensamento. O pensamento computacional, segundo Wing (2008) e Wing (2006) é um conjunto de habilidades não só acessíveis aos cientistas da computação, mas a todos. Sua principal habilidade é a abstração. Tem muitas similaridades ao pensamento matemático, de engenharias e científico.

que seriam necessários para o aprendizado de técnicas de programação. Nesse sentido, o documento aponta que:

> A experiência nesse campo envolve o conhecimento do universo dos computadores, o que não implica numa prática técnica, reservada aos profissionais da área, do mesmo modo que não é necessário saber o que acontece sob a capota de um automóvel para que nos utilizemos dele. As qualidades de um bom motorista são diversas, tais como conhecimento do código, respeito às regras elementares e uma certa competência, que lhe permite o domínio do veículo em todas as circunstâncias (BRASIL, 2000b, p. 58).

Com a revolução na vida e no trabalho, com a automação, digitalização e imersão em dados, "a escola precisa mudar, não só de conteúdos, mas aceitando novos elementos que possibilitem a integração do estudante ao mundo que o circunda" (BRASIL, 2000b, p. 61). Devido a quantidade e a variedade de informações atualmente, é importante que o cidadão saiba analisar, estabelecer relação, sintetizar e avaliar (BRASIL, 2000b, p. 61) e desenvolva

> (...) interesses e capacidades, criando condições para a sua inserção num mundo em mudança e contribuindo para desenvolver as capacidades que deles serão exigidas em sua vida social e profissional. Em um mundo onde as necessidades sociais, culturais e profissionais ganham novos contornos, todas as áreas requerem alguma competência em Matemática e a possibilidade de compreender conceitos e procedimentos matemáticos é necessária tanto para tirar conclusões e fazer argumentações, quanto para o cidadão agir como consumidor prudente ou tomar decisões em sua vida pessoal e profissional (BRASIL, 2000, p. 40).

Aprender Matemática contextualizada e relacioná-la com outros conhecimentos possibilita o desenvolvimento de competências e habilidades estruturantes ao pensamento do aluno capacitando-o a se comunicar em linguagens específicas, após análise e avaliação, viabilizando conclusões e tomadas de decisão para outras ações necessárias (BRASIL, 2002, p. 111).

Na seção 2.4, apresentamos as competências e habilidades descritas na BNCC e suas ponderações com o trabalho a partir de dados.

## 2.4. Base Nacional Comum Curricular

A BNCC é um documento normativo. Sua proposta é a orientação na elaboração de currículos sobre as áreas de **linguagens e suas tecnologias**, **matemática e suas tecnologias**, **ciências da natureza e suas tecnologias**,

**ciências humanas e sociais aplicadas** e **formação técnica e profissional** enumerando competências e habilidades relativas a cada uma.

> Na BNCC, competência é definida como a mobilização de conhecimentos (conceitos e procedimentos), habilidades (práticas, cognitivas e socioemocionais), atitudes e valores para resolver demandas complexas da vida cotidiana, do pleno exercício da cidadania e do mundo do trabalho (BNCC, 2018, p. 9).

Cada área do conhecimento apresenta competências específicas e habilidades para seu desenvolvimento. Abaixo foram listadas as dez competências gerais que se encontram descritas na BNCC, todas comentadas para mostrar possiblidades no trabalho com dados.

A primeira competência reflete sobre o melhor entendimento da sociedade através dos dados sociais, econômicos e científicos.

> Valorizar e utilizar os conhecimentos historicamente construídos sobre o mundo físico, social, cultural e digital para entender e explicar a realidade, continuar aprendendo e colaborar para a construção de uma sociedade justa, democrática e inclusiva (BNCC, 2018, p. 9).

Para a formação de uma sociedade mais justa, democrática e inclusiva é importante melhorar a criticidade dos alunos diante das mais diversas situações, com o desenvolvimento de uma postura investigativa, como proposto na segunda competência a seguir.

> Exercitar a curiosidade intelectual e recorrer à abordagem própria das ciências, incluindo a investigação, a reflexão, a análise crítica, a imaginação e a criatividade, para investigar causas, elaborar e testar hipóteses, formular e resolver problemas e criar soluções (inclusive tecnológicas) com base nos conhecimentos das diferentes áreas (BNCC, 2018, p. 9).

O desenvolvimento da criatividade para auxiliar no processo de investigação possibilita a análise de dados ir além e adentrar outras áreas. Exercícios de criatividade com dados podem ser conferidos no âmbito visual (PINHEIRO, 2018), (CANQUERINO, 2019) e no textual (LIU et al., 2018) com o desenvolvimento de inteligências artificiais, como na terceira competência.

> Valorizar e fruir as diversas manifestações artísticas e culturais, das locais às mundiais, e também participar de práticas diversificadas da produção artístico-cultural (BNCC, 2018, p. 9).

Além das Artes, a comunicação dos resultados obtidos no processo investigativo é fundamental e suscita o uso de diferentes recursos, sejam eles textuais, gráficos, matemáticos, tecnológicos etc. como citado na quarta competência.

> Utilizar diferentes linguagens – verbal (oral ou visual-motora, como Libras, e escrita), corporal, visual, sonora e digital –, bem como conhecimentos das linguagens artística, matemática e científica, para se expressar e partilhar informações, experiências, ideias e sentimentos em diferentes contextos e produzir sentidos que levem ao entendimento mútuo (BNCC, 2018, p. 9).

Nas competências anteriores, o desenvolvimento de criticidade, de habilidades investigativas e comunicativas é fundamental. Enquanto para muitos o trabalho com dados pode ser pouco compreensível, os algoritmos que trabalham com eles não são neutros, podem propagar preconceitos, ressaltar problemas em registros históricos, gerando questões a serem discutidas com a ascensão das máquinas (KAUFMAN, 2019) (PIERRO, 2018) (TUNES, 2019c). A não neutralidade dos algoritmos pode ser abordada através da quinta competência.

> Compreender, utilizar e criar tecnologias digitais de informação e comunicação de forma crítica, significativa, reflexiva e ética nas diversas práticas sociais (incluindo as escolares) para se comunicar, acessar e disseminar informações, produzir conhecimentos, resolver problemas e exercer protagonismo e autoria na vida pessoal e coletiva” (BNCC, 2018, p. 9).

Se os algoritmos de Inteligência Artificial não são neutros, então estatísticas geradas sobre questões sociais precisam ser mais bem compreendidas para o pleno desenvolvimento da cidadania. Uma temática proposta pela sexta competência.

> Valorizar a diversidade de saberes e vivências culturais e apropriar-se de conhecimentos e experiências que lhe possibilitem entender as relações próprias do mundo do trabalho e fazer escolhas alinhadas ao exercício da cidadania e ao seu projeto de vida, com liberdade, autonomia, consciência crítica e responsabilidade (BNCC, 2018, p. 9).

O cidadão tomará melhores decisões baseado em dados. Através de métodos científicos, trabalho com coleta, processamento e comunicação de informações objetivas oriundas de dados as questões argumentativas podem ser melhoradas, envolvendo questões humanas, ambientais e regionais. Assim como é proposto na sétima competência a seguir:

> Argumentar com base em fatos, dados e informações confiáveis, para formular, negociar e defender ideias, pontos de vista e decisões comuns que respeitem e promovam os direitos humanos, a consciência socioambiental e o consumo responsável em âmbito local, regional e global, com posicionamento ético em relação ao cuidado de si mesmo, dos outros e do planeta (BNCC, 2018, p. 9).

Em competências anteriores, foram destacados trabalhos com dados sociais, econômicos e artísticos além do desenvolvimento da postura crítica e comunicativa. Como a saúde física e emocional é parte fundamental de qualquer cidadão, a oitava competência pode ser o momento para discutir dados nessas temáticas.

> Conhecer-se, apreciar-se e cuidar de sua saúde física e emocional, compreendendo-se na diversidade humana e reconhecendo suas emoções e as dos outros, com autocrítica e capacidade para lidar com elas (BNCC, 2018, p. 10).

Os preconceitos, saberes distintos e identidades culturais podem ser trabalhados a partir do momento em que são (re)conhecidos. Na nona competência nota-se a proposta de quantificar os aspectos culturais e a partir disso, processá-los em diversas áreas como a Estatística e/ou Ciência de Dados.

> Exercitar a empatia, o diálogo, a resolução de conflitos e a cooperação, fazendo-se respeitar e promovendo o respeito ao outro e aos direitos humanos, com acolhimento e valorização da diversidade de indivíduos e de grupos sociais, seus saberes, identidades, culturas e potencialidades, sem preconceitos de qualquer natureza” (BNCC, 2018, p. 10).

Por fim, a última competência também abre espaço para o trabalho com dados para refletir sobre o cidadão consciente e sua participação no meio social (ENGEL, 2018). Condizente com as anteriores exploradas até o momento.

> Agir pessoal e coletivamente com autonomia, responsabilidade, flexibilidade, resiliência e determinação, tomando decisões com base em princípios éticos, democráticos, inclusivos, sustentáveis e solidários (BNCC, 2018, p. 10).

Observando as competências listadas acima e os comentários que procuram indicar como é possível abordar o trabalho com dados, notamos que a Estatística está presente em praticamente todas as abordagens. As observações feitas não visam o esgotamento da competência e aplicabilidades relacionadas a ela, mas propõem uma forma de abordagem a partir das áreas em questão neste trabalho.

### 2.4.1. O método de trabalho da Base Nacional Comum Curricular

O trabalho investigativo pode ser realizado para confirmar ou refutar diversos meios de comunicação, com notícias diárias baseadas em dados diversos sendo uma abordagem proveitosa à Estatística escolar. A forma de manipular essa massa de dados pode ser facilitada com a tecnologia como ferramenta, com o intuito de verificar o cotidiano, junto com o trabalho de esclarecimento do porquê tais métodos funcionam.

> Em lugar de pretender que os jovens apenas aprendam o que já sabemos, o mundo deve lhes ser apresentado como campo aberto para investigação e intervenção quanto a seus aspectos sociais, produtivos, ambientais e culturais (BNCC; 2018; p. 463).

Na área de Matemática e suas tecnologias o fundamental não deve ser somente na utilização de estratégias envolvendo conceitos ferramentas e processos para a resolução de problemas, mais do que isso, a possibilidade de "(...) formulá-los, descrever dados, selecionar modelos matemáticos e desenvolver o pensamento computacional, por meio da utilização de diferentes recursos da área" (BNCC; 2018, p. 470) é algo proveitoso que deve ter espaço na escola com o objetivo de desenvolver a criticidade nos seus estudantes.

Com o intuito de proporcionar momentos não focados somente nas disciplinas e em suas especificidades, a BNCC (2018) propõe situação de trabalhos mais colaborativos e fornece alguns exemplos, sendo que dentre eles há as *oficinas* promovendo produção e tratamento estatístico na articulação entre teorias e práticas. Além dos *Núcleos de estudos* trabalhando com pesquisas e momentos de debates, disseminando conhecimentos por meio de eventos.

### 2.4.2. Matemática e suas Tecnologias

Segundo a BNCC (BRASIL, 2018) as unidades de conhecimento, ou campos, são definidas como **Números, Álgebra, Geometria, Grandezas e Medidas, Probabilidade e Estatística**. Para estabelecer melhores conexões com conhecimentos anteriores, o documento propõe a ampliação e aprofundamento do que foi desenvolvido até o 9º ano do ensino fundamental. O uso da tecnologia seja

na utilização de calculadoras ou planilhas eletrônicas é incentivado desde os anos iniciais do ensino fundamental para que nos anos finais, os alunos tenham a possibilidade de serem estimulados ao desenvolvimento do pensamento computacional.

No ensino fundamental, os alunos têm a possibilidade de aprender em Probabilidade os seguintes tópicos: espaço amostral de eventos equiprováveis, princípio multiplicativo ou simulações para estimar a probabilidade relacionada a um evento. Referente à Estatística, os estudos não ficam limitados à interpretação das informações divulgadas pela mídia, mas além disso, planejamento e execução de pesquisa amostral, trabalhando com medidas de tendência central (média, moda e mediana) e de dispersão (amplitude, variância e desvio padrão) e comunicando os resultados por meio de relatórios com gráficos adequados. (BNCC; 2018)

Portanto "(...) os estudantes devem desenvolver habilidades relativas aos processos de investigação, de construção de modelos e de resolução de problemas" (BNCC; 2018, p. 519). Todas descritas em cinco competências, escritas e comentadas no começo das seções a seguir.

Nas subseções 2.4.2.1 a 2.4.2.5 discutiremos essas competências e algumas das habilidades relacionadas ao ensino de Estatística e de Probabilidade. Cabe notar que as questões algébricas na Estatística, como modelos lineares, transformações de escala etc., não foram todas incluídas no texto, pois, ainda que possíveis de serem estudadas, elas não são o foco do presente trabalho.

#### 2.4.2.1. Competência específica 1

> Utilizar estratégias, conceitos e procedimentos matemáticos para interpretar situações em diversos contextos, sejam atividades cotidianas, sejam fatos das Ciências da Natureza e Humanas, ou ainda questões econômicas ou tecnológicas, divulgados por diferentes meios, de modo a consolidar uma formação científica geral *(BNCC, 2018, 523).*

Investigar e estabelecer conjecturas a respeito de diferentes conceitos e propriedades matemáticas, empregando recursos e estratégias como observação de padrões, experimentações e tecnologias digitais, identificando a necessidade, ou não, de uma demonstração cada vez mais formal na validação das referidas conjecturas.

O desenvolvimento da interpretação e compreensão de informações originadas a partir de diversos meios de comunicação é essencial para tal competência, através

do aprimoramento de criticismo e reflexão. Há questões a serem trabalhadas, como amostragens e formas de generalização adotadas em pesquisas estatísticas e, com isso, se apropriar das informações sobre economia, sociedade, outras ciências etc. (BNCC; 2018).

Das cinco habilidades relacionadas a esta competência, duas são destacadas pela proposta no trabalho com Estatística e Ciência de Dados, sendo uma delas a EM13MAT102[6] que destaca habilidades relacionadas a análise de gráficos e métodos de amostragem, identificando inadequações que possam induzir ao erro. Já a EM13MAT104 se refere ao conhecimento sobre taxas e índices com o objetivo de investigar o processo pelo qual são definidos.

#### 2.4.2.2. Competência específica 2

> Articular conhecimentos matemáticos ao propor e/ou participar de ações para investigar desafios do mundo contemporâneo e tomar decisões éticas e socialmente responsáveis, com base na análise de problemas de urgência social, como os voltados a situações de saúde, sustentabilidade, das implicações da tecnologia no mundo do trabalho, entre outros, recorrendo a conceitos, procedimentos e linguagens próprios da Matemática *(BNCC, 2018, 523).*

Além do processo interpretativo da competência anterior, esta se refere ao planejamento de pesquisas, buscando estabelecer métodos para a execução do projeto pretendido, com reflexão sobre o processo no viés matemático, com colaboração entre estudantes e professores.

> Ela deve também fornecer condições para o planejamento e execução de pesquisas, identificando aspectos consensuais ou não na discussão de projetos, com base em princípios solidários, éticos e sustentáveis, valorizando a diversidade de opiniões de grupos sociais e de indivíduos e sem quaisquer preconceitos. (BNCC; 2018, p. 526)

As habilidades específicas pertinentes a Estatística e CD são a EM13MAT202 sobre a amostragem e formas de comunicação dos resultados de uma análise estatística e EM13MAT203 que é relacionada ao uso de tecnologias digitais para aplicar conhecimentos matemáticos e tomar decisões a partir deles.

[6] Para entender esse código veja que EM significa Ensino Médio. O número 13 se refere a faixa etária para o trabalho de tal habilidade. A sigla MAT se refere a Matemática. O número 101 diz que é a primeira referente a competência 1, no caso da segunda o número seria 102 e assim por diante.

#### 2.4.2.3. Competência específica 3

> Utilizar estratégias, conceitos e procedimentos matemáticos, em seus campos – Aritmética, Álgebra, Grandezas e Medidas, Geometria, Probabilidade e Estatística –para interpretar, construir modelos e resolver problemas em diversos contextos, analisando a plausibilidade dos resultados e a adequação das soluções propostas, de modo a construir argumentação consistente (BNCC, 2018, p. 523).

De forma a complementar as duas anteriores, essa ressalta as questões do processo, pois

> No Ensino Médio, os estudantes devem desenvolver e mobilizar habilidades que servirão para resolver problemas ao longo de sua vida; por isso, as situações propostas devem ter significado real para eles (BNCC; 2018, p. 527).

O “como fazer” está relacionado mais do que somente na questão tecnicista, na qual um problema é proposto e com ele já se sabe qual conteúdo será utilizado. “Esse processo envolve analisar os fundamentos e propriedades de modelos existentes, avaliando seu alcance e validade para o problema em foco” (BNCC; 2018, p. 527). E a tecnologia é citada como possibilidade para melhora na participação ativa dos estudantes no processo de resolução de problemas.

Para o trabalho com dados há as habilidades EM13MAT303 no trabalho com porcentagem, EM13MAT315 sobre o trabalho com problemas que podem ser expressos por meio de algoritmos, EM13MAT316 sobre o uso de medidas de posição (média, moda e mediana) e de variabilidade (amplitude, variância e desvio padrão).

#### 2.4.2.4. Competência específica 4

> Compreender e utilizar, com flexibilidade e fluidez, diferentes registros de representação matemáticos (algébrico, geométrico, estatístico, computacional etc.), na busca de solução e comunicação de resultados de problemas, de modo a favorecer a construção e o desenvolvimento do raciocínio matemático (BNCC, 2018, p. 523).

Tal competência é desenvolvida a partir da exploração de diferentes representações de um mesmo objeto matemático. Na BNCC (2018) é essencial que haja o uso de pelo menos duas formas para cada tópico abordado, de modo a incentivar os estudantes na escolha da forma mais conveniente para cada situação.

A habilidade EM13MAT406 se refere ao uso de linguagens de programação, a EM13MAT408 ressalta o uso de *software*s estatísticos que possibilitem a construção de tabelas e gráficos de dados coletados em pesquisas e, por fim, a EM13MAT409 sugere o uso de gráficos diversos, buscando aquele mais adequado à análise.

#### 2.4.2.5. Competência específica 5

> Investigar e estabelecer conjecturas a respeito de diferentes conceitos e propriedades matemáticas, empregando recursos e estratégias como observação de padrões, experimentações e tecnologias digitais, identificando a necessidade, ou não, de uma demonstração cada vez mais formal na validação das referidas conjecturas (BNCC, 2018, p. 523).

Esta última faz referência a investigação, exploração de conjecturas, experimentação, presença de tecnologias digitas e a verificação quanto a necessidade de se refinar demonstrações para a validação do processo.

> As habilidades vinculadas a essa competência assumem um importante papel na formação matemática dos estudantes que, mediante investigações, devem formular conjecturas, refutá-las ou validá-las e comunicar com precisão suas conclusões (BNCC; 2018, p. 532).

Por ser relacionada a buscar e questionar, a experimentação se utiliza de materiais concretos, apoios visuais e tecnologias digitais na busca de contraexemplos, ou argumentos para validação de hipóteses.

Dentre as habilidades, destaca-se a EM13MAT510 que propõe o trabalho com duas variáveis numéricas e o uso de uma reta para descrever uma possível relação, introduzindo a possibilidade da discussão sobre modelagem dos dados.

Em resumo, a Estatística está presente na apresentação do material da BNCC, nem sempre de modo explícito, mas subjacente às práticas propostas.

A elaboração de um plano de ação voltado à maioria da população passa pela Educação Básica e pode ser impulsionada pelo uso de Tecnologias Digitais (TD). A obtenção dos dados, o processamento e a criação de recursos visuais para a comunicação são facilitados com as TD, possibilitando que os "(...) estudantes se concentrem mais no planejamento da investigação, criação e interpretação da representação dos dados, construção de modelos e tomadas de decisão" (BEN-ZVI; 2017, p.33, *tradução nossa*) propondo uma pedagogia baseada em modelagem ou em projetos.

Ensinar como examinar os próprios dados de forma construtiva pode promover o aprendizado de habilidades relacionadas a dados, de tal forma que as pessoas consigam utilizar aqueles que lhes são fornecidos ao seu redor e que também sejam relevantes para as próprias vidas (WILD; UTTS; HORTON, 2018). "Professores de estatística poderiam enfatizar a utilidade da estatística quando ensinassem, de forma direcionada, a sua audiência" (WILD; UTTS; HORTON, 2018, p. 17, *tradução nossa*).

"Se considerarmos que a educação estatística precisa incluir novas perspectivas e uso de dados, a ideia do *Big Data* deveria ser abordada na escola" (FRANÇOIS; MONTEIRO, 2018, p. 2, *tradução nossa*, grifo próprio) possibilitando o desenvolvimento consciente e a melhora da participação cidadã de cada um.

No ensino de Estatística é conhecido e incentivado o trabalho com dados, de forma altamente experimental, sendo praticamente um processo artístico resultado de diversas habilidades. Tais processos são facilitados com o uso das tecnologias, pois o foco pode ser direcionado para a compreensão de conceitos e melhoria do letramento em dados, ao invés de um aprendizado segmentado em processos, abordagens mecanicistas e ferramentas (BEN-ZVI, 2017).

Visto que normalmente o aprendizado da estatística recai sobre casos isolados, técnicas pouco significativas com baixa relevância, entediantes e mecânicas (BEN-ZVI, 2017) (BEN-ZVI; FRIEDLANDER, 1996), a adição de tecnologias digitais e áreas como a Ciência de Dados podem promover uma melhora no ensino e aprendizagem. Para explorar essas melhorias, TD no ensino de Estatística podem ser mais frequentemente abordadas e embasam a discussão do Capítulo 3 sobre artigos ensinando conceitos básicos de estatística apoiados em *software* para facilitar a prática.

## 2.5. Educação Estatística e Tecnologias Digitais

Dentre as mudanças contemporâneas da estatística, Biehler (2019) ressalta a Análise Exploratória de Dados proposta por J. Tukey, facilitada pelo uso do computador, por causa da simplicidade em explorar simulações, modelos mais complexos e na facilidade de uso de dados reais. Entretanto, para evitar uma simples exploração de programas ou linguagens de programação, com a seleção da ferramenta apropriada, o foco será no aprendizado estatístico.

Os *software*s em estatística possuem uma vasta história, dentre alguns destaques há a linguagem S mostrando algumas preocupações bastante atuais, pois os desenvolvedores dela queriam facilidades em implementar a própria linguagem e usar os conceitos da análise exploratória de dados (BECKER, 1994), posteriormente, outros *softwares* propuseram mais do que a manipulação via linha de comandos, também trouxeram uma interface gráfica, possibilitando a facilidade do trabalho e ensino estatísticos. Por exemplo o CODAP, Tinkerplot, Fathom, R e Python (BIEHLER, 2019).

Essas tecnologias mudaram as formas de trabalho, pesquisa, ensino e aprendizagem em estatística pelas formas distintas de acesso, exploração e visualização de dados, automação dos seus cálculos, permitindo a ilustração de conceitos abstratos com a possibilidade de explorar além das fórmulas e suas manipulações.

> "(...) milhares de estudantes do ensino médio ainda usam calculadoras mais do que computadores para suas análises, limitando suas habilidades em ir além do simples cálculo, ou ganhar qualquer significado real do método de trabalho que eles provavelmente encontrarão no mundo real" (WILD; UTTS; HORTON, 2018, p. 31, *tradução nossa*).

Entretanto, a variedade de recursos e possibilidades já é explorada há anos, como é o caso no trabalho de Tukey (1965), o qual imaginava as vantagens do uso do computador para a análise de dados. Haveria possibilidade na melhoria do trabalho, com o aumento do acesso a diversos tipos de dados (volume e variedade) e o processamento mais rápido do que o habitual (velocidade). Muito similar ao que hoje é o *Big Data*. Outra possibilidade técnica apresentada por Tukey (1965) é a facilidade de geração de gráficos que deviam ser mais bem explorados.

O uso de *software* para o ensino e aprendizagem de estatística continua até hoje, afinal, “(...) ferramentas modelam a forma com que vemos o mundo, e ferramentas de computação estatística não são exceção” (McNAMARA, 2018, p. 2, *tradução nossa*).

Contudo, na perspectiva de *software* como ferramentas, há aquelas mais apropriadas para algumas situações. Por isso, McNamara (2018) propõe a categorização em *software* para o aprender estatístico ou o fazer estatístico.

Aqueles destinados ao aprender estatística possuem um começo bastante simplificado, com algumas possibilidades de representações e análises mais simplificadas, porém são limitados no quesito variedade de análises. Por outro lado, os destinados para o fazer estatístico possuem um sistema mais desafiador para novos usuários, mas sua capacidade de processamento e possibilidade de análises é gigantesca. Como colocar os dez critérios aqui no meio?

Por fim, McNamara (2018) elenca dez critérios para avaliar a possibilidade de uso de um *software* para o aprendizado estatístico, e Biehler (2019) utiliza esses itens para avaliar o Tinkerplot, o Fathom e o CODAP. Esse último será utilizado para exibir algumas práticas com Análise Exploratória de Dados na Educação Básica, devido a suas diversas facilidades para exploração de gráficos, ser gratuito, de código aberto e possuir versão em português.

Portanto, no próximo capítulo, serão revisados trabalhos que utilizaram o CODAP para promover o estudo da Estatística em nível escolar. Com isso, serão observadas as possibilidades da Análise Exploratória de Dados e até a exploração de um algoritmo utilizado na área de Inteligência Artificial.

# 3. CODAP – um *software* para o ensino da Estatística

"No mundo repleto de dados, todas as pessoas instruídas precisam entender ideias estatísticas e conclusões, para enriquecer sua vida profissional e pessoal" (WILD; UTTS; HORTON, 2018, p. 16, *tradução nossa*) e ao conhecer o mundo digital e suas possibilidades é factível gerar, processar e compreender tantas informações emergentes.

Entre os diversos recursos disponíveis para tratar dados, destacamos neste trabalho o *software Common Online Data Analysis Plataform*[7] (*CODAP)* por ser *online,* ser de código aberto e possibilitar a Análise Exploratória de Dados de maneira interativa**,** com o mínimo de instrução prévia. Seu desenvolvimento ocorreu a partir de outros dois programas utilizados no ensino de Estatística: Fathom e Tinkerplots (BIEHLER, 2019) mas**,** diferente deles, tem a proposta de ser gratuito (CODAP, 2020).

Com ênfase no trabalho com dados, o CODAP trabalha com quatro formatos[8] e é possível o manuseio simples e rápido para obter as medidas de posição e variação através de gráficos ou mesmo de fórmulas. Também há a possibilidade de extensão do ambiente através de *plugins* desenvolvidos por outras pessoas, como é o caso do *Sampler*, o qual possui algumas possibilidades, dentre elas a de fazer amostragem aleatória em uma dada base de dados. O *software* possui versão em português atualizada pelo autor deste texto (CODAP, 2023).

Neste capítulo, algumas possibilidades com o CODAP na Educação Básica serão exibidas (seção 3.1 e 3.2) e, depois, ele será utilizado para apresentar duas Análises Exploratórias de Dados (seção 3.3 e seção 3.4), além de auxiliar na ilustração do algoritmo da Árvore de Decisão (seção 3.5).

Exemplos ou tutoriais sobre o funcionamento do CODAP podem ser vistos em Caem IME USP (2020a) (2020b) (em português), ou *Enhancing Statistics Teacher Education with E-Modules* (2018) (com legendas em português).

## 3.1. O CODAP e a Análise Exploratória de Dados

O CODAP foi utilizado em pesquisas para auxiliar na Análise Exploratória de Dados (BONANGELO; CORDANI, 2022) (ENGEL, 2018), introduzir conceitos de

[7] https://codap.concord.org/app/
[8] .codap, .json, .csv, .txt

Ciência de Dados para alunos da Educação Básica (BUDDE et al., 2020), (FRISCHEMEIER et al., 2021) e na abordagem inicial de um algoritmo de Inteligência Artificial (BIEHLER; FLEISCHER, 2021).

A Análise Exploratória de Dados proposta por Tukey (1977) trouxe um foco na exploração de gráficos e possibilita o trabalho com conjunto de dados multivariados, além da interatividade facilitada pelo uso de computadores (BIEHLER, 2018). O CODAP é uma ferramenta utilizada para essas possibilidades e facilidades citadas, como é o caso do trabalho desenvolvido por Bonangelo e Cordani (2022).

Bonangelo e Cordani (2022) relatam um minicurso *online*, com duração de quatro horas, oferecido de forma síncrona em um Centro de Aperfeiçoamento ao Ensino de Matemática no Instituto de Matemática e Estatística da Universidade de São Paulo. Nele foram coletadas, em tempo real, 44 observações dos participantes, com o objetivo de explorar o questionamento: “Caso fabricássemos luvas, seria necessário a diferenciação por gênero dos tamanhos pequeno, médio e grande?”. Tais dados foram analisados e ilustraram o uso do Gráfico de Pontos (*Dotplot*) e do Gráfico de Caixa (*Boxplot*) na comparação entre grupos.

O CODAP facilitou a criação dos gráficos, a exploração das medidas resumo, possibilitou a interação entre os dados brutos e os pontos presentes nos gráficos. Além da possibilidade de todas as análises feitas poderem ser compartilhadas através de um *link* disponibilizado na hora do curso, permitindo o estudo individual de todos os participantes em seus próprios dispositivos sem alterar o trabalho de outros. Os detalhes sobre a Análise Exploratória dos Dados serão detalhados na seção 3.3.

Como parte do curso “Promovendo Engajamento Cívico Através da Exploração de Evidências” (*Tradução nossa*), ou sua sigla em inglês: ProCivicStat[9], Engel (2018) descreve três atividades práticas desenvolvidas com o auxílio do CODAP para as análises de dados. As três atividades são intituladas (*tradução nossa*):

1) “Alguns tão ricos outros tão pobres – Distribuição de Renda na Europa” (p. 3)
2) “Como podemos descrever a situação da população mundial?” (p. 4)
3) “Os juízes no futebol europeu são racialmente tendenciosos?” (p. 6)

Após comentar cada uma das três atividades, explorando questionamentos, possibilidades de gráficos para ilustrar o processo e análises a serem realizadas,

[9] http://www.iase-web.org/islp/pcs/

Engel (2018) destaca a facilidade de uso do CODAP por parte dos participantes, os quais não precisaram de muitas instruções prévias para usarem o *software* e a facilidade em alguns processos, como a comparação entre distribuições, a investigação e comparação entre grupos etc.

Por fim, Budde et al. (2020) propõem uma atividade interdisciplinar (Educação Estatística e Ciência da Computação) para 14 alunos da Educação Básica (17-18 anos), com o intuito de desenvolver o raciocínio estatístico e, posteriormente, desenvolverem um curso de Ciência de Dados para os estudantes. A base de dados utilizada foi obtida a partir do estudo JIM[10] com informações sobre tempo de lazer e em mídias sociais de estudantes alemães de 12 a 19 anos, incluindo mais de 80 questões com resposta qualitativa. O CODAP foi utilizado para auxiliar na análise de dados com as diferentes representações, facilitando a comparação entre grupos, exploração da relação entre duas variáveis qualitativas, exibição de frequências absoluta e relativa etc., sempre com o objetivo de responder as questões propostas pelos pesquisadores.

Os autores ressaltam que o “CODAP serviu como uma valiosa ferramenta digital para uma primeira exploração de dados” (BUDDE et al., 2020, p. 6) assim como a recepção positiva do CODAP pela maioria dos estudantes, devido a suas facilidades de uso.

## 3.2. O CODAP e a Árvore de Decisão

Biehler et al. (2020) propuseram uma outra forma para ir além na Análise Exploratória de Dados de bases multivariadas. Através da introdução da Árvore de Decisão, eles utilizaram as facilidades do CODAP e fizeram uma proposta, para alunos dos últimos anos do ensino médio, adentrarem a linguagem de programação Python para melhorarem suas análises estatísticas.

A Árvore de Decisão é um algoritmo utilizado na Estatística e na Ciência de Dados, devido a facilidade de entendimento do seu funcionamento e o poder de predição a partir de algumas informações prévias. O que é esse algoritmo, como funciona, objetivos de uso e quais a ideias que o embasam serão discutidas e

[10] https://www.mpfs.de/startseite/

exemplificadas na seção 3.4. A Árvore de Decisão será abordada a partir de conteúdos vistos nas aulas de Estatística na Educação Básica.

No trabalho de Biehler et al. (2020) os autores utilizaram a base de dados do estudo JIM com informações sobre tempo de lazer e em mídias sociais de estudantes alemães de 12 a 19 anos. Após uma redução na quantidade de variáveis de mais de 80 para somente 15, propuseram algumas atividades aos alunos. Uma atividade foi tentar prever a frequência que as pessoas jogavam jogos *online* a partir das variáveis disponíveis, montando manualmente uma Árvore de Decisão. A atividade foi produtiva, pois permitiu que os alunos expusessem suas reflexões e analisassem valores para verificar a qualidade dos algoritmos montados por eles. Entretanto, a Árvore de Decisão é feita de forma manual no CODAP e com algumas restrições, por isso utilizaram a linguagem de programação Python para automatizar o processo e verificar mais possibilidades do algoritmo.

## 3.3. Uma breve análise descritiva de dados no CODAP

O CODAP será apresentado através de uma breve análise estatística de uma base de dados obtida em um minicurso oferecido de modo online no Centro de Aperfeiçoamento ao Ensino de Matemática (CAEM IME USP, 2020a) (CAEM IME USP, 2020b). Outros detalhes podem ser conferidos em Bonangelo e Cordani (2021).

Neste minicurso, foi solicitado aos inscritos que medissem a palma da mão direita e que enviassem o valor de modo *online*, de forma que pudessem ser analisados em tempo real pelo autor desta dissertação. Outras informações solicitadas foram idade em anos completos e gênero.

O conjunto recebido totalizou 44 observações referentes aos participantes do curso, cada um com as seguintes informações: Tamanho do palmo da mão em centímetros (cm), idade em anos e gênero, sendo que as duas primeiras variáveis são quantitativas e a terceira é qualitativa (Feminino, Masculino e Outro).

Referente às variáveis, duas observações devem ser feitas:

1) relativo a gênero, foram obtidas respostas somente das categorias masculino e feminino;
2) a medida do palmo da mão foi obtida de forma que a mão estivesse totalmente espalmada;

3) para os tamanhos medidos, foi utilizado um método de arredondamento para que valores com a parte decimal menor, ou igual a quatro décimos truncassem (por exemplo: 19,3cm seria 19cm) e, caso contrário, o valor seria o maior e mais próximo inteiro (23,8cm seria 24cm, por exemplo). Desta forma, a abordagem de construção de gráficos e de medidas é facilitada.

Nesta oficina virtual, em que a análise seria feita com computador, a inclusão e uso de decimais é simples. Contudo, a proposta aos professores também era mostrar a facilidade de replicar o experimento em sala de aula de forma totalmente analógica. Por isso, o arredondamento possibilitaria a ênfase que se desejava atribuir ao raciocínio estatístico e não ao trabalho com medidas até a primeira casa decimal. É também um momento em que se pode haver uma discussão sobre arredondamento, caso o professor ache pertinente.

Toda a coleta de dados só tem sentido se tiver como propósito responder a uma pergunta. Neste caso, o questionamento poderia ser: “ao fabricar luvas, temos que levar em conta o gênero ou não?” e, através de uma coleta de dados, seria investigado a distinção entre o tamanho da mão dos dois gêneros.

Uma análise estatística para variáveis quantitativas geralmente é iniciada com cálculos de medidas de posição e de variação, bem como com gráficos informativos do comportamento dos dados coletados. Essas análises pertencem ao que se denomina estatística descritiva, que reúne, portanto, medidas resumo (posição e variação) e gráficos. Para a análise com CODAP vamos explorar inicialmente as ferramentas gráficas para o conjunto dos dados coletados.

O Gráfico de Pontos (*Dotplot*) é bastante elucidativo para o comportamento de dados quantitativos e é fácil de ser construído. No CODAP, com os dados coletados, sem qualquer distinção entre grupos, o *Dotplot* é apresentado na Figura 2.

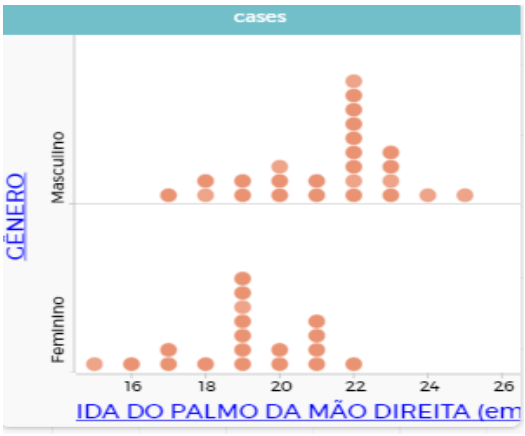
**Figura 2** - *Dotplot* da Medida do Palmo da Mão Direita


Fonte: Bonangelo (2023)

O Gráfico de Pontos é uma representação univariada e cada ponto representa o tamanho do palmo da mão de uma pessoa. Rapidamente, se percebe que a menor medida de mão é de 15 cm, enquanto a maior é de 25cm, gerando, portanto, uma primeira medida de variabilidade, a amplitude (= máximo – mínimo), cujo valor é 10 cm. Para esta visualização conjunta de ambos os gêneros, há indícios de maior concentração de valores observados entre 19cm e 22cm.

Após a separação dos grupos (Figura 3), observou-se que a menor medida do palmo da mão (mínimo) e a maior (máximo) das mulheres é de 15cm e 22cm, respectivamente. Para os homens, o mínimo é de 17cm e o máximo é de 25cm. Isso indica que a amplitude feminina é 7cm e a masculina é 8cm (variação maior entre o grupo masculino). O grupo masculino apresenta valores globalmente superiores e o valor mais frequente em cada um dos grupos, a Moda, que é uma medida de posição, reforça o comentário, pois o valor mais frequente nas mulheres é de 19cm (7 valores), enquanto no dos homens é de 22cm (9 valores).

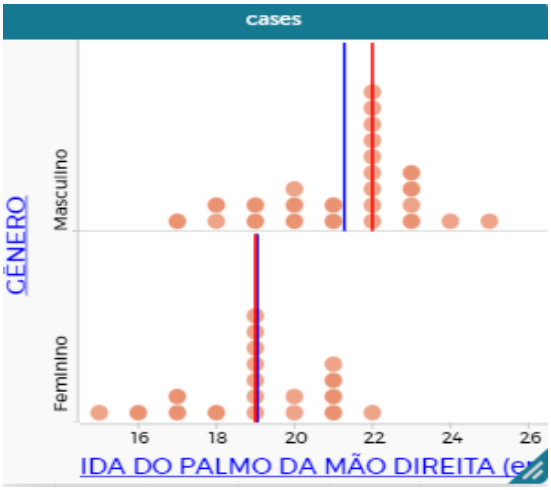
**Figura 3** - Gráfico de Pontos dos grupos feminino e masculino




Fonte: Captura de tela do CODAP. (BONANGELO (2023))

Além da Moda obtida por observação visual no *Dotplot*, o CODAP possibilita a adição de outras medidas de posição no gráfico, como é o caso da média (em azul) e da mediana (em vermelho) (Figura 4).

**Figura 4** – Média (azul) e mediana (vermelho) adicionadas ao Gráfico de Pontos




Fonte: Bonangelo (2023)

A média e mediana masculinas são de 21,3cm e 22cm, respectivamente, enquanto para o feminino são 19,05cm e 19cm. Portanto, há várias evidências empíricas que reforçam a sugestão de que o tamanho da mão dos homens é em geral maior do que a das mulheres. Pelo Gráfico de Pontos, é imediata a observação de que o grupo masculino se encontra geralmente à direita do feminino.

Outro gráfico que é importante ser construído numa análise descritiva de dados é o *Boxplot* (Gráfico de Caixa), que será gerado no CODAP juntamente com o *Dotplot,* a média e a mediana. Vale ressaltar que o Gráfico de Caixa ainda não está frequentemente disponível nos textos da escola básica, devido a sua ausência no PCN e inserção na BNCC.

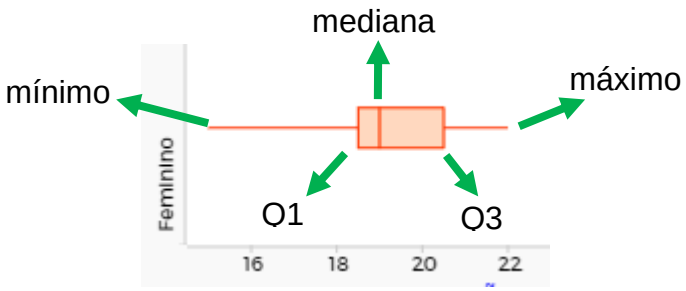
Assim como o Gráfico de Pontos, o *Boxplot* é também de natureza univariada e permite a comparação dos valores da variável “tamanho do palmo da mão” estratificando pela variável “gênero”, ou seja, separando dois grupos: o estrato feminino e o masculino.

A facilidade de construção desse gráfico é que ele necessita somente de cinco medidas de posição determinadas a partir das observações da variável, sendo elas: Mínimo, 1º Quartil (Q1), Mediana, 3º Quartil (Q3) e Máximo.

Sua construção é feita através de uma caixa e duas hastes. A caixa possui arestas limitadas pelo Q1 e o Q3 (o que delimita os 50% centrais dos dados observados). As hastes de cada lado das arestas vão até os valores extremos (Mínimo e Máximo) no caso da versão simplificada do *Boxplot*, proporcionando a informação sobre a amplitude, medida de variação já vista no Gráfico de Pontos.

Uma outra medida de variabilidade decorrente da construção do *Boxplot* é o chamado Intervalo Interquartil (IIQ) que é a diferença entre o 3º e o 1º quartis, facilmente identificável no gráfico uma vez que corresponde ao tamanho da caixa. O traço vertical dentro da caixa corresponde ao valor da mediana (Figura 5).

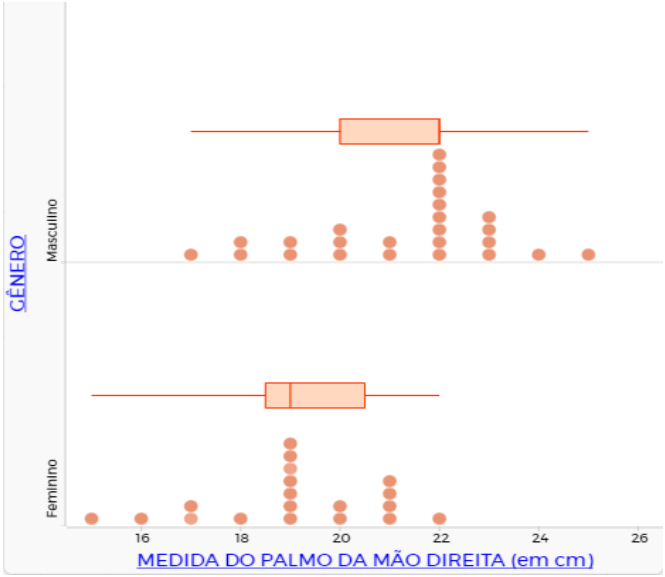
**Figura 5** – *Boxplot* no CODAP




Fonte: Bonangelo (2023)

Com o objetivo de enfatizar os gráficos de caixa e de pontos, foram retiradas as linhas azul e vermelha adicionadas anteriormente na Figura 3, resultando no que pode ser visto na Figura 6.

**Figura 6** – *Boxplot* e Gráfico de Pontos da medida do palmo das mãos separados em grupos




Fonte: Bonangelo (2023)

O IIQ, ou seja, o tamanho das caixas de 2cm para ambos os grupos, indica que os 50% centrais masculinos variam de 20 a 22cm, enquanto o feminino de 18,5 a 20,5, reforçando a premissa da medida masculina ser globalmente superior à feminina. Agora, é importante analisar as partes da caixa para verificar a dispersão de ambos os estratos.

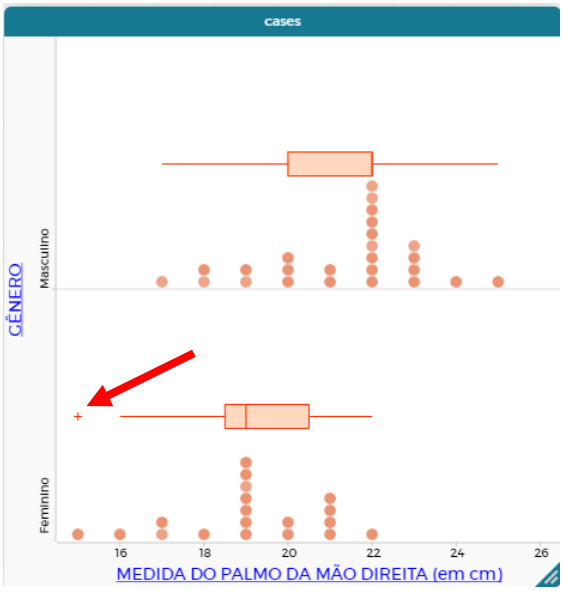
No grupo masculino, Q3 e a Mediana coincidem em 22cm (mesmo valor da Moda), mostrando grande concentração de observações neste valor. Enquanto Q1 se distancia 2cm da mediana, sugerindo maior dispersão de observações neste intervalo e caracterizando alguma assimetria desta variável para o grupo masculino. Tal situação se repete no feminino.

No grupo feminino a diferença entre Q1 e a Mediana (0,5cm) é bem menor do que entre a Mediana e o Q3 (1,5cm), indicando maior concentração de mulheres no intervalo à esquerda da Mediana, o que qualifica uma certa assimetria desta variável para o estrato feminino, ao mesmo tempo que as outras estão mais dispersas entre a Mediana e Q3.

Por fim, as hastes de ambas as caixas, que vão das arestas até os valores mínimos e máximos na versão simplificada do *Boxplot,* descrevem os 50% não centrais, os 25% com as menores medidas e os 25% com as maiores. Entretanto, nessa situação, suas representações são fortemente influenciadas por valores discrepantes, ou seja, se por acaso alguém digitasse tamanhos da palma da mão de 40cm, ou 3cm, geraria um problema de visualização e interpretação do gráfico, por causa da distorção da caixa que tais dados poderiam ocasionar.

Para verificar se há valores discrepantes no conjunto de dados, o *Boxplot* pode ser aprimorado para sua versão original. Estes dados, se existirem, são usualmente chamados de *outliers*, na versão em inglês. No CODAP é feita a representação de tais pontos por um sinal no gráfico representado por + (Figura 7). Antes de prosseguir, note que o *Boxplot* formado no grupo masculino na Figura 7 não apresenta o traço da mediana dentro da caixa, isto ocorreu porque a mediana coincidiu com o Q3 devido à alta concentração de pontos em um determinado valor (22cm nesse caso).

**Figura 7** - Pontos discrepantes no *Boxplot*




Fonte: Bonangelo (2023)

Tal marcação dos pontos é determinada através da medida de variação IIQ, ou seja, o tamanho da caixa. Assim, um ponto é considerado discrepante caso ele esteja além do final de cada haste. Para ser *outlier* inferior (aquém da haste da esquerda) o valor discrepante deverá ser menor do que “Q1-1,5IIQ” (com valores de 15,5cm e

17cm para os grupos femininos e masculinos, respectivamente). Para ser *outlier* superior (além da haste da direita) o valor discrepante deverá ser maior do que “Q3+1,5IIQ” (com valores de 23,5cm e 25cm para os grupos femininos e masculinos, respectivamente). A Figura 8 obtida a partir do grupo feminino exibe um valor discrepante à esquerda observado a partir do sinal de + na observação 15cm (este valor é inferior ao obtido para Q1-1,5 IIQ (de 15,5cm).

**Figura 8** - Ilustração dos limites das hastes no *Boxplot* original


Fonte: Bonangelo (2021)

As medidas até aqui obtidas podem ser organizadas na Tabela 1, elaborada com informações do CODAP sem a necessidade de programar nenhuma linha de código. Foi suficiente o uso de poucas operações para obtê-las.

**Tabela 1** - Análise Descritiva do Tamanho do Palmo da Mão (por Gênero)

| | Feminino (cm) | Masculino (cm) |
|---|---|---|
| Média | 19,05 | 21,28 |
| Mínimo | 15,00 | 17,00 |
| Q1 | 18,50 | 20,00 |
| Mediana | 19,00 | 22,00 |
| Q3 | 20,50 | 22,00 |
| Máximo | 22,00 | 25,00 |
| Amplitude | 7,00 | 8,00 |
| Intervalo Interquartílico (IIQ) | 2,00 | 2,00 |
| Há pontos discrepantes? | Sim | Não |

Fonte: Bonangelo (2023)

Ao que tudo indica, na amostra trabalhada, o grupo masculino apresenta globalmente o tamanho da mão maior do que o grupo feminino, ideia reforçada pelas medidas descritivas de posição, com variabilidade similar em ambos os grupos, ou seja, caso houvesse a pretensão de fabricação de luvas, então seria necessário levar em conta o gênero, assim como também proposto em Cordani e Fontes (2018). Nesse estudo também foi explorada a mesma questão, porém com uma amostra (n=288) diferente, composta por 164 medidas do gênero feminino e 124 do gênero masculino.

As medidas em centímetros de modelos Pequeno (P), Médio (M) e Grande (G) são sugeridas na Tabela 2, segundo a amostra analisada neste capítulo. Com o tamanho M produzido em maior quantidade e os P e G em menor.

**Tabela 2** – Proposta de tamanho de luvas

| Modelo | Medida para o grupo Feminino (em cm) | Medida para o grupo Masculino (em cm) |
|---|---|---|
| P | 15 a 18 | 17 a 20 |
| M | 18 a 20 | 20 a 22 |
| G | 21 a 22 | 22 a 25 |

Fonte: Bonangelo (2023)

Tais valores se justificam pelas informações discutidas a partir dos gráficos analisados até o momento. Note que os modelos de luva de tamanho P tanto para grupo feminino quanto para o masculino correspondem aos valores 25% menores, representados pela haste esquerda dos *Boxplots*. Os tamanhos M são os 50% centrais representados pela caixa e os tamanhos G são os 25% representados pelas hastes a direita, que são os maiores valores dessa amostra.

A conclusão é a mesma ao utilizar o Gráfico de Pontos e o Gráfico de Caixa para a comparação dos grupos, com o último oferecendo mais informações resumidas do que o primeiro.

Há outras possibilidades de análise exploratória nesta base de dados, mas o intuito não é esgotá-las e sim verificar possíveis procedimentos utilizando os recursos do CODAP. Observa-se que até aqui o processo foi bastante simplificado, contudo, suficiente para realizar uma análise exploratória de dados, comparando diversos gráficos e representações no mesmo espaço visual, o que demonstra uma qualidade do *software*.

A partir do CODAP é possível realizar outras análises exploratórias e estudar outros tópicos em estatística, além da possibilidade de introduzir conceitos básicos sobre alguns tópicos da Inteligência Artificial ou Ciência de Dados. Um exemplo de algoritmo que pode ser estudado é a Árvore de Decisão.

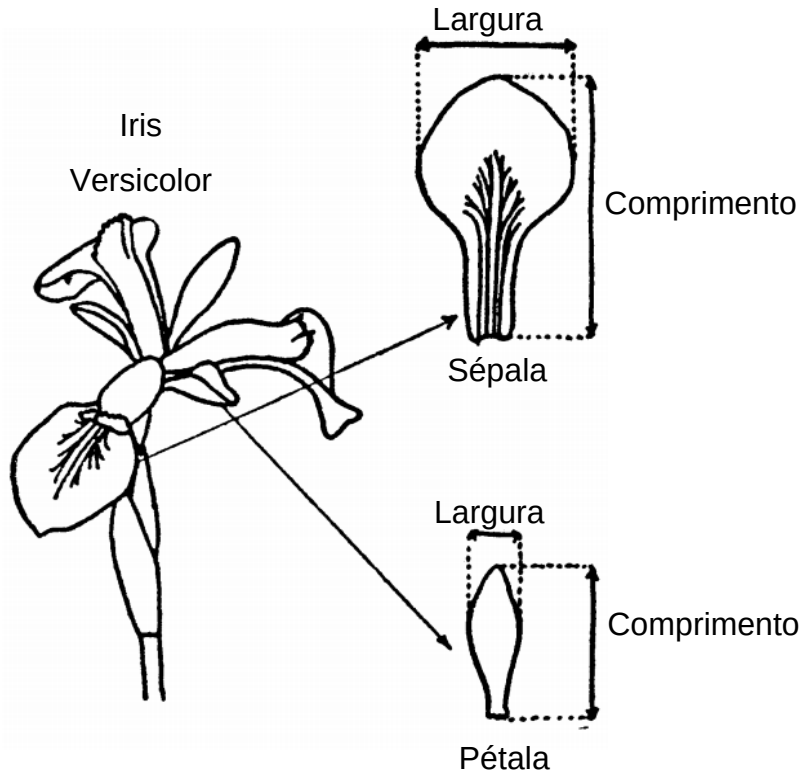
A Árvore de decisão é um exemplo de algoritmo que pode ser estudado no CODAP, assim, ampliando as possibilidades de estudo em base de dados multivariadas. Esse algoritmo será utilizado em uma base de dados com importância histórica, mas antes, será feita uma Análise Exploratória de Dados para melhor compreensão das observações que compõem o conjunto de dados.

## 3.4. Uma breve análise descritiva de dados no CODAP – o conjunto Iris

Nesta seção, uma conhecida base de dados será analisada com o apoio do CODAP. Ela foi obtida pelo botânico inglês Edgar Anderson (1897 - 1969) e usada pelo famoso cientista *sir* R. A. Fisher (1890 – 1962) no artigo "*The use of multiples measurements in taxonomic problems*", publicado em 1936. O estudo de Fisher teve como objetivo desenvolver uma técnica de análise multivariada, que vem a ser uma área da estatística que trabalha com várias variáveis observadas no mesmo indivíduo (ou mesma unidade amostral).

A base de dados é composta por 150 flores de três espécies: *Iris Setosa, Iris Versicolor* e *Iris Virgínica*. Cada uma possuí quatro variáveis numéricas que se referem a medidas das próprias flores em quatro dimensões: comprimento e largura da Sépala e comprimento e largura da pétala (todas em cm). Um exemplo é ilustrado na Figura 9 a seguir no caso de uma *Iris Versicolor*.

**Figura 9** - Sépalas e Pétalas de uma Versicolor




Fonte: ANDERSON, 1936, p. 488 (adaptado)

Uma análise visual do arquivo que contém o conjunto de medidas permite perceber sua ordem de grandeza, embora não identifique o comportamento de cada variável.

Através de uma análise descritiva, tal padrão em cada medida pode ser mais bem explorado a partir de gráficos e do cálculo de medidas estatísticas (chamadas de medidas resumo) que ajudam a conhecer as variáveis, ou seja, as medidas de posição (como Média, Mediana, 1º Quartil, 3º Quartil, Mínimo e Máximo) e as de variabilidade (como Amplitude e Intervalo Interquartil (IIQ)).

Inicialmente, os cálculos das medidas de posição foram obtidos diretamente do CODAP, como feito na seção 3.3. Já as medidas de variabilidade foram determinadas a partir das medidas de posição (ou seja, Amplitude = Máximo – Mínimo, IIQ = Q3 – Q1). Estas informações foram organizadas na Tabela 3, Tabela 4, Tabela 5 e Tabela 6, o que nos permite fazer algumas comparações e/ou considerações, começando pelas sépalas e depois abordando as pétalas.

**Tabela 3** – Comprimento da Sépala (CS) (n=150)

| | Setosa | Versicolor | Virgínica |
|---|---|---|---|
| Mínimo | 4,3 | 4,9 | 4,9 |
| Média | 5,0 | 5,9 | 6,6 |
| Q1 | 4,8 | 5,6 | 6,2 |
| Mediana | 5,0 | 5,9 | 6,5 |
| Q3 | 5,2 | 6,3 | 6,9 |
| IIQ | 0,4 | 0,7 | 0,7 |
| Máximo | 5,8 | 7,0 | 7,9 |
| Amplitude | 1,5 | 2,1 | 3,0 |

Fonte: Bonangelo (2023)

A variável CS é a que possui a maior estrutura (em média) se comparada às outras. Os valores das três espécies apresentam certa semelhança em grandeza, com a *Setosa* apresentando valores um pouco menores e, também, com menor variabilidade (sugerida tanto pela amplitude quanto pelo IIQ). Há uma certa simetria, percebida pela proximidade entre média e mediana. A outra medida da Sépala, largura da Sépala (LS), possui informações organizadas na Tabela 4 e comentadas logo a seguir, sugerindo semelhanças entre CS e LS.

**Tabela 4** – Largura da Sépala (LS) (n=150)

| | Setosa | Versicolor | Virgínica |
|---|---|---|---|
| Mínimo | 2,3 | 2,0 | 2,2 |
| Média | 3,4 | 2,8 | 3,0 |
| Q1 | 3,1 | 2,5 | 2,8 |
| Mediana | 3,4 | 2,8 | 3,0 |
| Q3 | 3,7 | 3,0 | 3,2 |
| IIQ | 0,6 | 0,5 | 0,4 |
| Máximo | 4,4 | 3,4 | 3,8 |
| Amplitude | 2,1 | 1,4 | 1,6 |

Fonte: Bonangelo (2023)

Da mesma forma que CS, a variável LS apresenta certa semelhança na ordem de grandeza nas três espécies e uma certa simetria devido à proximidade de valores de média e mediana. Entretanto, diferente de CS, a *Setosa* se apresenta com valores

um pouco maiores e com maior variabilidade (visto tanto pela amplitude quanto pelo IIQ).

Aparentemente, as sépalas possuem medidas semelhantes entre as espécies quando consideradas cada uma de suas dimensões, mas isso não se mantém nas pétalas, as quais possuem suas medidas estatísticas organizadas nas Tabelas 5 e 6 e comentadas nos parágrafos subsequentes.

**Tabela 5** – Comprimento da Pétala (CP) (n=150)

| | Setosa | Versicolor | Virgínica |
|---|---|---|---|
| Mínimo | 1,0 | 3,0 | 4,5 |
| Média | 1,5 | 4,3 | 5,6 |
| Q1 | 1,4 | 4,0 | 5,1 |
| Mediana | 1,5 | 4,4 | 5,6 |
| Q3 | 1,6 | 4,6 | 5,9 |
| IIQ | 0,2 | 0,6 | 0,8 |
| Máximo | 1,9 | 5,1 | 6,9 |
| Amplitude | 0,9 | 2,1 | 2,4 |

Fonte: Bonangelo (2023)

Para esta variável, comprimento da pétala (CP), os valores das três espécies não são homogêneos quanto à grandeza, com a *Setosa* apresentando valores menores que as demais. A variável apresenta uma certa simetria, percebida através dos valores próximos entre média e mediana. Através das medidas de variabilidade, percebe-se que a *Setosa* apresenta valores mais concentrados.

Após observar 3 dimensões das 4 disponíveis no conjunto de dados, falta verificar quais as características que podem ser observadas quanto a largura das pétalas, com suas medidas estatísticas organizadas na Tabela 6.

**Tabela 6** – Largura da Pétala (LP) (n=150)

| | Setosa | Versicolor | Virgínica |
|---|---|---|---|
| Mínimo | 0,1 | 1,0 | 1,4 |
| Média | 0,2 | 1,3 | 2,0 |
| Q1 | 0,2 | 1,2 | 1,8 |
| Mediana | 0,2 | 1,3 | 2,0 |
| Q3 | 0,3 | 1,5 | 2,3 |
| IIQ | 0,1 | 0,3 | 0,5 |
| Máximo | 0,6 | 1,8 | 2,5 |
| Amplitude | 0,5 | 0,8 | 1,1 |

Fonte: Bonangelo (2023)

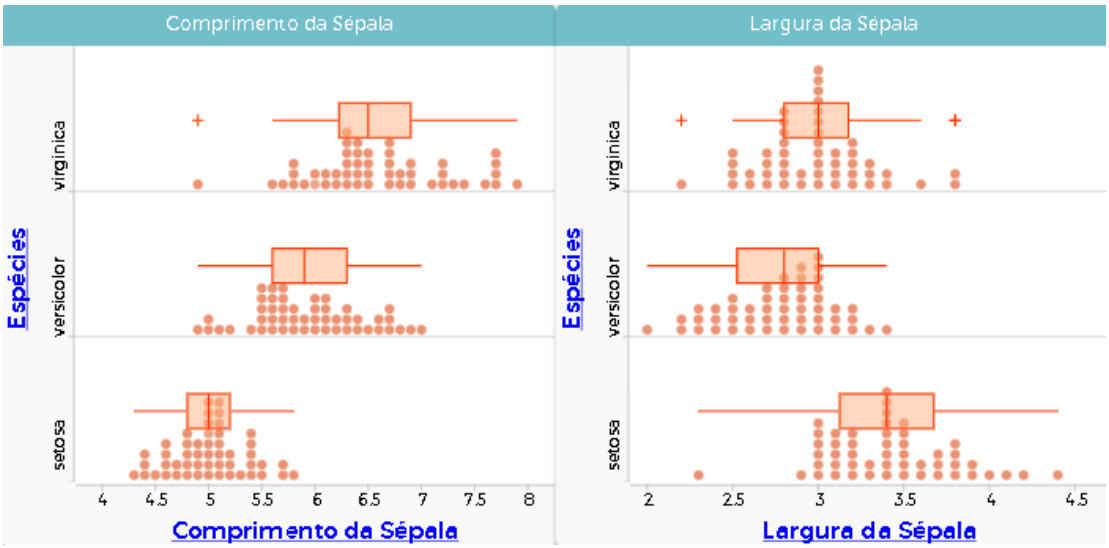
Para esta variável, largura da pétala (LP), os valores das três espécies não são homogêneos quanto à grandeza, apresentando três padrões distintos, com valores inferiores para a *Setosa* e superiores para a *Virgínica*. A exemplo das variáveis anteriores, aqui também se nota uma certa simetria, percebida através dos valores próximos entre média e mediana. Através das medidas de variabilidade, percebe-se que analogamente ao CP, a *Setosa* apresenta valores mais concentrados para a LP.

Em resumo, através das medidas estatísticas organizadas nas Tabela 4, Tabela 5, Tabela 6 e Tabela 7 pode-se observar: (1) CS apresenta a maior estrutura em média; (2) As sépalas apresentam maior homogeneidade em suas dimensões, diferente das pétalas que apresentam maior heterogeneidade; e (3) as quatro variáveis parecem apresentar simetria em cada uma das três espécies, percebidas pela proximidade entre valores de média e mediana.

Mesmo que as tabelas forneçam várias oportunidades de caracterização das variáveis e possam ser analisadas detidamente, esses comentários serão enriquecidos com a abordagem gráfica.

Já foram vistos gráficos na análise do primeiro conjunto de dados e voltaremos a eles: o Gráfico de Pontos (*Dotplot*) e o Gráfico de Caixa (B*oxplot*). A apresentação dos dois superpostos não precisaria ser repetida, (pois bastaria o B*oxplot*), mas vamos repeti-la por entender que é um recurso pedagógico que facilita a compreensão do aluno e a interpretação do comportamento dos dados. Os gráficos (Figura 10 e Figura 11) fazem comparação entre comprimento e largura tanto para sépalas como para pétalas.

**Figura 10** – Comprimento e Largura das Sépalas




Fonte: Bonangelo (2023)

Na análise da Sépala, os gráficos (Figura 9) respaldam os comentários feitos a partir da exploração das tabelas, ilustrando a posição relativa das espécies entre si. A *Setosa* possui os menores valores de CS embora com alguma intersecção observada pelo Gráfico de Caixa, pois mais do que 75% dos comprimentos dessas flores (todas as observações abaixo de Q3) são inferiores a 25% dos comprimentos das demais espécies (essas últimas representadas pelas hastes inferiores). Os 50% centrais (caixa) se destacam à esquerda no gráfico de CS e à direita no de LS, o que não ocorre com as outras duas espécies.

Percebe-se um padrão entre as *Versicolors* e as *Virgínicas* no comprimento e na largura, no qual as *Versicolor* são normalmente menores do que as *Virgínicas.* No entanto, para ambas as variáveis, se verifica que há intersecção entre os valores das espécies, em menor grau em CS (cerca das 75% menores *Virgínicas* e 100% das *Versicolors*) ou maior grau em LS com interseção entre várias observações, principalmente as 50% centrais de ambas as flores. Para a espécie *Virgínica* observam-se *outliers* como definido na seção anterior na p. 9 tanto para comprimento quanto para largura.

Em relação à simetria, observada anteriormente pela proximidade entre média e mediana, os Gráficos de Caixa da Setosa, tanto para CS quanto para LS, corroboram com essa hipótese. No caso das *Versicolors* e das *Virgínicas* a metade

inferior da caixa em CS indica maior concentração de observações em medidas ligeiramente menores dos 50% centrais, enquanto se inverte no caso de LS.

Portanto, os resultados referentes às sépalas podem ser resumidos da seguinte forma:

1) há indícios de simetria na distribuição das plantas através do Gráfico de Caixa (Figura 9).

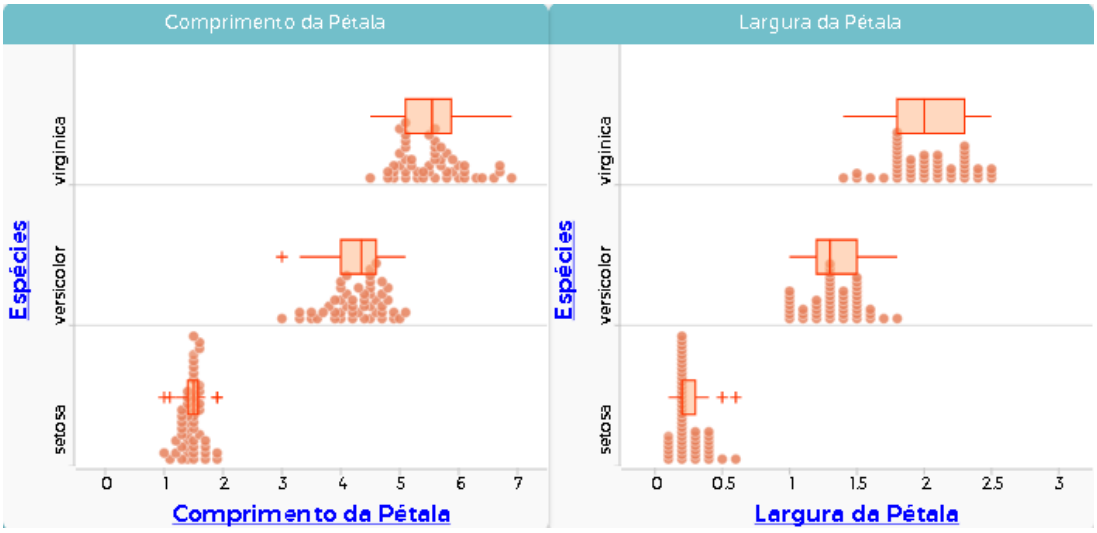
2) Nas medidas de comprimento e largura, a análise gráfica mostrou sempre mais proximidade entre *Versicolors* e *Virgínicas*, embora haja uma intersecção não desprezível com as medidas das *Setosas.*

3) No caso do comprimento, as *Setosas* têm os menores valores (globalmente) e isso se inverte no caso da largura.

4) As *Versicolors* são menores do que as *Virgínicas*, porém com interseção não desprezível. E todas essas observações são bastante distintas das dimensões das pétalas, que serão comentadas a seguir.

Nos Gráficos do comprimento e da largura das pétalas (Figura 10) a espécie *Setosa* se destaca com valores bem inferiores aos das outras duas espécies. Tanto que o máximo dela é inferior ao mínimo das outras, distinguindo-a das demais com relação a comprimento e a largura das pétalas. Característica que já poderia ser observada na Tabela 5 e na Tabela 6, mas agora mais perceptíveis na representação gráfica.

**Figura 11** - Comprimento e Largura das Pétalas




Fonte: Bonangelo (2023)

É imediato verificar graficamente que a *Setosa* possui variabilidade menor dos que as demais para as duas variáveis. Também é possível ver a presença de *outliers*, em CP e em LP. Verifica-se uma certa homogeneidade da espécie em relação às suas medidas de pétala, com poucos casos discrepantes.

As outras duas espécies possuem aspectos distintos das *Setosas*, como as *Virgínicas* que são globalmente maiores do que as outras duas nas dimensões de pétalas e, ao comparar *Versicolor*s e *Virgínicas*, em ambas as variáveis, o valor máximo das *Versicolors* não supera o 1º Quartil (Q1) das *Virgínicas*. Ou seja, há semelhança entre as dimensões de cerca de 25% das menores *Virgínicas* com as maiores *Versicolors*. Há ocorrência de *outliers* na variável “comprimento” para *Setosa* e *Versicolor*. No caso da largura há ocorrência de *outlier* somente para *Setosa*. Há também *outlier* na *Versicolor* no CP, sendo o menor valor considerado no gráfico como o distinto do grupo, mas ainda sim superior a todas as *Setosas*.

Portanto, os resultados referentes às pétalas podem ser resumidos da seguinte forma:

(1) As *Setosas* se destacam por possuírem os menores valores para CP e LP, isoladas a esquerda do eixo nos gráficos explorados.

(2) As *Virgínicas*, por outro lado, possuem valores globalmente maiores do que as demais.

(3) Não há medidas em comum entre as *Setosas* com as outras duas espécies.

(4) Entretanto, cerca de 25% das menores *Virgínicas* possuem semelhança com as maiores *Versicolor*s.

(5) Com relação a pontos discrepantes, há casos na *Setosa* em CP, em LP e na *Versicolor* no CP.

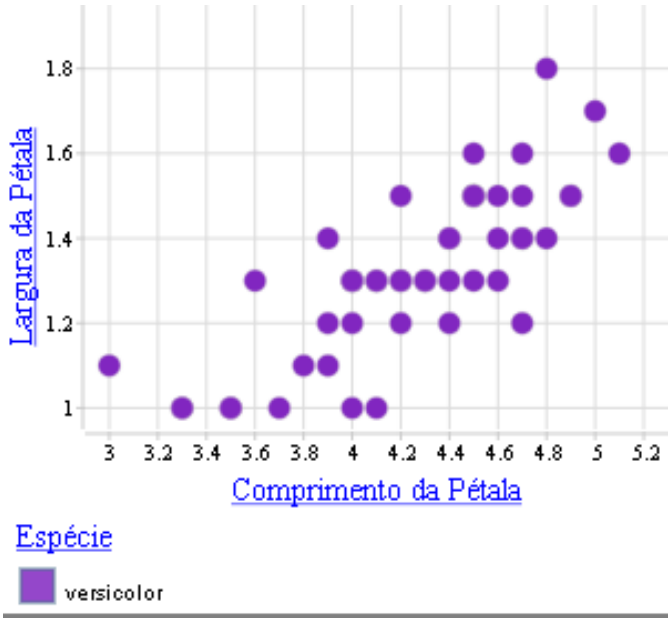
Da mesma forma que foi realizado até então, buscando a descrição das dimensões das sépalas e pétalas através das medidas de comprimento e largura com abordagens univariadas, o Gráfico de Dispersão será construído, proporcionando uma abordagem bivariada para explorar a relação entre duas variáveis.

O Gráfico de Dispersão é um gráfico de coordenadas cartesianas que mostra o comportamento conjunto de duas variáveis quantitativas – por exemplo, comprimento e largura de pétalas ou de sépalas. Vamos propor aqui inicialmente, numa abordagem didática, a construção de um gráfico de dispersão entre comprimento e largura de pétalas, para uma só espécie (*Versicolor*), mostrando a sua construção no CODAP e fazendo uma análise da relação apresentada entre essas

duas variáveis. O professor pode usar tal construção no *software*, como também reproduzi-la com lápis e papel, assim como todos os gráficos apresentados até o momento.

O Gráfico de Dispersão construído a partir das dimensões das pétalas das *Versicolors* pode ser observado a seguir (Figura 12). No gráfico estão representados 36 pontos e não 50 como foi descrito anteriormente. Isso decorre de pares repetidos, e, portanto, sobrepostos. Esse detalhe será considerado na discussão que se seguirá.

**Figura 12** - O Gráfico de Dispersão das Versicolors



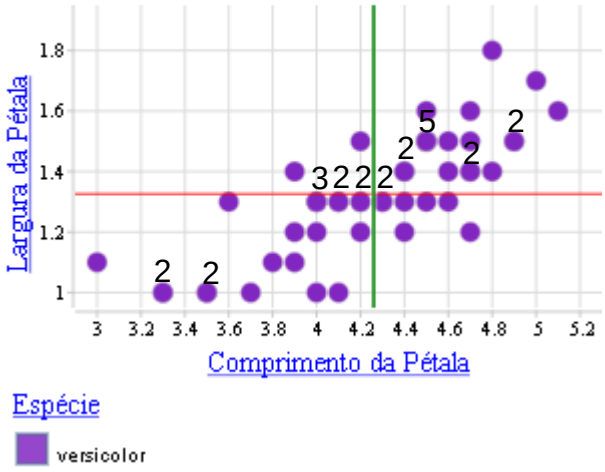

Fonte: Bonangelo (2023)

No entanto, antes vamos pensar em como analisamos o comportamento conjunto dessas duas variáveis em uma sala de aula. Levando em conta a sugestão do Relatório GAISE (BARGAGLIOTTI et al., 2020) é possível fazer três abordagens de acordo com a formação prévia dos alunos: **Abordagem A** – estudantes, no início da formação em estatística, analisariam o gráfico informalmente, procurando por algum comportamento explícito, por exemplo alguma tendência, com o intuito de perceber se existe ou não alguma associação entre comprimento e largura de pétalas, para descrevê-la com palavras. Eventualmente, os alunos poderiam tentar traçar uma reta a fim de verificar a hipótese de que “se o comprimento aumenta, então a largura também aumenta?”.

Na **Abordagem B** – estudantes que já passaram pela Abordagem A, tendo construído o gráfico de dispersão, e tendo discutido uma eventual tendência de comportamento conjunto das duas variáveis (linear, por exemplo), podem tentar descobrir quão forte seria essa relação, através de uma medida quantitativa, que seria a *Razão de Contagens entre Quadrantes* - RCQ (em inglês Quadrant Count Ratio (QCR)). A **Abordagem C** será um incremento para o estudo da correlação com o acréscimo do Coeficiente de Correlação Linear de Pearson, o qual será comentado após a exploração da RCQ.

A RCQ pode ser construída e analisada de modo exploratório, usando a média das variáveis ou a mediana. A Figura 13 foi construída usando a média de cada variável para delimitar os quadrantes, sendo a média do comprimento da pétala de *Versicolor* igual a 4,26 cm e média de largura igual a 1,33 cm e como RCQ não é uma medida utilizada na escola básica, porém com grande valor pedagógico, será detalhada aqui a sua construção. Os valores replicados estão representados sobre o ponto em caso de sobreposição.

**Figura 13** - Gráfico de Dispersão das Versicolors dividido em quadrantes




Fonte: Bonangelo (2023)

Analisando os quadrantes, os pontos no Quadrante 1 (superior direito) mostram flores com valores maiores do que a média tanto para comprimento como para largura de pétalas. Já os do Quadrante 3 (Inferior esquerdo) mostram flores com valores menores do que a média para ambas as variáveis. Para os Quadrantes pares temos uma relação inversa, ou seja, os pontos no Quadrante 2 (superior esquerdo) mostram

plantas com comprimento menor do que a média e largura maior do que a média, já no Quadrante 4 (inferior direito) é o contrário do segundo quadrante (comprimento maior do que a média e largura menor do que a média).

Como descrito no Relatório GAISE “uma medida de correlação é uma quantidade que mede a direção e a força de uma associação entre duas variáveis quantitativas” (BARGAGLIOTTI et al., 2020, p. 61, *tradução nossa*). Essa associação é dita positiva se houver mais pontos nos quadrantes ímpares, sugerindo uma relação diretamente proporcional entre as duas variáveis, enquanto, caso os pontos sejam majoritários nos quadrantes pares, há indícios de uma associação inversamente proporcional entre as duas variáveis.

No caso das *Iris Versicolor* (Figura 13) podemos obter RCQ verificando sua posição em cada quadrante, levando em conta as réplicas de algumas flores com mesmas dimensões para comprimento e largura das pétalas, totalizando 50 *Versicolors* distribuídas nos quatro quadrantes (Tabela 7).

**Tabela 7** – Quantidade de Versicolors por quadrante (n=50)

| Quadrante | Contagem |
|---|---|
| 1º | 20 |
| 2º | 2 |
| 3º | 21 |
| 4º | 7 |
| Total | 50 |

Fonte: Bonangelo (2023)

Observa-se então que os quadrantes ímpares (1 e 3) contribuem com 41 pontos, indicando uma associação positiva e os quadrantes pares com nove mostrando uma associação negativa. Portanto, a expressão que definirá RCQ (BARGAGLIOTTI et al., 2020) será definida em função dessas quantidades:

$$RCQ = \frac{n_{Q1} + n_{Q3} - n_{Q2} - n_{Q4}}{n}$$

sendo $n_{Q1}$, $n_{Q2}$, $n_{Q3}$ e $n_{Q4}$ a quantidade de observações nos quadrantes 1, 2, 3, 4 respectivamente e “n” o número total de observações. Nota-se também que RCQ não tem dimensão.

Usando os valores obtidos na Tabela 7 podemos calcular a medida de associação entre comprimento e largura da pétala para *Versicolor*:

$$RCQ_{Versicolors,\ CPxLP} = \frac{n_{Q1} + n_{Q3} - n_{Q2} - n_{Q4}}{n}$$

$$= \frac{20 + 21 - 2 - 7}{50}$$

$$= \frac{41 - 9}{50}$$

$$= 0{,}64$$

ou seja, esses dados indicam que há uma associação positiva entre as variáveis comprimento e largura de Pétalas (análise exploratória do Gráfico de Dispersão) e o valor obtido da RCQ quantifica essa associação positiva (não perfeita) entre essas duas variáveis.

Se todas as observações estivessem nos quadrantes Q1 e Q3, o RCQ= +1 indicaria uma perfeita associação direta, pois enquanto uma variável aumenta, a outra também aumenta. Contudo, se todas estivessem nos quadrantes Q2 e Q4, o RCQ= -1 apontaria uma perfeita associação inversa, dado que uma variável aumenta e a outra diminui. Assim temos que RCQ varia de -1 a +1. No entanto se os dados estiverem igualmente distribuídos entre os quadrantes, o RCQ será 0 e seria um indicativo de não haver relação linear entre as variáveis.

Os dados acima indicam uma associação positiva entre as variáveis comprimento e largura de Pétalas (análise exploratória do gráfico de dispersão na Figura 13) e o valor obtido do RCQ quantifica essa associação positiva (não perfeita) entre essas duas variáveis. Dessa forma, o estudo exploratório pode ser melhorado num trabalho conjunto com a função linear, proporcionando uma excelente oportunidade de trabalho simultâneo entre questões estatísticas e algébricas (BARGAGLIOTTI et al., 2020).

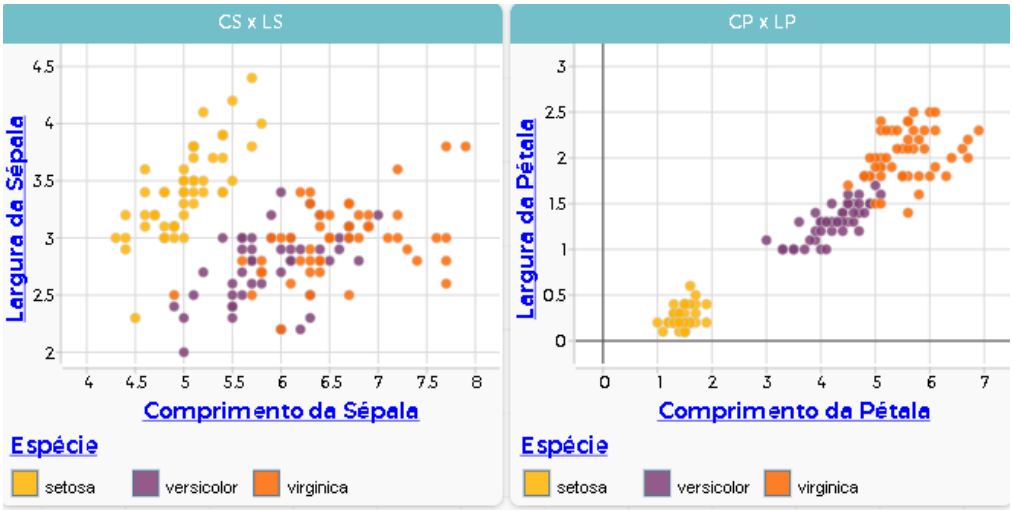
Para verificar tendências de distribuição dos pontos, a **Abordagem C** sugere a introdução do coeficiente de Coeficiente de Correlação Linear de Pearson, também uma medida sem dimensão e variando entre (-1 e +1), considerando a distância dos pontos à média.

O Coeficiente de Correlação Linear de Pearson é mais poderoso do que a RCQ e de abordagem mais formal. Alguns de seus valores em relação as espécies de Iris serão oferecidos na Tabela 8, para comparação e breve abordagem de suas propriedades.

O Coeficiente de Correlação Linear de Pearson pode ser uma ferramenta para entender a comportamento conjunto das variáveis Comprimento e Largura tanto para Sépalas quanto para Pétalas não só das Versicolors, como também das outras

espécies. Através do Gráfico de Dispersão podemos ter uma representação do seu efeito, por isso, o coeficiente será abordado após a Figura 13 a seguir.

**Figura 14** - Gráfico de Dispersão CS x LS e CP x LP




Fonte: Bonangelo (2023)

A partir de tal representação, além de ser possível reafirmar os resultados encontrados nas análises gráficas anteriores (com o Gráfico de Caixa e de Pontos), algumas observações podem ser propostas: (1) As dimensões das sépalas parecem ter uma relação linear mais acentuada no caso da *Setosa*, mas as outras duas parecem mais dispersas no gráfico. (2) As pétalas possuem relações entre comprimento e largura que sugerem um crescimento conjunto mais uniforme, característica observada devido ao menor espalhamento de pontos pelo gráfico, sugerindo uma relação linear.

Se o RCQ fosse calculado para ambos os gráficos, considerando todas as 150 flores e a média para delimitar os quadrantes, obteríamos $RCQ_{CSxLS}$= -0,16 (gráfico CS x LS na Figura 13) e $RCQ_{CPxLP}$= 0,93 (gráfico CP x LP na Figura 13), isto é, enquanto o $RCQ_{CSxLS}$ é próximo de zero, indicando maior homogeneidade do espalhamento de observações nos quatro quadrantes, mas com sinal negativo, por possuir mais observações nos quadrantes pares; o $RCQ_{CPxLP}$ indica uma associação positiva próxima de um, resultado da maior quantidade de flores nos quadrantes 1 e 3.

O Coeficiente de Correlação Linear de Pearson (amostral) definido a seguir, denominado por **r**, pode informar melhor a existência, ou não, de relação linear entre

as variáveis. Por exemplo, x poderia ser o comprimento da pétala e y a largura da pétala,

$$r = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{dp(x) \cdot dp(y)},$$

onde $\bar{x}$ *e* $\bar{y}$ são as médias das variáveis x e y, $dp(x)$ e $dp(y)$ são os desvios padrões para x e y; por fim, $n$ é a quantidade de observações no conjunto de dados.

O valor de **r** é limitado entre -1 e 1, com algumas características que podem ser ilustradas através dos valores organizados na Tabela 8 e referentes à análise exploratória das três espécies em relação às variáveis comprimento e largura das sépalas e/ou pétalas. Observa-se que: **(1) (*r* positivo, ou negativo)** neste caso todos são positivos, indicando relação diretamente proporcional entre as variáveis; caso houvesse valores negativos a relação seria inversamente proporcional. **(2) (**$r$ próximo dos extremos**)** quanto mais próximo de +1 (ou -1) a relação é dita mais forte, além dos pontos se aproximarem de uma reta (por exemplo as *Versicolor*s na Figura 13). **(3) (**$r$ e o espalhamento dos pontos**)** quanto mais próximo de 0, mais os pontos parecem espalhados (por exemplo as *Virgínicas* na Figura 13). **(4) (**$r$ próximo de 0**)** quanto mais próximo de 0 é o valor do Coeficiente, mais fraca é a relação linear e caso seja igual a 0 é dito que não há relação linear.

**Tabela 8** – Coeficiente de Correlação Linear de Pearson das Iris (n=150)

| | CS x LS | CP x LP |
|---|---|---|
| Setosa | 0,75 | 0,31 |
| Versicolor | 0,53 | 0,79 |
| Virgínica | 0,46 | 0,32 |

Fonte: Bonangelo (2023)

Portanto, os resultados referentes aos Gráficos de Dispersão podem ser resumidos em: (1) As variáveis comprimento e largura em uma mesma estrutura (sépala ou pétala) são diretamente proporcionais. (2) A relação linear entre as dimensões das sépalas das *Setosas* $r = 0{,}75$ e das pétalas das Verisolors $r = 0{,}79$ são ditas mais fortes por estarem mais próximos do extremo 1. (3) Para os valores mais próximos de 1, há indicativo de maior força de uma relação linear e os pontos se localizam mais próximos a uma reta, enquanto os mais próximos de 0 sugerem uma relação linear mais fraca e aparentemente mais espalhados pelo Gráfico de Dispersão da Figura 13.

O Coeficiente de Correlação Linear de Pearson e o Gráfico de Dispersão podem ser mais explorados em Holmes (2001), Magalhães e Lima (2015) e Morettin e Singer (2020), bem como sua forma analítica e interpretações em outros contextos.

Por fim, para reunir e resumir todas as informações formadas até aqui com o estudo das tabelas, Gráficos de Pontos, de Caixa e de Dispersão:

1) As três espécies mantêm as dimensões (comprimento e largura) de pétalas e o comprimento das sépalas em uma ordem crescente começando pelas *Setosas* e terminando com as *Virgínicas*, como visto nas Tabela 3, Tabela 4, Tabela 5 e Tabela 6. A exceção é referente à largura das sépalas, das *Setosas* que apresentam maiores dimensões e variação.
2) Todas as quatro variáveis têm comportamento homogêneo nas três espécies, com poucos valores discrepantes.
3) As pétalas parecem caracterizar melhor cada espécie, havendo algumas medidas semelhantes entre *Versicolor*s *e Virgínicas*, ambas características observáveis através dos Gráficos de Pontos e os de Caixa nas Figuras 23 e 24.
4) As dimensões de uma mesma estrutura (sépala ou pétala) neste caso são diretamente proporcionais, como verificado nos Gráficos de Dispersão (Figura 14).

Algumas das particularidades das três espécies foram observadas em relação à dimensão de suas estruturas. A continuação do estudo de mais detalhes sobre correlação poderia ser explorada em uma análise mais avançada, porém não é o intuito do texto esgotar tal tema; o objetivo era a descrição, com conhecimentos possíveis de serem explorados na escola básica, do estudo de tabelas e de gráficos produzidos pelo próprio CODAP, passíveis de serem feitos sem ele também.

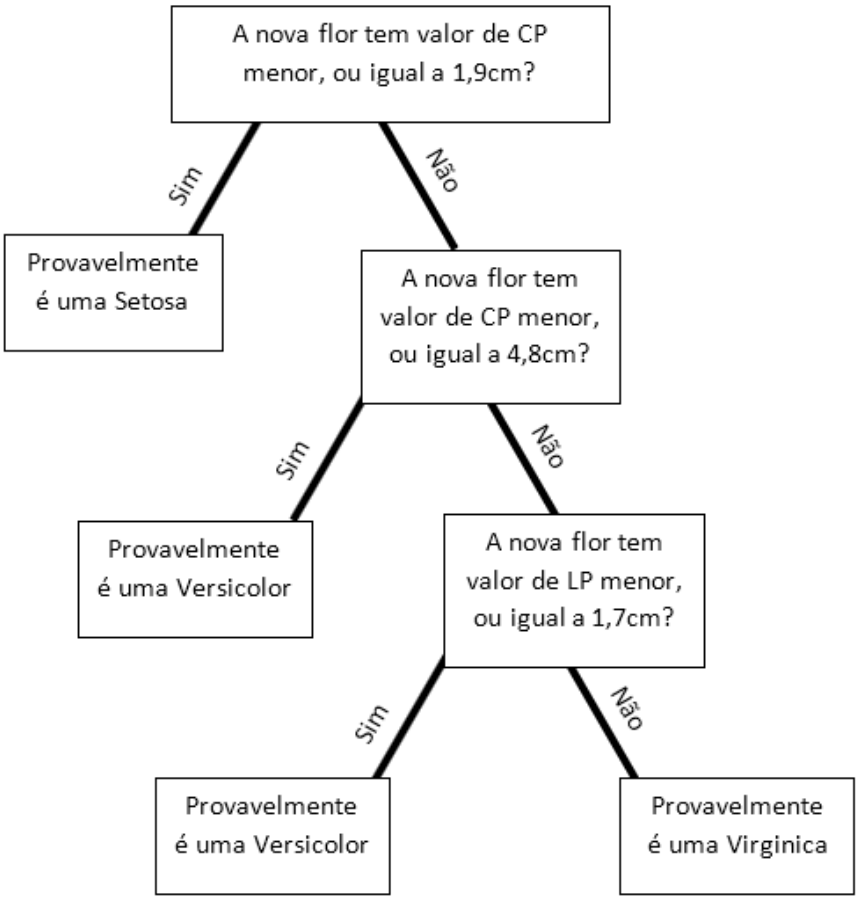
Por outro lado, as tabelas, os gráficos e as medidas estatísticas exploradas nesta análise estão presentes nas ementas da Educação Básica, e podem também ser exploradas para abordar alguns dos algoritmos citados no começo do texto, ainda de forma elementar, como é o caso da Árvore de Decisão que será abordada no próximo capítulo.

## 4. A Árvore de Decisão no CODAP

A Árvore de Decisão pertence a um grupo de classificadores que sistematizam o processo para determinar a qual classe um dado pertence. O estudo de classificadores é justificado a partir de duas situações: (1) na investigação da estrutura preditiva de algum problema, ou (2) na criação de classificadores mais assertivos (BREIMAN et al., 1984). Por exemplo, caso seja encontrada uma nova Iris, sua espécie (*Setosa*, *Versicolor* ou *Virgínica*) pode ser determinada a partir das suas medidas: comprimento de sépala (CS), largura de sépala (LS), comprimento de pétala (CP) e largura de pétala (LP).

A Figura 15 é uma representação de uma Árvore de Decisão feita para auxiliar na classificação da variável “espécie” de uma Iris desconhecida. Com as dimensões de sépala e de pétala da Iris desconhecida, o primeiro questionamento apresentado em um **nó** é referente ao valor de CP: “A nova flor tem valor de CP menor, ou igual a 1,9cm?”. Em caso afirmativo, segue-se para o nível inferior pelo “Sim” e a flor em questão “Provavelmente é uma *Setosa”.* Caso contrário, segue-se pelo “Não” e buscam-se outros critérios para definir qual é a espécie da Iris desconhecida.

**Figura 15** - Árvore de Decisão para classificar a espécie de uma Iris

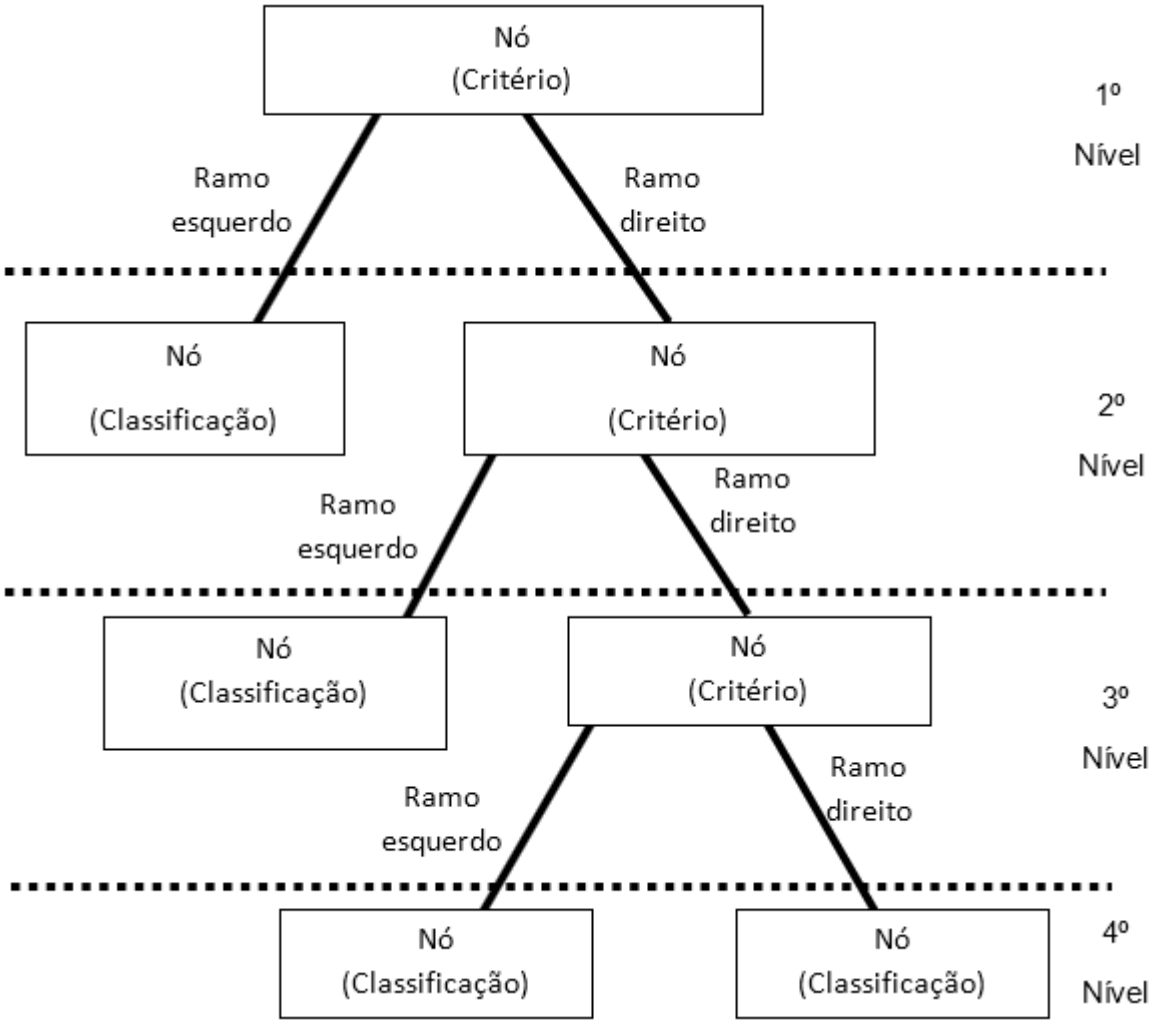



Fonte: Bonangelo (2023)

Nota-se que a representação fornecida (Figura 15) auxilia na investigação da estrutura preditiva, ou seja, indica quais variáveis podem ser utilizadas para caracterizar as espécies de Iris desconhecidas. Tais resultados são semelhantes ao que foi estudado no capítulo anterior e que será comentado ainda neste capítulo.

A representação do modelo possui três componentes importantes: (1) Os **nós** possuem um critério ou uma classificação. (2) Os nós podem ser divididos em **níveis**, sendo a Figura 15 formada por quatro níveis. Por fim, (3) os **ramos** são a ligação entre nós de níveis diferentes. Uma representação destes elementos pode ser conferida na Figura 16.

**Figura 16** - Árvore de Decisão para classificar uma Iris




Fonte: Bonangelo (2023)

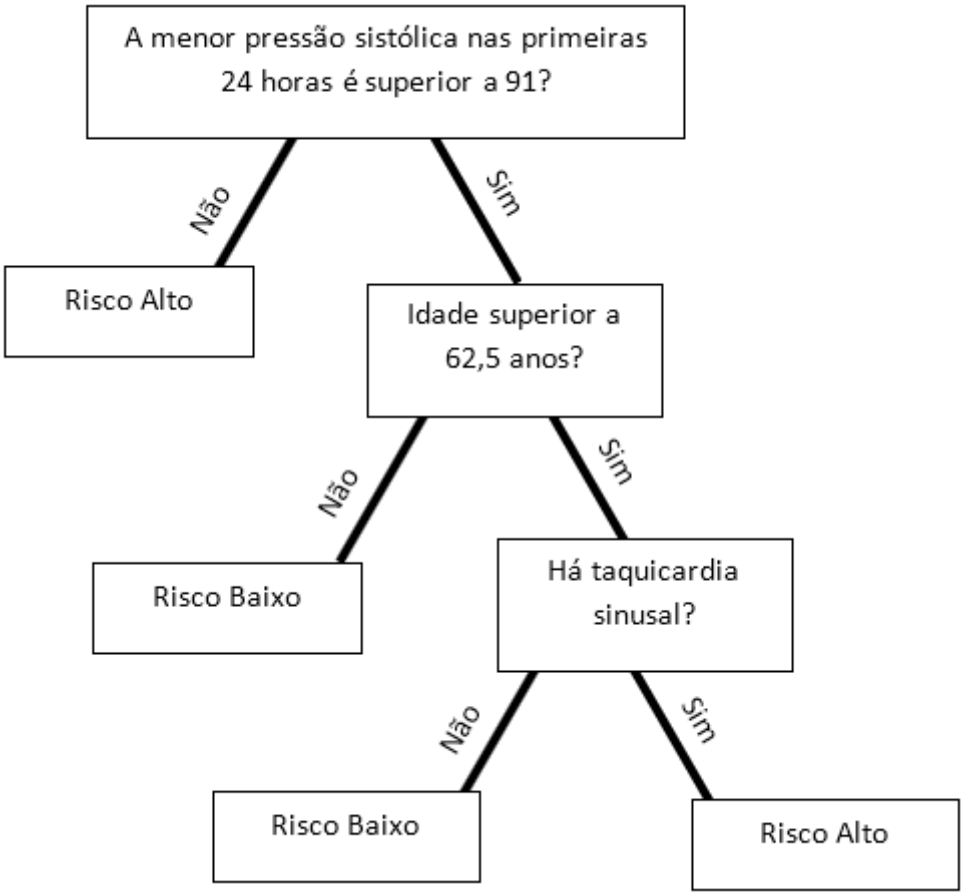
Portanto, este capítulo será destinado ao estudo e à construção do modelo da Árvore de Decisão a partir dos dados analisados no capítulo anterior. O objetivo será: (1) ilustrar a estrutura preditiva; (2) utilizar os conteúdos da Estatística passíveis de serem estudados na Educação Básica para melhor compreensão da Árvore de Decisão; (3) utilizar medidas simples para verificar a qualidade do classificador.

## 4.1. Uma breve apresentação da Árvore de Decisão

A Árvore de Decisão, segundo Breiman et al. (1984), é utilizada pelas ciências sociais desde antes da década de 1980 e, com o avanço dos computadores, foram ferramentas importantes no estudo de diversas variáveis.

A utilização da Árvore de Decisão é ilustrada por Breiman et al. (1984) em três situações, sendo uma delas a verificação se um paciente tem alto ou baixo risco de ataque cardíaco, resultando em atendimento hospitalar imediato ou não. Para montar o classificador (Figura 17), os pesquisadores dispunham de 19 variáveis do paciente, medidas nas primeiras 24 horas.

**Figura 17** - Árvore de Decisão para classificar o risco de ataque cardíaco




Fonte: Breiman et al. (1984) (Adaptado)

Portanto, os pesquisadores identificaram quais variáveis seriam mais importantes de serem verificadas, para classificar o risco em alto ou baixo de um paciente ter problema cardíaco. Também é importante notar que não foi necessário o uso de todas as dezenove variáveis à disposição, somente a identificação das mais

explicativas para o problema. Para verificar mais detalhes desse exemplo e outras duas situações, consultar Breiman et al. (1984).

## 4.2. A Árvore de Decisão na Educação Básica

Os pesquisadores Biehler et al. (2020) utilizaram a Árvore de Decisão para verificar se estudantes dos anos finais da Educação Básica conseguiam prever a frequência que as pessoas jogavam jogos online a partir de quinze variáveis disponíveis. A produção do classificador foi feita pelos alunos de maneira manual e permitiu que justificassem suas montagens. Posteriormente, outras medidas foram utilizadas para verificar a qualidade da Árvore de Decisão. Com essa experiência, os autores fizeram os alunos determinarem variáveis que fossem mais importantes para responderem ao que foi solicitado, ou seja, mais uma abordagem para bases de dados multivariados.

Por fim, a partir da análise das Iris feita no capítulo anterior, quais das medidas seriam interessantes para verificar uma nova espécie? Como determinar a qualidade do classificador? E como esse processo pode ser explicado a partir de tópicos que possam ser abordados nas aulas de estatística na Educação Básica? Essas perguntas serão exploradas nas seções a seguir.

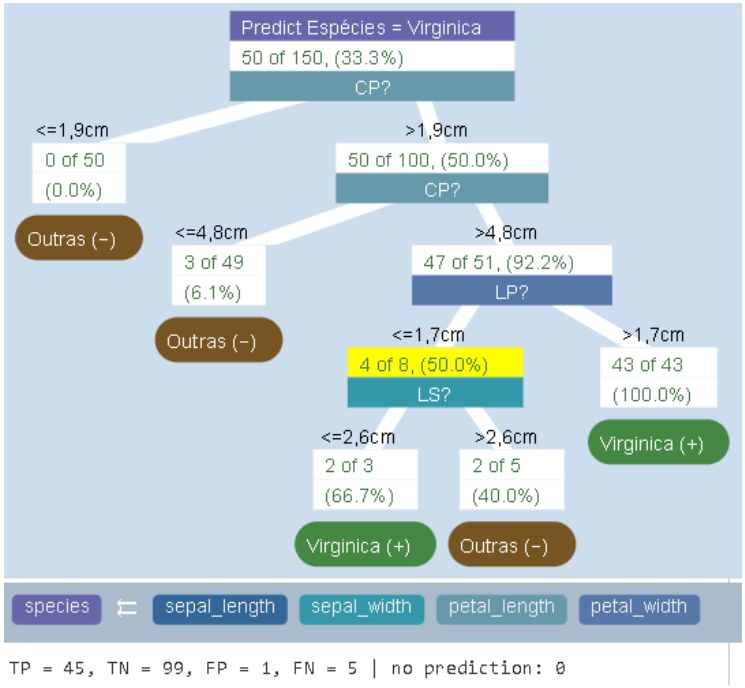
## 4.3. A Árvore de Decisão e as Iris

No CODAP é possível criar uma Árvore de Decisão a partir do *plugin* Arbor[11]. As vantagens são (1) a possibilidade de fazer diferentes classificadores e observar suas métricas de eficiência; (2) aproveitar a interatividade do classificador com os dados brutos. Contudo, as desvantagens são (1) não há como gerar um classificador de modo automático, todos devem ser feitos manualmente; (2) só é possível criar classificadores binários. Ou seja, ou algo a ser classificado é ou não é o que se deseja, como é o caso apresentado na Figura 17, a qual verifica se o risco é alto ou baixo. Enquanto na Figura 15 há a possibilidade de uma Iris ser classificada como *Setosa*, *Versicolor* ou *Virgínica*.

[11] https://codap.xyz/plugins/arbor/

Para além de verificar padrões na base de dados Iris, ao construir uma Árvore de Decisão, será possível verificar novas observações e estimar sua espécie a partir das 150 analisadas anteriormente. Por exemplo, na Figura 18 a seguir, o classificador verificará se uma Iris é *Virgínica* ou não. Para isso, serão utilizadas as dimensões das flores: comprimento de Sépala (CS), largura de Sépala (LS), comprimento de pétala (CP) e largura de pétala (LP).

**Figura 18** - Árvore de Decisão para classificar Virgínicas




Fonte: Bonangelo (2023)

Em cada nível há a proporção de *Virgínicas* naquele nó e um questionamento sobre as medidas das flores, para verificar se a flor é ou não uma *Virgínica*.

- No primeiro nível é verificador o valor de CP
    - Se CP é menor, ou igual a 1,9cm (<= 1,9cm), então a flor é de outra espécie. Não há nenhuma *Virgínica* nessa condição (0 de 50).
    - Se CP é maior do que 1,9cm (>1,9cm) o processo continua. Há 50 *Virgínicas* de 100 flores nessa condição (50 de 100).
- No segundo nível, ainda é analisada a medida de CP.

    - Se CP é menor, ou igual, a 4,8cm, (<=4,8cm), então a flor é de outra espécie. Há três *Virgínicas* de 49 flores que respeitam essa regra (3 de 49), ou seja, seriam classificadas erroneamente.
    - Se CP é maior do que 4,8cm (>4,8), o processo continua. Há 47 *Virgínicas* de 51 flores que respeitam essa regra (47 de 51).
- No terceiro nível indaga-se a variável LP.
    - Se LP for maior do que 1,7cm (>1,7cm), então a flor é uma *Virgínica*. Há 43 *Virgínicas* de 43 flores que respeitam essa regra (43 de 43). Ou seja, seriam classificadas de forma correta.
    - Se LP for menor, ou igual, a 1,7cm (1,7cm), então o processo continua. Há quatro *Virgínicas* de oito flores que respeitam essa regra (4 de 8).
- No quarto nível a variável LS é observado.
    - Se o valor de LS for menor, ou igual, a 2,6cm (2,6cm), então a flor é classificada como *Virgínica*. Há duas *Virgínicas* de três flores que respeitam essa regra (2 de 3). Ou seja, há uma flor que seria erroneamente classificada como *Virgínica*.
    - Se o valor de LS for maior do que 2,6cm então seria uma outra espécie. Há duas *Virgínicas* de cinco flores que respeitam essa regra (2 de 5) e seriam classificadas erroneamente.

Sobre a classificação, já é possível observar que algumas *Virgínicas* são classificadas *Virgínicas* e outras não. A partir disso são compreensíveis os valores sob a Figura 4 representando a Árvore de Decisão gerada com o auxílio do CODAP. Os valores são TP = 45, TN = 99, FP = 1, FN = 5 e *no prediction* = 0.

Ao todo eram 50 *Virgínicas* e 45 delas foram classificadas como *Virgínicas*, elas são as **Verdadeiras Positivas**, tradução do inglês para *True Positives* (TP). Das 150 Iris da base de dados, 100 não eram *Virgínicas* e 99 foram classificadas como de outra espécie, são as **Verdadeiras Negativas**, tradução do inglês para *True Negative* (TN).

Se 100 Iris não eram *Virgínicas* e somente 99 foram classificadas como outra espécie, então há uma que foi dita *Virgínica* mesmo não sendo. Portanto, a observação que foi classificada como *Virgínica* e é outra espécie foi um **Falso Positivo**, tradução do inglês para *False Positive* (FP). De forma análoga, se há 50 *Virgínicas* na base de dados e somente 45 são Verdadeiras Positivas, então as 5

restantes são **Falsos Negativos**, tradução do inglês para *False Negative* (FN). Por fim, como todas as 150 flores podem ser classificadas em *Virgínicas* ou de outra espécie, então a quantidade de Iris **não classificadas** é 0, tradução do inglês para *no prediction.*

Para exemplificar o classificador com dados reais, a Tabela 9 a seguir contém três observações de Iris, retiradas da base de dados analisada na seção anterior. Tais dados podem ser utilizados para ilustrar o funcionamento do classificador proposto na Figura 18.

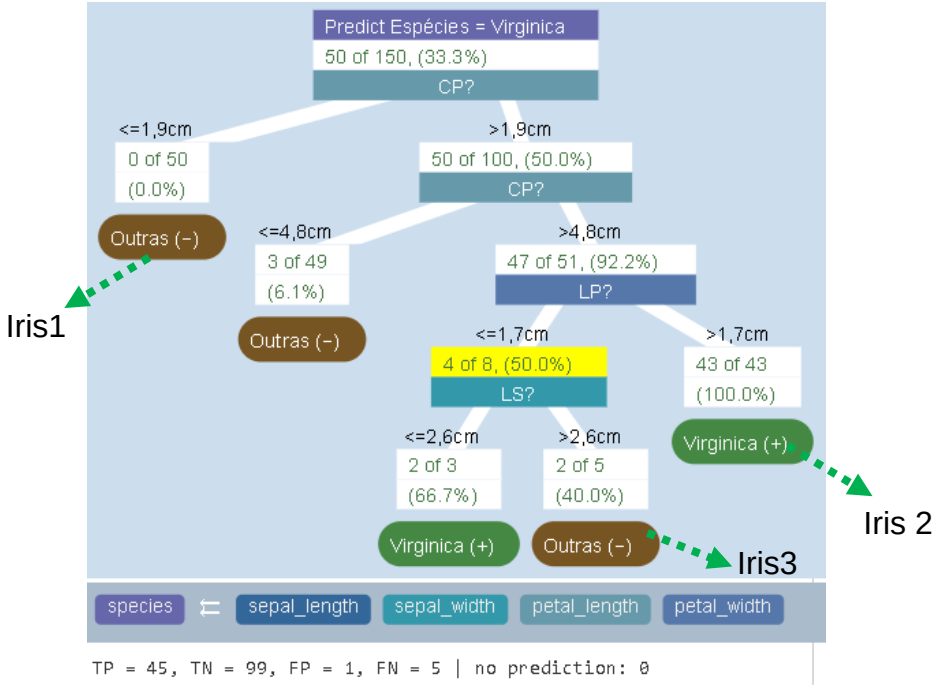
**Tabela 9** – Três flores Iris com Espécie desconhecida para testar a Árvore de Decisão

| Flor | Espécie | Comprimento da Sépala (CS) (em cm) | Largura da Sépala (LS) (em cm) | Comprimento da Pétala (CP) (em cm) | Largura da Pétala (LP) (em cm) |
|---|---|---|---|---|---|
| Iris 1 | ? | 5 | 3,4 | 1,6 | 0,4 |
| Iris 2 | ? | 5,8 | 2,8 | 5,1 | 2,4 |
| Iris 3 | ? | 6 | 2,7 | 5,1 | 1,6 |

Fonte: Bonangelo (2023)

Portanto, seguindo a Árvore de Decisão disposta na Figura 18, obtém-se a Figura 19 a seguir, verificando quais das três Iris são *Virgínicas* ou não.

**Figura 19** – Teste com a Árvore de Decisão para classificar Virgínicas




Fonte: Bonangelo (2023)

Portanto, das três Iris dispostas na Tabela 9, somente uma foi corretamente classificada como *Virgínica* (Tabela 10). A Iris 1 com certeza não é *Virgínica*. A Iris 3 pode ser uma *Virgínica* que foi classificada como de outra espécie (Falso Negativo) ou ser realmente de outra espécie. Portanto, outros classificadores são necessários para analisar se as outras duas Iris são *Setosa*s ou *Versicolor*s.

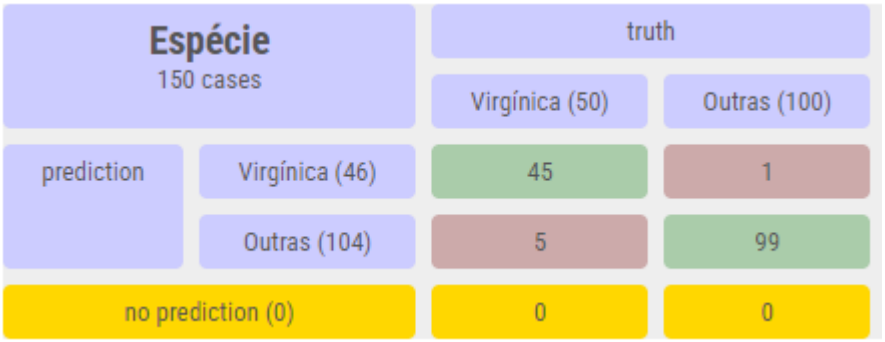
**Tabela 10** – Duas flores Iris com Espécie desconhecida para testar a Árvore de Decisão

| Flor | Espécie | Comprimento da Sépala (CS) (em cm) | Largura da Sépala (LS) (em cm) | Comprimento da Pétala (CP) (em cm) | Largura da Pétala (LP) (em cm) |
|---|---|---|---|---|---|
| Iris 1 | ? | 5 | 3,4 | 1,6 | 0,4 |
| Iris 2 | Virgínica | 5,8 | 2,8 | 5,1 | 2,4 |
| Iris 3 | ? | 6 | 2,7 | 5,1 | 1,6 |

Fonte: Bonangelo (2023)

Antes de utilizar outros classificadores para as outras duas espécies de Iris, é necessário saber o quão bom é esse classificador de *Virgínicas*. Uma primeira proposta para medição da qualidade da Árvore de Decisão foi apresentada através das medidas TP, TN, FP e FN, os quais podem ser organizados na Matriz de Confusão gerada pelo próprio CODAP (Figura 20).

**Figura 20** – Matriz de Confusão da Árvore de Decisão (Virgínica)




Fonte: Bonangelo (2023)

A Matriz de Confusão é uma tabela de dupla entrada formada pelos valores reais nas colunas e os classificados nas linhas. A leitura ocorre da seguinte forma: A quantidade real de *Virgínicas* era 50 flores, 45 foram classificadas como *Virgínicas*

(TP) e cinco como Outras (FN). A quantidade real de outras era 100 observações (50 *Versicolor*s e 50 *Virgínicas*), dentre elas, uma foi classificada como *Virgínica* (FP) e 99 foram classificadas corretamente como Outras (TN).

A diagonal principal da Matriz de Confusão reflete os valores corretamente classificados, enquanto a diagonal secundária reflete os valores incorretamente classificados. Com isso, surge uma medida para verificar a qualidade da Árvore de Decisão, a Acurácia.

De acordo com Bruce e Bruce (2019) a **acurácia** é uma medida do erro total, com valores possíveis no intervalo fechado de extremos 0 e 1. Em termos matemáticos é a razão entre os valores corretamente classificados (total dos elementos da diagonal principal) e a quantidade de dados:

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

No caso do classificador de *Virgínicas*, a acurácia será de

$$acurácia = \frac{45 + 99}{45 + 99 + 1 + 5} = 0{,}96.$$

Portanto, é esperado que em 96% dos casos a Árvore de Decisão para verificar se é ou não *Virgínica*s, faça a classificação correta.

Outra medida que pode ser citada é a precisão. A **precisão** mede a proporção de positivos corretamente classificados, com valores possíveis no intervalo fechado de extremos 0 e 1. Essa medida é a razão entre os Verdadeiros Positivos e o total dos Verdadeiros Positivos com os Falsos Positivos (total da linha predita pelo classificador), ou

$$Precisão = \frac{TP}{TP + FP}$$

No classificador de *Virgínicas* a precisão é de

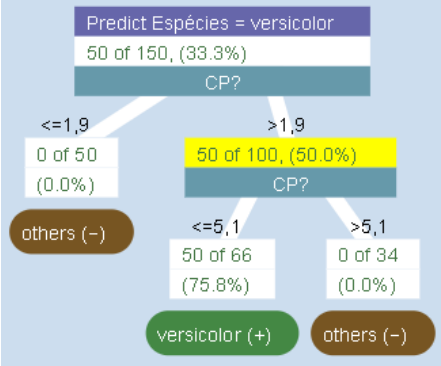
$$Precisão = \frac{45}{45 + 1} = 0{,}98$$

A precisão de 98% significa que entre os 46 classificados como *Virgínica*, somente 45 realmente são. Há outras medidas que podem ser exploradas, mas fogem ao escopo deste texto. No material de Bruce e Bruce (2019) é possível consultar outras medidas.

Com isso, é possível analisar um modelo de Árvore de Decisão para classificar *Versicolor*s e outra para *Setosas*, assim como as Matrizes de Confusão, as medidas de Acurácia e Precisão.

Um possível classificador para as *Versicolor*s é o que está representado na Figura 21.

**Figura 21** - Árvore de Classificação das Versicolors




Fonte: Bonangelo (2023)

Ao utilizá-lo para classificar as duas Iris restantes da Tabela 2, tem-se que a Iris 3 é classificada como *Versicolor*, enquanto a Iris 1 se mantém como “Outras” (Tabela 11).

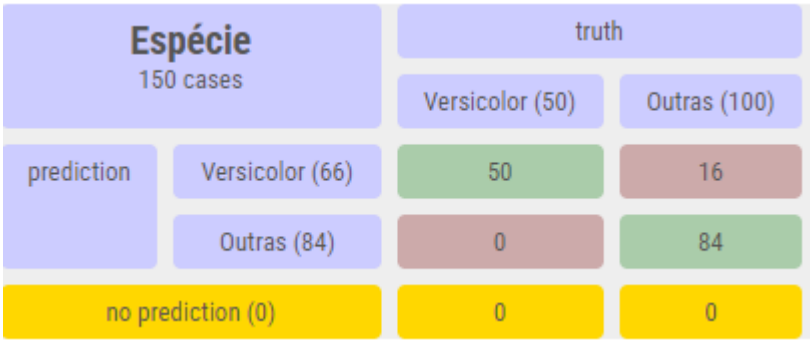
**Tabela 11** – Uma flor Iris com Espécie desconhecida para testar a Árvore de Decisão

| Flor | Espécie | Comprimento da Sépala (CS) (em cm) | Largura da Sépala (LS) (em cm) | Comprimento da Pétala (CP) (em cm) | Largura da Pétala (LP) (em cm) |
|---|---|---|---|---|---|
| Iris 1 | ? | 5 | 3,4 | 1,6 | 0,4 |
| Iris 2 | Virgínica | 5,8 | 2,8 | 5,1 | 2,4 |
| Iris 3 | Versicolor | 6 | 2,7 | 5,1 | 1,6 |

Fonte: Bonangelo (2023)

Em relação a Matriz de Confusão (Figura 22), obtém-se as medidas TP = 50, TN = 84, FP = 16, FN = 0 e *no prediction* = 0. Ou seja, das 50 *Versicolors* disponíveis na base de dados, todas foram corretamente classificadas e das 100 de outras espécies (*Setosa*s e *Virgínicas*) 84 foram classificadas corretamente, resultando em uma acurácia de 89% (134 de 150). Em compensação, a precisão é de 76% (50 de 66), isto é, a cada 100 classificados como *Versicolors*, somente 76 estão corretos.

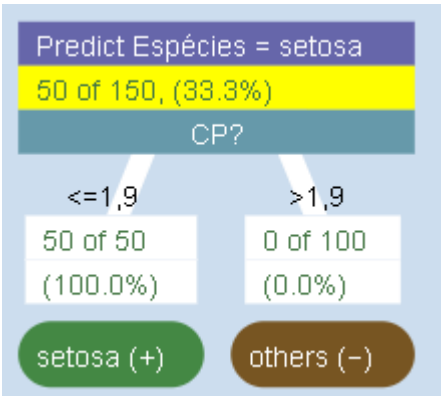
**Figura 22** – Matriz de Confusão da Árvore de Decisão (Virgínica)




Fonte: Bonangelo (2023)

As *Setosa*s são facilmente classificadas, pois basta utilizar a medida de CP para verificar quais são *Setosas* e quais não (Figura 23).

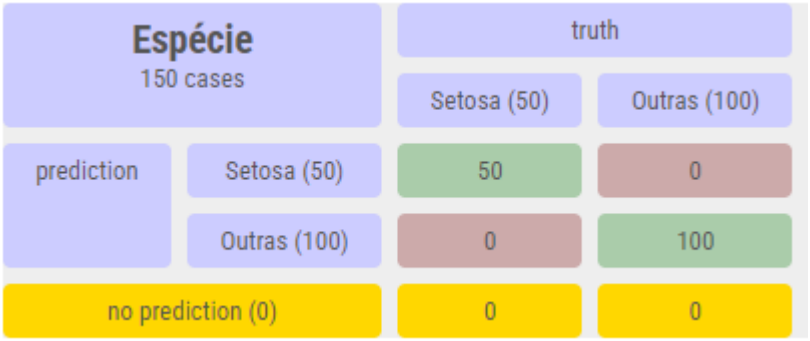
**Figura 23** - Sugestão de Árvore de Classificação das Setosas




Fonte: Bonangelo (2023)

Devido a exatidão desse classificador, a Matriz de Confusão (Figura 24) é bastante simples, com os valores: TP = 50, TN = 100, FP = 0, FN = 0 e *no prediction* = 0. As medidas de Acurácia e Precisão são ambas de 100% para esse classificador.

**Figura 24** – Matriz de Confusão da Árvore de Decisão (Virgínica)




Fonte: Bonangelo (2023)

Portanto, a Iris 1, ainda sem espécie nas tabelas anteriores, pode ser classificada como *Setosa* segundo o classificador de *Setosas* (Tabela 12).

**Tabela 12** – Todas as flores Iris com Espécie classificadas pela Árvore de Decisão

| Flor | Espécie | Comprimento da Sépala (CS) (em cm) | Largura da Sépala (LS) (em cm) | Comprimento da Pétala (CP) (em cm) | Largura da Pétala (LP) (em cm) |
|---|---|---|---|---|---|
| Iris 1 | Setosa | 5 | 3,4 | 1,6 | 0,4 |
| Iris 2 | Virgínica | 5,8 | 2,8 | 5,1 | 2,4 |
| Iris 3 | Versicolor | 6 | 2,7 | 5,1 | 1,6 |

Fonte: Bonangelo (2023)

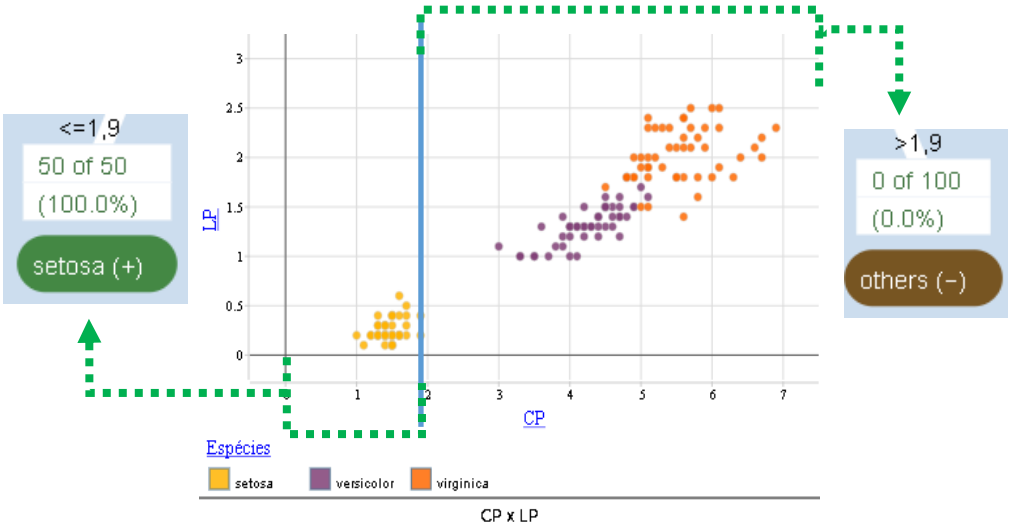
Portanto, falta averiguar como esses classificadores podem ser explicados, pois sua construção é manual no CODAP. Como a Árvore de Decisão é um algoritmo que traz uma nova perspectiva para base de dados multivariados, então deve haver relação com a Análise Exploratória de Dados feita anteriormente.

A variável “espécie” é qualitativa e foi determinada através das variáveis quantitativas relativas às dimensões (comprimento e largura) das sépalas e pétalas. Para estudar CS, LS, CP e LP foram utilizados Gráficos de Pontos e de Caixas como gráficos univariados; Gráficos de Dispersão para uma perspectiva bivariada.

Uma das conclusões da Análise Exploratória foi a caracterização das Iris através da medição das pétalas. Com isso, basta observar como todos os classificadores foram iniciados, a partir da medida do comprimento da pétala.

Uma possibilidade de visualização é a partir do Gráfico de Dispersão CP x LP (Figura 25). Ao se dividir na Árvore de Decisão o conjunto de Iris em dois grupos, um com CP<=1,9cm e outro com CP>1,9cm, duas regiões são definidas no Gráfico de Dispersão.

**Figura 25** - Gráfico de Dispersão ilustrando a Árvore de Decisão (Setosa)

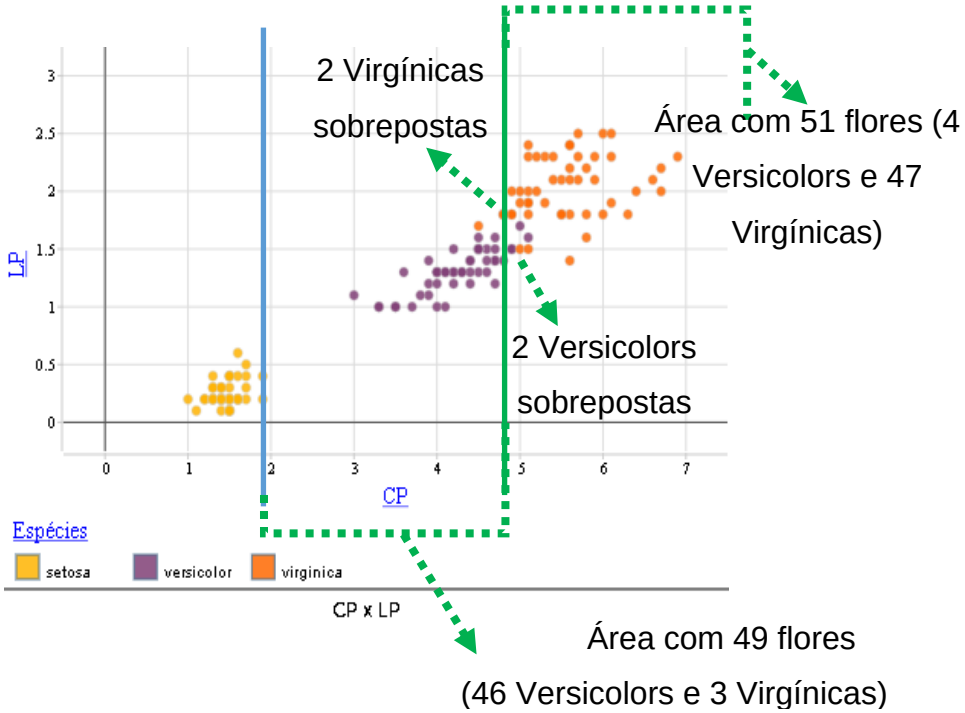



Fonte: Bonangelo (2023)

Claramente à esquerda da divisão em 1,9cm se encontram as 50 *Setosas* das 150 Iris. As 100 flores restantes possuem medidas de CP em comum, o mesmo para LP.

Ao criar outra divisão a partir de 4,8 cm (Figura 26) outras duas regiões são formadas. À esquerda há 46 *Versicolors* e três *Virgínicas* com CP<=4,8cm, enquanto do lado direito as quatro *Versicolors* restantes e as outras 47 *Virgínicas*. Coincidindo com as condições da Árvore de Decisão feita para classificar *Versicolors*, pois primeiro é verificado se CP é menor, ou igual, a 1,9cm (uma provável *Setosa*), caso contrário, analisa-se CP novamente, verificando em qual região a medida de CP da flor se encontra. Caso CP<=4,8cm, então há maior chance (46 em 49) de ser uma *Versicolor*.

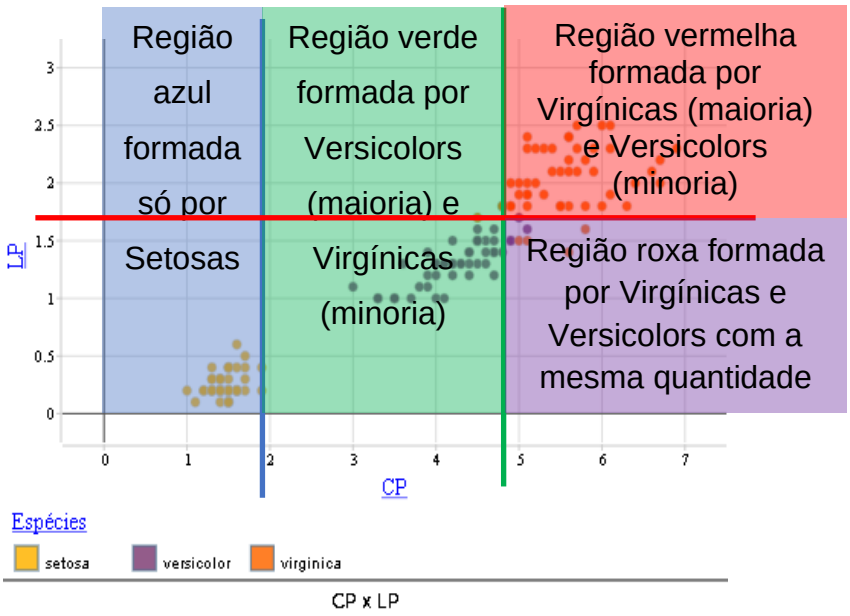
**Figura 26** - Gráfico de Dispersão ilustrando a Árvore de Decisão (Versicolor)




Fonte: Bonangelo (2023)

Agora que a variável CP foi separada em três regiões, é inviável criar mais uma divisão e esperar diferenciar melhor as *Versicolors* e as *Virgínicas*. Portanto, uma divisão pode ser feita na variável LP, para observações com LP <=1,7cm e observações com LP >1,7cm, obtendo mais uma região que pode ser representada como na Figura 27 a seguir.

**Figura 27** - Gráfico de Dispersão ilustrando a Árvore de Decisão (as 3 espécies)




Fonte: Bonangelo (2023)

A Figura 27 ilustra o funcionamento da Árvore de Decisão representada na Figura 15 que iniciou este capítulo. Por fim, vale destacar:

1) a Árvore de Decisão é um algoritmo utilizado na Estatística, na Ciência de Dados e na Inteligência Artificial e é de fácil compreensão;
2) alguns programas e linguagens de programação criam automaticamente uma Árvore de Decisão, baseados em conceitos estatísticos muito mais avançados do que os abordados neste texto. Entretanto, foi possível abordar diversos conceitos intrínsecos ao classificador com os conceitos passíveis de serem abordados nas aulas de Estatística da Educação Básica. Por exemplo os Gráficos de Pontos, de Caixa e de Dispersão, as medidas de posição e de variabilidade, proporção e ilustrações de conceitos aparentemente complexos, como o Coeficiente de Correlação Linear de Pearson;
3) os algoritmos aqui ilustrados são classificadores binários, ou seja, não há a opções distintas, como “talvez”, ou “não sei”, mostrando que o funcionamento desses algoritmos no dia a dia não é independente ou inteiramente autônomo. Há erros, muita estatística e probabilidade para medi-los. Ainda há a necessidade de intervenção humana para planejar e avaliar os novos modelos.

A Estatística e a Matemática para produzir uma Árvore de Decisão está aquém da proposta deste texto, além de não ser o único algoritmo utilizado nas tecnologias citadas no início do texto. Entretanto, ela é uma das opções de mais fácil compreensão e passível de ser esquematizada após uma análise da dados objetiva como a realizada neste texto. Outros algoritmos podem ser estudados bem como explicados a partir de conceitos escolares, uma proposta interessante para trabalhos futuros.

## 5. Considerações Finais

A Educação Estatística se mostra cada vez mais imprescindível para a sociedade imersa em dados, pois todos estão suscetíveis às notícias de diversas fontes, sejam elas confiáveis ou não. Por isso, criar discernimento é fundamental e é possível através da melhor compreensão do que se faz com esses tais dados.

Áreas como a Inteligência Artificial e a Ciência de Dados se apresentam nos tempos atuais e poucos sabem suas abrangências. Ambas usam dados e conseguem exercer grande influência no comportamento humano. Portanto, o entendimento da atualidade deve ser embasado em conhecimentos estatísticos.

A Estatística está presente na escola em todas as fases da Educação Básica, segundo a Base Nacional Comum Curricular (BNCC). Através da Educação Estatística, há estudos sobre o ensino e o aprendizado da Estatística, considerando características afetivas e cognitivas (CAZORLA; KATAOKA; SILVA, 2010). A partir desses conhecimentos, seria interessante a melhora do entendimento das tecnologias atuais.

A busca pelo entendimento das tecnologias atuais a partir de conhecimentos estatísticos estudados na Educação Básica foi a questão proposta no início deste trabalho: “De que forma os conhecimentos estatísticos aprendidos na Educação Básica podem explicar o funcionamento, uso e possibilidades das tecnologias atuais, como a Inteligência Artificial?”.

Para responder o questionamento principal deste texto foram apresentadas as definições e uma breve contextualização das áreas da Estatística, Educação Estatística, Inteligência Artificial, Ciência de Dados e *Big Data*. Entender essas áreas não é somente estudar as questões técnicas, mas sim ter ciência das questões sociais influenciadas por cada uma, como discutido no Capítulo 1.

O Capítulo 2 aborda o início do estudo da Estatística na escola. Um planejamento mundial, adentrando o Brasil na década de 1950, sendo as aulas de Estatística na Educação Básica normalizadas para todo o território nacional na década de 1990 através dos Parâmetros Curriculares Nacionais (PCNs) e obrigatória para todos da Educação Básica segundo o documento conhecido por Base Nacional Comum Curricular (BNCC).

A BNCC, vigente atualmente, é composta por habilidades e competências condizentes com os conteúdos a serem estudados na Estatística abordada na

Educação Básica. Esse documento também discorre sobre a importância do uso de tecnologias para o ensino e aprendizado, mesmo sem definir claramente o que e como utilizá-las.

As tecnologias para o ensino e aprendizagem de Estatística na escola são abordadas por diversos autores (WILD; UTTS; HORTON, 2018), (FRANÇOIS; MONTEIRO, 2018), (BEN-ZVI, 2017), (BIEHLER, 2019), (McNAMARA, 2018). As principais considerações dos pesquisadores são com o uso para facilitar processos mecânicos permitindo mais tempo para reflexão e desenvolvimento de habilidades relacionadas à Estatística; facilitar a interatividade entre representações; auxiliar no estudo de bases de dados multivariadas.

No capítulo 3 foram exibidas duas Análise Exploratórias de Dados com o auxílio do *software* CODAP, evidenciando diversos assuntos possíveis de serem abordados na Educação Básica, como os gráficos univariados (Gráfico de Pontos e de Caixa), as Medidas de Posição (Média, Mediana, Mínimo, Máximo, 1º e 3º Quartis), as de Variação (Amplitude, Intervalo Interquartil), gráfico bivariado (Gráfico de Dispersão), a Razão de Contagem entre os Quadrantes e o Coeficiente de Correlação Linear de Pearson.

Por fim, a Árvore de Decisão foi abordada no capítulo 4. O objetivo de estudar a Árvore de Decisão é para investigar a estrutura preditiva da base de dados e criar classificadores mais assertivos (BREIMAN et al., 1984). Ela é utilizada na Estatística, na Inteligência Artificial, na Ciência de Dados assim como na Educação Básica, esse último é o caso do trabalho de Biehler et al. (2020).

Biehler et al. (2020) estudaram a Árvore de Decisão para classificar uma variável qualitativa a partir de outras também qualitativas, enquanto este texto abordou a classificação de uma variável qualitativa a partir de variáveis quantitativas.

Para construir uma Árvore de Decisão é necessário conhecimentos avançados de Matemática e Estatística, porém, o CODAP junto com o *plugin* Arbor possibilitou a criação manual de modelos, de forma simplificada, assim como a exploração das métricas para avaliar a qualidade das Árvores criadas.

A Matriz de Confusão, a acurácia e a precisão são meios para verificar a eficácia de classificadores, como as Árvores de Decisão. Para entendimento desses assuntos, basta a leitura de tabelas e de proporções, ambos abordados na Educação Básica.

Logo, a partir de gráficos, medidas estatísticas, leitura de tabelas e proporções simples foi possível explorar o algoritmo da Árvore de Decisão; também é importante ressaltar o funcionamento probabilístico desse algoritmo, pois há chances de classificar de forma correta ou não, diferente da perspectiva determinística que algoritmos de Inteligência Artificial aparentam possuir.

## 6. Referências


ANDERSON, Edgar. The Species Problem in Iris. **Annals of the Missouri Botanical Garden**, Missouri, Estados Unidos, v. 23, n. 3, p. 488, 1936. Disponível em: https://www.jstor.org/stable/pdf/2394164.pdf?refreqid=excelsior%3A86b06cba398fa7168204c54efb3dc288. Acesso em: 18 fev. 2021.

BARGAGLIOTTI, Anna *et al*. **Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)**: A Framework for Statistics and Data Science Education. 2. ed. Virginia: National Council of Teachers of Mathematics, 2020. 116 p. ISBN 978-1734223514.

BECKER, Richard A. **A Brief History of S**. Canadá, 1994. Disponível em: http://www.math.uwaterloo.ca/~rwoldfor/software/R-code/historyOfS.pdf. Acesso em: 20 set. 2020.

BEN-ZVI, D.; FRIEDLANDER, A. Statistical Thinking in a Technological Environment. In: IASE ROUNDTABLE CONFERENCE, 1996, Granada, Espanha. *Anais*... LOCAL: IASE, 1996. p. 45-55

BEN-ZVI, D. Big data inquiry: Thinking with data. *In*: FERGUSON, R. *et al*. **Innovating Pedagogy 2017**: Exploring new forms of teaching, learning and assessment, to guide educators and policy makers. 6. ed. Inglaterra: Institute of Educational Technology, 2017. p. 32-36. ISBN 9781473024328. Disponível em: https://iet.open.ac.uk/file/innovating-pedagogy-2017.pdf. Acesso em: 28 nov. 2019.

BIEHLER, Rolf. Software for learning and for doing statistics and probability - Looking back and looking forward from a personal perspective. In: CONGRESO INTERNACIONAL VIRTUAL DE EDUCACIÓN ESTADÍSTICA (CIVEEST), 3., 2019, Granada, Espanha. **Anais**.... Granada, Espanha: 2019. p. 1-14.

BNCC, **Base Nacional Comum Curricular**: Ensino Médio. Brasil: 2018. p. 576. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=85121-bncc-ensino-medio&category_slug=abril-2018-pdf&Itemid=30192 Acesso em: 8 out. 2019

boyd d.; Crawford, K. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. **Information, Communication & Society**, Reino Unido, v. 15, n. 5, p.662-679, 2012. DOI http://dx.doi.org/10.1080/1369118X.2012.678878

BRASIL. MEC. Secretaria de Educação Média e Tecnológica. **Parâmetros curriculares nacionais (Ensino Médio)**. Parte III - Ciências da Natureza, Matemática e suas Tecnologias. Brasília, 2000a.

_____. MEC. Secretaria de Educação Média e Tecnológica. **Parâmetros curriculares nacionais (Ensino Médio)**. Parte II - Linguagens, Códigos e suas Tecnologias. Brasília, 2000b.

_____. MEC. Secretaria de Educação Média e Tecnológica. **PCN+ Ensino Médio:** Orientações educacionais complementares aos parâmetros curriculares nacionais. Ciências da Natureza, Matemática e suas Tecnologias. Brasília, 2002.

BIEHLER, Rolf; FLEISCHER, Yannik. Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. **Teaching Statistics**, Estados Unidos da América, p. 133-142, 2021.

BIEHLER, Rolf et al. Data Science Education in Secondary Schools: Teaching and Learning Decision Trees with CODAP and Jupyter Notebooks as an Example of Integrating Machine Learning into Statistics Education. In: PROCEEDINGS OF THE STATISTICA EDUCATION (IASE), 3., 2019, Voorbor, Holanda. Anais.... Holanda: 2020. p. 1-6.

BIEHLER, Rolf. Software for learning and for doing statistics and probability - Looking back and looking forward from a personal perspective. *In*: CONGRESO INTERNACIONAL VIRTUAL DE EDUCACIÓN ESTADÍSTICA (CIVEEST), 3., 2019, Granada, Espanha. **Anais**[...]. Granada, Espanha, 2019. p. 1-14.

BONANGELO, Rafael Vieira; CORDANI, Lisbeth Kaiserlian. Decifrando o gráfico de caixa (Box-plot) com uso do CODAP - relato de uma oficina. In: International Conference on Teaching Statistics (ICTOS 11), 11., 2022, Rosario, Argentina. Anais [...]. São Paulo, Brasil, 2022. p. 1-6.

BREIMAN, L. Statistical Modeling: The Two Cultures. **Statistical Science**, Estados Unidos, v. 16, n. 3, p. 199-231, 2001.

BRUCE, Peter; BRUCE, Andrew. Estatística Prática para Cientistas de Dados: 50 conceitos essenciais. Brasil: Alta Books, 2019. 376 p.

BUDDE, Lea et al. Data science education in secondary school: how to develop statistical reasoning when exploring data using CODAP. *In*: IASE Roundtable, 2020, Holanda. **Proceedings** [...]. Alemanha, 2020, p. 1-6.

BUSH, Heather M.; DADDYSMAN, Jennifer; CHARNIGO, Richard. Improving Outcomes with Bloom's Taxonomy: From Statistics Education to Research Partnerships. **Journal of Biometrics & Biostatistics**, Bélgica, v. 5, n. 4, p. 1-3, 2014. DOI 10.4172/2155-6180.1000e130. Disponível em: https://www.hilarispublisher.com/open-access/improving-outcomes-with-blooms-taxonomy-from-statistics-education-to-research-partnerships-2155-6180.1000e130.pdf. Acesso em: 29 maio 2020.

CAEM. **Centro de Aperfeiçoamento do Ensino de Matemática**. São Paulo, Brasil, 2020. Disponível em: https://www.ime.usp.br/caem/index.php. Acesso em: 13 jan. 2020.

CAEM IME USP. Decifrando o gráfico boxplot (diagrama de caixa) no Ensino Médio (Parte 1). Brasil, 2020a. Disponível em: https://youtu.be/2cpfvtZ2D5o. Acesso em: 21 dez. 2020.

CAEM IME USP. Decifrando o gráfico boxplot (diagrama de caixa) no Ensino Médio (Parte 2). Brasil, 2020b. Disponível em: https://youtu.be/FIo3YMLU5Eg. Acesso em: 21 dez. 2020.

CALDAS, M. S.; CLAUDINO SILVA, E. C. Fundamentos e aplicação do Big Data: como tratar informações em uma sociedade de yottabytes. **Bibliotecas Universitárias: pesquisas, experiências e perspectivas**, v. 3, n. 1, 13 abr. 2016.

CANQUERINO, M**. Artista plástico ensina Inteligência Artificial a gerar desenho de observação**. Brasil, 5 nov. 2019. Disponível em: https://jornal.usp.br/universidade/eventos/artista-plastico-ensina-inteligencia-artificial-a-gerar-desenho-de-observacao/. Acesso em: 11 dez. 2019.

CARVALHO, J. B. P. F. de. Propostas curriculares de Matemática. In: BARRETO, E. S. de S. (coord.). **As propostas curriculares oficiais**. São Paulo: Fundação Carlos Chagas, 1995. p. 46-58.

CAZORLA, I. M.; KATAOKA, V. Y.; SILVA, C. B. da. Trajetória e perspectivas da Educação Estatística no Brasil: um olhar a partir do GT12. In: LOPES, C. E.; COUTINHO, C. Q. S.; ALMOLOUD, S. A. (Org.). **Estudos e reflexões em Educação Estatística**. Campinas: Mercado de Letras, 2010. p. 19-44.

CHAMBERS, J. M. Greater or lesser statistics: a choice for future research. **Statistics and Computing**, Springer, Estados Unidos, v. 3, n. 4, p. 182-184, dez. 1993.

CHANCE, B. *et al*. The Role of Technology in Improving Student Learning of Statistics. **Technology Innovations in Statistics Education**, Califórnia, Estados Unidos, v. 1, p. 1-26, 2007.

CLEVELAND, W. S. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. **International Statistical Review**, México, v. 69, n. 1, p. 21-26, abr. 2001. DOI 10.2307/1403527. Disponível em: https://www.jstor.org/stable/1403527. Acesso em: 29 ago. 2019.

CODAP. The CODAP Community. Estados Unidos da América, 2023. Disponível em: https://codap.concord.org/contributors/. Acesso em: 17 out. 2023.

CORDANI, L. K. Caminhos da educação estatística ao longo do tempo: uma leitura pessoal. **Jornal Internacional de Estudos em Educação Matemática**, Paraná, v. 8, n. 3, p. 157-182, 2014. Disponível em: https://revista.pgsskroton.com/index.php/jieem/article/view/3043. Acesso em: 4 jun. 2020.

(CODAP) THE CONCORD CONSORTIUM. **The CODAP Community**. Estados Unidos, 2023. Disponível em: https://codap.concord.org/contributors/. Acesso em: 01 dez. 2023.

(CODAP) THE CONCORD CONSORTIUM. **Getting Started with Graphs**. Estados Unidos, 2021. Disponível em: https://codap.concord.org/help/basics/graphs. Acesso em: 14 jan. 2021.

(CODAP) THE CONCORD CONSORTIUM. **Common Online Data Analysis Platform (CODAP)**. Estados Unidos, 2020. Disponível em: https://codap.concord.org/. Acesso em: 27 nov. 2020a.

(CODAP) THE CONCORD CONSORTIUM. **CODAP Toolbar**. Estados Unidos, 2020. Disponível em: https://codap.concord.org/about/. Acesso em: 11 dez. 2020b.

DONOHO, D. 50 Years of Data Science, Journal of Computational and Graphical Statistics. **Journal of Computational and Graphical Statistics**, Estados Unidos, v. 26, n. 4, p. 745-766, 19 dez. 2017. DOI 10.1080/10618600.2017.1384734. Disponível em: https://www.tandfonline.com/doi/pdf/10.1080/10618600.2017.1384734?needAccess=true. Acesso em: 28 jul. 2019.

DUARTE, Roberta. Como os robôs superaram os humanos no xadrez?. **Folha de São Paulo**, São Paulo, p. 1, 7 abr. 2021.

ENGEL, Joachim. Exploring Civic Statistics With CODAP. In: Challenges and Innovations in Statistics Education CHALLENGES AND INNOVATIONS IN STATISTICS EDUCATION, 1., 2017, Szeged, Hungria. *Proceedings*... Szeged, Hungria: Universidade de Szeged, 2018.

ENGEL, Joachim. Statistical literacy for active citizenship: A call for data science education. **Statistics Education Research Journal**, Estados Unidos, v. 16, n. 1, p. 44-49, 2017.

ENGEL, Joachim; ERICKSON, Tim; MARTIGNON, Laura. Teaching and Learning about Tree-Based Methods for Exploratory Data Analysis. In: INTERNACIONAL CONFERENCE ON TEACHING STATISTICS, 10., 2018, Kyoto, Japão. *Proceedings*... Voorburg, Holanda: International Statistical Institute, 2018. p. 1-6

ENHANCING STATISTICS TEACHER EDUCATION WITH E-MODULES. **CODAP Tutorial: Introduction to using CODAP**. Brasil, 2018. Disponível em: https://www.youtube.com/watch?v=aD5tLWld98w&list=PLq_mgFaS8OGYF1cKF4IrBl_0RwB_SWQk1. Acesso em: 1 dez. 2023.

FERRAZ, A. P. do C. M.; BELHOT, R. V. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. **Gestão & Produção**, São Carlos, v. 14, n. 2, p. 421-431, 2010.

FIENBERG, S. E. What Is Statistics?. **Annual Review of Statistics and Its Application**, Estados Unidos, v. 1, p. 1-9, 2014. DOI https://doi.org/10.1146/annurev-statistics-022513-115703. Disponível em: https://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-022513-115703. Acesso em: 22 nov. 2019.

FRANÇOIS, Karen; MONTEIRO, Carlos. Big Data Literacy. In: INTERNACIONAL CONFERENCE ON TEACHING STATISTICS, 10., 2018, Kyoto, Japão. *Proceedings*... Voorburg, Holanda: International Statistical Institute, 2018. p. 1-6

FRISCHEMEIER, Daniel *et al*. A first introduction to data science education in secondary schools:: Teaching and learning about data exploration with CODAP using survey data. **Teaching Statistics**, Estados Unidos da América, p. 182-189, 2021.

FRISCHMEIER, Daniel; BIEHLER, Rolf; ENGEL, Joachim. Competencies and dispositions for exploring micro data with digital tools. In: ROUNDTABLE CONFERENCE OF THE INTERNATIONAL ASSOCIATION OF STATISTICS EDUCATION, 13., 2016, Berlim, Alemanha. *Proceedings*... Berlim, Alemanha: International Statistical Institute e International Association of Statistics Education, 2016. p. 1-10

GOOGLE. **Inside Google Translate**. Estados Unidos, 9 jul. 2010. Disponível em: https://www.youtube.com/watch?v=_GdSC1Z1Kzs. Acesso em: 12 dez. 2019.

GOULD, Robert. Data Literacy is Statistical Literacy. **Statistics Education Research Journal**, Estados Unidos, v. 16, n. 1, p. 22-25, 2017.

HAYASHI, C. What is Data Science? Fundamental Concepts and a Heuristic Example. *In*: HAYASHI, C.; YAJIMA, K.; BOCK, HH.; OHSUMI, N.; TANAKA, Y.; BABA, Y. (ed.). **Data Science, Classification, and Related Method**. Tóquio, Japão: Springer, 1998. p. 40-51. ISBN 978-4-431-65950-1. DOI https://doi.org/10.1007/978-4-431-65950-1_3. Disponível em: https://link.springer.com/content/pdf/10.1007%2F978-4-431-65950-1.pdf. Acesso em: 19 ago. 2019.

HOLMES, Peter. Correlation: From Picture to Formula. **Teaching Statistics**, Estados Unidos, v. 23, n. 3, p. 67-71, 2001.

IBGE. **O IBGE**. Brasil, 2023. Disponível em: https://www.ibge.gov.br/acesso-informacao/institucional/o-ibge.html. Acesso em: 1 dez. 2023.

IBM. Analytics: The real-world use of big data: How innovative enterprises in the midmarket extract value from uncertain data. **IBM Global Business Services**, Nova Iorque, Estados Unidos, p. 1-4, abr. 2013. Disponível em: https://www.ibm.com/downloads/cas/E4BWZ1PY. Acesso em: 15 ago. 2019.

JOHN, V. The Term "Statistics.". **Journal of the Statistical Society of London**, Londres, Inglaterra, v. 46, n. 4, p. 656-679, dez. 1883. DOI 10.2307/2979311 https://www.jstor.org/stable/2979311. Disponível em: https://www.jstor.org/stable/i349460. Acesso em: 24 nov. 2019.

KAUFMAN, D. **Os algoritmos de inteligência artificial podem ser éticos?**: O uso da IA para decisões pré-programadas pelos humanos traz novos desafios éticos para a sociedade. Brasil, ago. 2019. Disponível em: https://epocanegocios.globo.com/colunas/IAgora/noticia/2019/08/os-algoritmos-de-incteligencia-artificial-podem-ser-eticos.html. Acesso em: 11 dez. 2019.

KENDALL, Maurice G. Introduction. *In*: KENDALL, Maurice G. **The advanced theory of statistics**. 2. ed. rev. Londres: Charles Griffin and Company, 1945. v. I, p. xi-xii.

KISH, L. Chance, Statistics, and Statisticians. **Journal of the American Statistical Association**, Estados Unidos, v. 73, n. 361, p. 1-6, mar. 1978.

KLUTKA, Justin; ACKERLY, Nathan; MAGDA, Andrew J. **Artificial Intelligence in Higher Education**: Current Uses and Future Applications. Estados Unidos da América: Wiley, 2018. 31 p.

KRATHWOHL, D. R. A Revision of Bloom's Taxonomy: An Overview. **Theory into Practice**, Columbus, v. 41, n. 4, p. 212-218, 2002.

LÓPEZ, G. M.; URREA, W. H. S. Pedagogía del dato: perspectiva desde la enseñanza de la estadística en la sociedad del dato. **Análisis**, Bogotá, v. 51, n. 94, p. 141-158, 2019. DOI https://doi.org/10.15332/s0120-8454.2019.0094.07. Disponível em: http://ojs3.usantotomas.edu.co/revistas_ustacolombia/index.php/analisis/article/view/4306/pdf. Acesso em: 25 jun. 2020.

LIU, B. et al. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. **Association for Computing Machinery**, Seul, Coreia do Sul, p. 783-791, out. 2018. DOI https://doi.org/10.1145/3240508.3240587. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3240508.3240587. Acesso em: 12 dez. 2019.

LOPES, C. A. E. **A probabilidade e a estatística no ensino Fundamental: uma análise curricular**. 1998. 127 p. Dissertação (Mestre em Educação) - Faculdade de Educação, Universidade Estadual de Campinas, Campinas, 1998.

MACGILLIVRAY, H. Statistics and data science are NOT branches of mathematics—or of any other discipline. **Teaching Statistics**, Estados Unidos, ano 2, v. 41, p. 41, 23 abr. 2019. DOI https://doi.org/10.1111/test.12197. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1111/test.12197. Acesso em: 15 ago. 2019.

MAGALHÃES, M. N.; LIMA, A. C. P. de. **Noções de Probabilidade e Estatística**. 7. ed. rev. São Paulo: Editora da Universidade de São Paulo, 2015. 408 p. ISBN 978-85-314-0677-5.

MARQUARDT, D. W. The Imporlance of Statistician. **Journal of the American Statistical Association**, Chicago, Estados Unidos, v. 82, n. 397, p. 1-7, Mar. 1987.

McKinsey. Big data: The next frontier for innovation, competition, and productivity. **McKinsey Global Institute**, Estados Unidos, p. 1-143, 1 jun. 2011. Disponível em: https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf. Acesso em: 15 ago. 2019.

MCNAMARA, Amelia. Key attributes of a modern statistical computing tool. **The American Statistician**, Estados Unidos, p. 1-30, 2018. DOI https://doi.org/10.1080/00031305.2018.1482784. Disponível em: https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1482784. Acesso em: 27 nov. 2020.

MINISTÉRIO DA EDUCAÇÃO (MEC). **Base Nacional Comum Curricular**. Brasil, [2018]. Disponível em: http://basenacionalcomum.mec.gov.br/a-base. Acesso em: 17 dez. 2019.

MOORE, D. S. Statistics Among the Liberal Arts. **Journal of the American Statistical Association**, Estados Unidos, v. 93, n. 444, p. 1253-1259, 1998.

MORETTIN, Pedro A.; SINGER, Julio M. **Estatística e Ciência de Dados**. Brasil: São Paulo, 2021. 356 p. Disponível em: https://www.ime.usp.br/~jmsinger/MAE0217/cdados2021junho01.pdf. Acesso em: 12 jun. 2021.

MORETTO, Vasco Pedro. A prova operatória: ressignificando a taxonomia de Bloom. In: MORETTO, Vasco Pedro. **Prova**: Um momento privilegiado de estudo, não um acerto de contas. 9. ed. Rio de Janeiro: Lamparina, 2010. cap. 10, p. 153-184. ISBN 978-85-98271-69-9.

PENTLAND, A. S. The data-driven society. **Scientific American**, Estados Unidos, v. 309, n. 4, p. 78-83. 2013

PEREIRA, Roberta Duarte. **Black Hole Weather Forecasting Using Deep Learning.** 2020. Dissertação (Mestrado em Astronomia) - Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, 2020. doi:10.11606/D.14.2020.tde-23102020-170517. Acesso em: 2023-12-01.

PIERRO, B. de. O mundo mediado por algoritmos: Sistemas lógicos que sustentam os programas de computador têm impacto crescente no cotidiano. 266. ed. São Paulo, Brasil: Pesquisa Fapesp, abr. 2018. Disponível em: https://revistapesquisa.fapesp.br/2018/04/19/o-mundo-mediado-por-algoritmos/. Acesso em: 11 dez. 2019.

PINHEIRO, J. Obra de arte criada por inteligência artificial será leiloada pela primeira vez. Brasil, 27 ago. 2018. Disponível em: https://canaltech.com.br/arte/obra-de-arte-criada-por-inteligencia-artificial-sera-leiloada-pela-primeira-vez-121052/. Acesso em: 11 dez. 2019.

RIEDER, G.; SIMON, J. Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. **Big Data & Society**, Califórnia, v.3, p. 1-6, 2016. DOI 10.1177/2053951716649398

ROBINSON, D. **What's the difference between data science, machine learning, and artificial intelligence?**. Estados Unidos da América, 9 jan. 2018. Disponível em: http://varianceexplained.org/r/ds-ml-ai/. Acesso em: 15 ago. 2019.

RODRIGUES, Márcio Urel; SILVA, Luciano Duarte da. Disciplina De Estatística Na Matriz Curricular Dos Cursos De Licenciatura Em Matemática No Brasil. **REVEMAT**, Florianópolis, SC, v. 14, p. 1-21, 2019.

RUSSEL, S. J.; NORVIG, P. **Artificial intelligence**: A modern Approach. 3. ed. Estados Unidos: Pearson, 2016. 1132 p. ISBN 9781292153964.

SANGIORGI, O. Propostas curriculares de Matemática e Estatística para cursos normais. **Atualidades Pedagógicas**, São Paulo, n. 41, p. 20-26, 1957.

SILVA, M. R. I. S. da; VALENTE, W. R. Da estatística educacional para a estatística: das práticas profissionais a um campo disciplinar acadêmico. **Educação e Pesquisa**, São Paulo, v. 41, n. 2, p. 443-459, 2015. DOI https://doi.org/10.1590/s1517-97022015041876. Disponível em: https://www.scielo.br/pdf/ep/v41n2/1517-9702-ep-41-2-0443.pdf. Acesso em: 20 jun. 2020.

TUKEY, John. **Exploratory Data Analysis**. 1. ed. Estados Unidos da América: Pearson, 1977. 712 p.

TUKEY, J. W. The Future of Data Analysis. **The Annals of Mathematical Statistics**, Ohio, v. 33, n. 1, p. 1-14, 1962.

TUKEY, John W. The technical tools of statistics. **The American Statistician**, Estados Unidos, v. 19, n. 2, p. 23-28, 1965.

TUNES, S. Imitação do cérebro: Inteligência artificial nasceu como campo científico nos anos 1940 em consequência de estudos matemáticos. **Pesquisa Fapesp**, São Paulo, Brasil, ed. 275, p. 24-25, 21 jan. 2019a. Disponível em: https://revistapesquisa.fapesp.br/wp-content/uploads/2019/01/018-025_CAPA-Intelig%C3%AAncia-Artificial_275_NOVO4-1.pdf. Acesso em: 11 dez. 2019.

_____. Terreno Fértil para a Inteligência Artificial: Com a crescente evolução da ferramenta tecnológica, consórcio de pesquisadores cria instituto dedicado a estabelecer parcerias entre universidades e empresas. **Pesquisa Fapesp**, São Paulo, Brasil, ed. 275, p. 18-24, 21 jan. 2019b. Disponível em: https://revistapesquisa.fapesp.br/wp-content/uploads/2019/01/018-025_CAPA-Intelig%C3%AAncia-Artificial_275_NOVO4-1.pdf. Acesso em: 11 dez. 2019.

_____. **Algoritmos parciais**: Como a inteligência artificial absorve padrões discriminatórios e o que a ciência pode fazer para evitar essas distorções. São Paulo, Brasil: Pesquisa Fapesp, 15 nov. 2019c. Disponível em: https://revistapesquisa.fapesp.br/2019/11/15/algoritmos-parciais/. Acesso em: 11 dez. 2019.

UNIVERSIDADE DE HELSINKI. **Elements of AI**. Finlândia, 2019. Disponível em: https://www.elementsofai.com/. Acesso em: 14 jan. 2020.

VALENTE, W. R. NO TEMPO EM QUE NORMALISTAS PRECISAVAM SABER ESTATÍSTICA. **Revista Brasileira de História da Matemática**, Brasil, n. 1, p. 357-368, 2007.

VIALI, L. O Ensino de Estatística e Probabilidade nos Cursos de Licenciatura em Matemática. In: Simpósio Nacional de Probabilidade e Estatística, 18., 2008, Estância de São Pedro. *Anais*... Estância de São Pedro, SP: ABE, 2008.

WENDLER, T.; GRÖTTRUP, S. Classification Models. In: WENDLER, T.; GRÖTTRUP, S. **Data Mining with SPSS Modeler**: Theory, Exercises and Solutions. Suíça: Springer, 2016. cap. 8, p. 713-984. ISBN 978-3-319-28707-2.

WILD, C. J.; UTTS, J. M.; HORTON, N. J. What Is Statistics?. *In*: BEN-ZVI, D.; MAKAR, K.; GARFIELD, J. (ed.). **International Handbook of Research in Statistics Education**. Suíça: Springer, 2018. cap. 1, p. 5-36. ISBN 978-3-319-66193-3.

YATES, F.; Theory and Practices in Statistics. **Royal Statistical Society**, Londres, p. 463-475, 1968.