# Aprendizagem de máquina básico: uma abordagem teórico-prática com Python

Giseldo Neo

versao alpha 0.1

25 de julho de 2024

# Sumário

# Capítulo 1

# Inteligência artificial

A "inteligência" tem definições que dependem do contexto. Isso pode trazer certa confusão no entendimento e delimitação do tema. Menos abrangente, porém mais confuso ainda, é o termo "inteligência artificial". Portanto, dado as diversas definições de inteligência artificial (IA), ou *artificial inteligence* em inglês, vamos delimitar um pouco o escopo da nossa inteligência em questão.

Nós humanos somos da espécie Homo-Sapiens. Espero que o leitor ainda o seja, pois esse texto pode estar sendo processado para treinar o mais novo modelo de IA, como por exemplo, o Gemini da Google, o chatGPT da openAI ou o copilot da Microsoft.

Homo-Sapiens vem do latim e significa homem sábio [**?**]. A importância da sapiência (que é um sinônimo de inteligência) é tamanha que define a nossa própria espécie. Porém, neste contexto consideramos que um animal, como o gato ou cachorro, também são dotados de inteligência. Mas não somente os mamíferos, uma abelha, então nem se fala, ela é praticamente uma cientista [**?**]. Portanto, seremos mais contidos e reservados quanto ao significado do termo inteligência.

O que confunde bastante é que "inteligência" e "artificial" são palavras que têm significado implícito para pessoas que não são da área de computação, naturalmente surge o desejo de médicos, advogados, engenheiros (só para citar alguns) de verificar como a "inteligência artificial" pode ser inserida na sua rotina diária. O meu dentista já quis saber como a IA iria afetar seus procedimentos odontológicos. Porém, ninguém nunca me perguntou em como a "Transformada de Fourier" poderia melhorar o seu dia-a-dia, mesmo sabendo que a transformada já é utilizada em vários domínios do conhecimento e com entusiasmo [**?**].

A "inteligência artificial" da computação está mais relacionada com a capacidade de realizar coisas que seres inteligentes (tais como, um gato, um bebê, uma abelha, ou um humano) realizam, como por exemplo puxar a mão (ou pata) instantaneamente ao tocar em uma superfície quente, realizar uma prova objetiva de anatomia, ou elaborar um recurso para a anulação de uma questão de concurso. Se conseguimos que

(a) Jhon MacCarthy

(b) Alan Turing

Figura 1.1: Jhon Maccarthy e Alan Turing

programas realizem ações realizadas por entidades dotadas de inteligencia, e realizamos isso de forma computacional, estamos próximos do significado desejado de "inteligência artificial".

O livro de Russel e Norvig é um dos livros mais lidos em todas as universidades do mundo e tem uma boa definição sobre o tema: "O campo da inteligência artificial [...] tenta não apenas compreender, mas também construir entidades inteligentes" (tradução nossa) [**?**]. Em outras palavras temos o audacioso objetivo de construir agentes dotados de inteligência.

A origem do termo "inteligência artificial", neste contexto, é atribuída a John McCarthy, professor de Matemática da Universidade Dartmouth College [**?**] (Figura 1.1), ele organizou uma conferência com duração de oito semanas com outros colegas em 1956, alguns anos após a segunda guerra, e desde então o termo vem sendo utilizado para designar parte de conteúdos estudados em ciência da computação. Porém, um pouco antes, o artigo seminal de Alan Turing, com quem trabalhou em conjunto, apresentava reflexões sobre a inteligência que uma máquina poderia possuir [**?**]. Temos vários exemplos de entidades dotadas de inteligência documentados em nossa sociedade, e remotam tempos bem antigos.

Foi na década de 1970 que o uso da IA começou a ser mais difundido. Uma das primeiras abordagens com relativo sucesso foram os Sistemas Especialistas (SE). Eles dependiam dos especialistas do domínio para transformar o conhecimento tácito (baseado em sua experiência) em explicito (formalizado, documentado), que era então codificado na forma de regras em lógica formal. O processo de aquisição desse conhecimento acabou sendo um grande obstáculo na adoção em massa dessa abordagem. Veja um exemplo de software que implementa um motor de inferência baseado na teoria dos SE na Figura 1.2

A superação de certas limitações permitiu o avanço de outras técnicas. Alguns dos avanços foram: o aumento da capacidade de processamento e armazenamento dos computadores, a geração de grandes volumes de dados, novidades científicas e

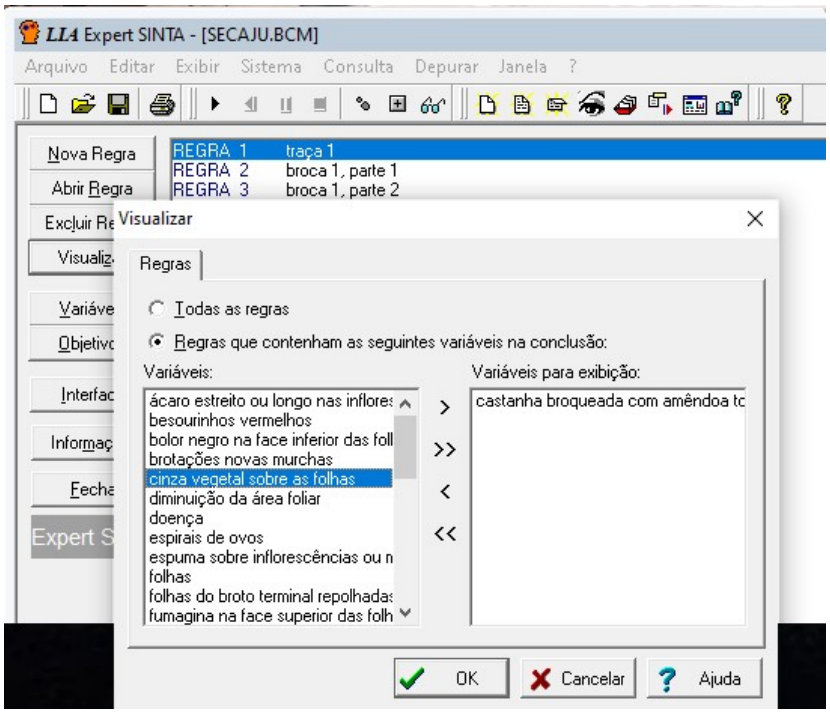
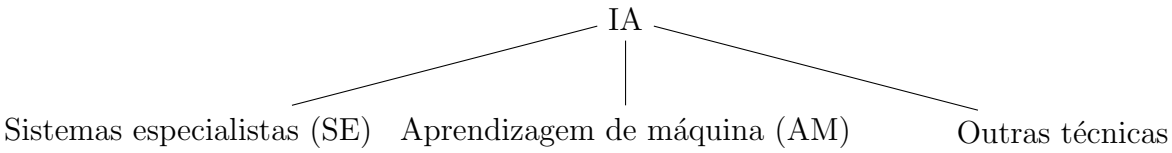



Figura 1.2: ExpertSinta. Uma interface de um Sistema Especialista

tecnológicos, chips supercondutores, eficiência energética, entre outras. A partir destes avanços, uma das técnicas que tem ganhado notoriedade é o Aprendizado de máquina (Figura 1.3).

Figura 1.3: AM é uma parte da IA



# Capítulo 2

# Aprendizado de Máquina

O Aprendizado de Máquina (AM) é uma subárea da IA motivada pelo desenvolvimento de softwares mais independentes da intervenção humana para extração do conhecimento, o que era uma dificuldade nos Sistemas Especialistas. Geralmente aplicações de AM utilizam indução para buscar por modelos capazes de representar o conhecimento existente nos dados.

Na Figura 2.1, é possível identificar alguns usos de AM integrado em diversas atividades cotidianas. São elas, (a) Um smartphone com um assistente de voz fornecendo atualizações meteorológicas. (b) Um sistema de casa inteligente ajustando o termostato com base nas preferências do usuário. (c) Um carro autônomo dirigindo em uma rua movimentada da cidade. (d) Uma plataforma de compras online recomendando produtos a um usuário com base em suas compras anteriores. Essa figura foi criada inclusive com o chatGPT, um chatbot que ganhou notoriedade sendo uma dos aplicativos que mais ganhou usuários rapidamente no mundo.

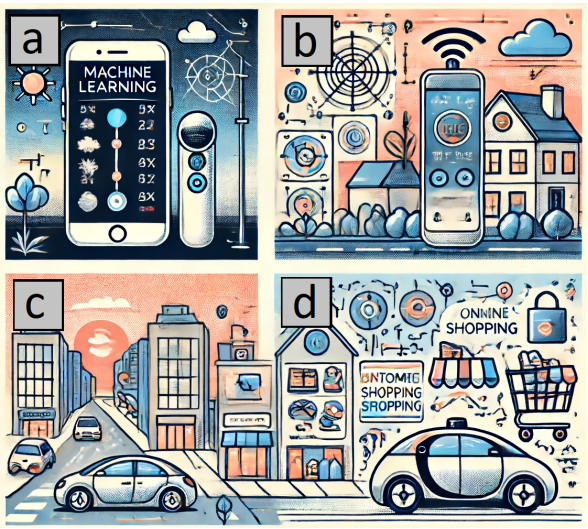



Figura 2.1: Exemplos AM

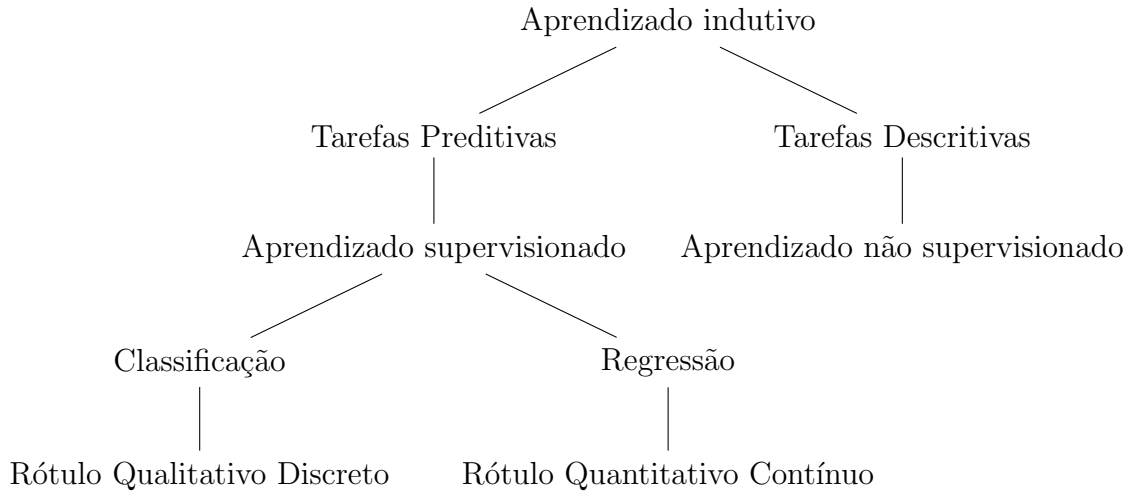
As tarefas de aprendizado de máquina podem ser divididas entre tarefas **pre-**

**ditivas** e **descritivas**.

As tarefas de aprendizado preditivas visam inferir o atributo alvo de uma nova entrada a partir da exposição prévia aos dados durante o treinamento do modelo.

As tarefas descritivas buscam extrair padrões e correlações, além disso, não existe esta distinção entre atributos alvo e preditivos, todos são possíveis preditores em tarefas descritivas.

Figura 2.2: Classificação de AM




Ambas as tarefas podem ser categorizadas sob o conceito de aprendizado indutivo, que é a capacidade de generalizar a partir de exemplos específicos, isto é, do conjunto de dados de treinamento.

Em se tratando de tarefas preditivas, os algoritmos poderão implementar tarefas de **classificação**, nas quais o atributo alvo é **qualitativo discreto**, ou de **regressão**, em que o atributo alvo é **quantitativo contínuo**. Detalhes em relação aos tipo e escalas dos atributos serão apresentados no próximo capítulo.

Já as tarefas descritivas podem ser: agrupamento, que busca por similaridades, associação, que busca por padrões frequentes, e sumarização, que resulta em um resumo do conjunto de dados. No entanto, tarefas descritivas estão fora do escopo deste livro.

# Capítulo 3

# Exemplo

Para exemplificar criaremos um modelo preditivo a partir de um conjunto de dados com 2 atributos preditores (X1 e X2) e um atributo alvo (Y). Existe uma função que gerou os dados de treino e ela é desconhecida. Essa função chamamos de "god function", g(x). Queremos encontrar uma outra função f(x), dentro de um um universo de funções disponíveis que mais se aproxima de g(x). A premissa é que o engenheiro de aprendizagem de máquina não conhece e nunca conhecerá a função g(x) que gerou os dados, mas íra dar um melhor chute para esta função, que se chamará f(x).

Primeiro vamos tentar inferir esta função f(x) com nossa inteligência humana. Em seguida utilizaremos um modelo preditivo e vamos comparar se a técnica de inteligência artificial de aprendizagem de máquina chegou em um resultado similar.

| X1 | X2 | Y |
|---|---|---|
| -2 | -2 | 0 |
| -1 | -1 | 0 |
| 1 | 1 | 1 |
| 2 | 2 | 1 |

Tabela 3.1: Dados Fictícios

Na Tabela 3.1 $X1$ e $X2$, são dois vetores, juntos eles formam uma matriz de preditores $X_{[}pred]$.

Y é um vetor e é o atributo alvo.

Um conjunto de dados de forma geral é uma Matriz Xde dimensão d, que pode ser representada por Xij. Onde i é o enésimo elemento e ja o énesimo atributo.

Utilize a sua intuição. A partir dos dados de treino, para um nova observação (X1=3 e X2=3) qual seria o valor de Y?

# Capítulo 4

# AM e Python

A seguir um exemplo de modelo preditivo em Python. O modelo utiliza o algoritmo SVM e o conjunto de dados iris, que é um conjunto de dados conhecido e bastante utilizado como em demonstrações em outros livros e sites.

Listing 4.1: Exemplo de código que usa AM

```
        from sklearn import svm
        from sklearn.datasets import iris
        iris = load_iris()
        X = iris.data
        y = iris.target
        model = svm.SVC()
        model.fit(X, y)
        model.predict([[2., 2., 2., 2.]])
```

# Capítulo 5

# Estatística Básica

Um conjunto de dados (geralmente para fins de análise) é organizado em uma estrutura tabular no formato de linhas e colunas. Cada coluna é chamado de atributo ou variável e cada linha é chamada de observação, ou registro, ou instância.

A Tabela 5.1 apresenta um exemplo de conjunto de dados com dados de clientes. A primeira linha da tabela define o nome dos atributos, são eles: *Nome*, *Endereço*, *Telefone*, *Salario*, e *Concede crédito*; cada linha abaixo do nome da coluna representa um cliente.

Tabela 5.1: Tabela cliente.

| **Nome** | **Endereço** | **Telefone** | **Salario** | **Concede crédito** |
|---|---|---|---|---|
| Jose Carlos | Rua das alamedas[...] | 222 96666 | 1.000,00 | Sim |
| Alex Borges | Avenida Fernandes[...] | 333 96589 | 2.000,00 | Não |

# Capítulo 6

# Tipo e Escala de Atributos

O objetivo é descrever a tipologia das variáveis para que possamos utilizar corretamente as técnicas estatísticas e de aprendizagem de máquina. Em aprendizado de máquina, os atributos ou variáveis podem ser classificados em diferentes tipos e escalas, o que é importante para escolher o algoritmo mais adequado e preprocessar os dados corretamente.

Para apresentar uma estatística de resumo (tais como, média, mediana ou moda) ou algum gráfico (por exemplo, de barras ou de linhas) temos que conhecer o tipo do atributo para podermos selecionar qual a técnica mais adequada que devemos aplicar. Por exemplo, não podemos calcular a média para todos os atributos do cliente. Qual é a média do nome? ou a média do endereço, ou a média do telefone? Não faz sentido a estatística média para estes atributos. Pode parecer óbvio para estes casos porém nem tanto para outros.

Utilizando uma classificação teórica para esse tipo de atributo, podemos definir qual é o tipo possível de operação. Portanto, vamos acrescentar mais uma informação (teórica) a coluna da tabela para que possamos claramente definir o que deve, ou pode ser feito com ela, chamamos essa informação de tipo de dado, ou tipo de variável, que é o termo mais usado.

Conhecendo estes tipo e sua aplicação correta não cometeremos o erro de por exemplo, calcular a média de uma variável do tipo qualitativa ordinal. Além disso, é interessante reportar nos estudos científicos (artigos) uma tabela com o tipo das variáveis utilizado, pois isso facilita muito o entendimento do conjunto dos dados para o leitor interessado, caso o artigo tenha utilizado um conjunto de dados.

**Tipos de variável**

Uma variável (ou atributo) pode ser de dois tipos, ou quantitativo (também chamado de métrico) ou qualitativo (também chamado de não métrico, ou categórico).

**Escalas da variável (Mensuração)**

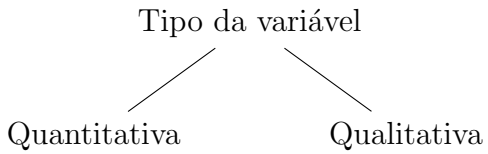
Uma variável tem uma escala de mensuração.

**Escalas da variável (contagem ou categorização)**

Um variável tem uma escala ou de contagem ou de categorização. Se quantitativo de contagem; se qualitativo de categorização.

Nesta seção utilizaremos o termo variável com o significado de atributo. A ciência de dados é uma área interdisciplinar e utiliza técnicas de estatística, computação e matemática. Em estatística, uma variável é uma característica do que foi observado em uma amostra. Essa mesma característica é chamado de parâmetro quando está relacionado a população. Já em computação variável tem uma outra conotação, é um espaço em memória que armazena um valor que pode ser digitado pelo usuário ou inicializado. Porém em análise de dados uma variável também é chamado de atributo, é este conceito que iremos utilizar.

Uma variável (ou atributo) pode ser de alguma forma (a) mensurada. Além disso, ela ainda pode ser ou (b) contada ou (c) categorizada. Chamamos isso de escalas. Além destas três escalas, (a) mensuração, (b) contagem e (c) categorização, classificamos as variáveis em dois tipos, quantitativas ou numéricas, e qualitativas, ou não numéricas. Se a variável for quantitativa ela tem uma escala de contagem, se for qualitativa uma escala de categorização. Ambas tem obrigatoriamente uma escala de mensuração.




Na Tabela ?? as variáveis são: id, issuekey, title, descripton e storypoints. Elas foram registradas nessa tabela após os processos de mensuração, contagem ou categorização.

```
        import pandas as pd
        df = pd.read_csv('7764.csv')
        df.head()
```

| id | issuekey | created | title | description |
|---|---|---|---|---|
| **0** | 29688087 | 2020-01-17 | Update templates for website... | Relates to &232 and #6109 Go... |
| **1** | 29682716 | 2020-01-16 | Make sure that we Capture ... | This was raised in the PM ... |
| **2** | 29644971 | 2020-01-15 | Propose new IA for Brand ... | ## Goals\nPropose new IA for... |
| **3** | 29494181 | 2020-01-10 | Cache ‘node_modules‘ for ... | # UPDATE NOTE: This MR ... |
| **4** | 29437529 | 2020-01-09 | Disable all remaining unn ... | Similar to new site... |

Tabela 6.1: Tabela cliente atualizada com novas colunas e novos registros.

| **Nome** | **Endereço** | **Telefone** | **Estado civil** | **Altura** |
|---|---|---|---|---|
| Giseldo Neo | Rua das Alam[..] | 222 66666 | casado | 1,80 |
| Alex Barros | Avenida Ferna[..] | 333 6589 | solteiro | 1,70 |
| Pedro Alves | Alameda dos Anj[..] | 888 5879 | casado | 1,50 |
| Miguel Peixoto | BR 259, trecho[..] | | solteiro | 1,79 |

É útil definir de qual tipo é determinada variável, pois, existem técnicas adequadas para cada tipo. Para realizar uma análise estatística descritiva, elaborar um gráfico para um artigo científico, ou aplicar uma técnica de pré-processamento em um modelo preditivo é necessário entender de qual tipo é cada variável e quais as suas escalas, pois existem determinadas técnicas para determinados fins. Por isso, vamos entender estas classificações teóricas.

Neste contexto, uma variável pode ser **quantitativa** ou **qualitativa** (Figura **??**). Além disso, uma variável quantitativa também é chamada de métrica, e a variável qualitativa de não métrica ou categórica [**?**].

A variável quantitativa é expressa geralmente como um número. Porém, existem casos em que números também expressam variáveis qualitativas, logo cada caso deve ser analisado individualmente. Já a variável do tipo qualitativa está relacionado ao pertencimento do valor mensurado a um universo. Um exemplo de variável qualitativa é o estado civil do cliente, que pode ser solteiro ou casado. Para continuarmos, vamos atualizar nossa tabela de cliente com duas novas variáveis, estado civil e altura, assim teremos variáveis dos tipos qualitativa e quantitativas na mesma tabela, o que é bem comum (Tabela 6.1).

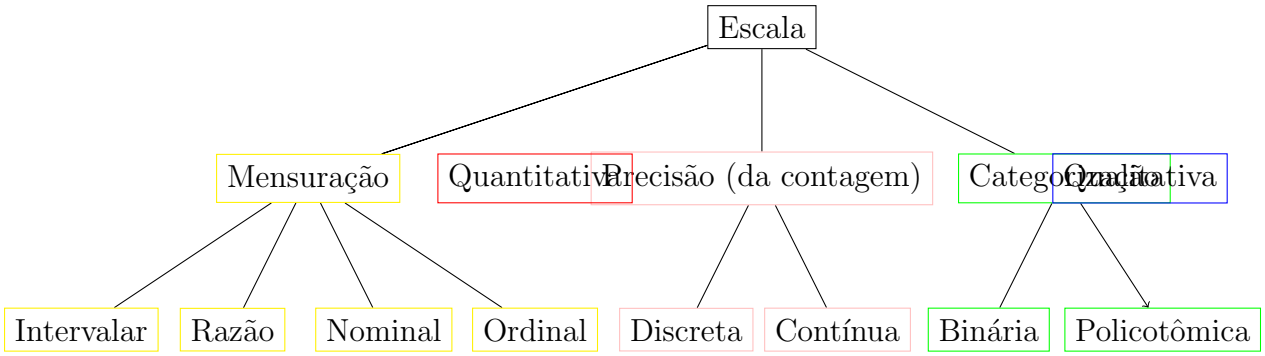
Para exemplificar, tipificaremos (ou classificaremos) as variáveis (ou características) que foram medidas, contadas (com determinado grau de precisão) ou categorizadas dos clientes em uma outra tabela para registrar estes dados sobre as variáveis (se quantitativo ou qualitativo). Veja na Tabela 6.2 o resultado dessa tipificação, para cada variável informamos se a variável é qualitativa ou quantitativa na segunda coluna, já na primeira coluna temos o nome da variável.

No entanto, ainda temos uma outra classificação das variáveis, relacionada as

Tabela 6.2: Tipo das variáveis da tabela cliente.

| **Nome da variável** | **Tipo da variável** |
|---|---|
| Nome | qualitativa |
| Endereço | qualitativa |
| Telefone | qualitativa |
| Estado Civil | qualitativa |
| Altura | quantitativa |

Figura 6.1: Escala da variável




**escalas**. As escalas são 3: **mensuração**, **precisão** (da contagem) e **categorização**. As opções da escala mensuração são: **intervalar** e **razão** para as variáveis quantitativas e **nominal** e **ordinal** para as variáveis qualitativas. Veja na Figura 6.1, o tipo das variáveis e as escalas de mensuração de cada tipo. Além das escalas de mensuração, temos as escalas de precisão e as escalas de categorização. A escala de precisão é utilizada somente nas variáveis quantitativas com as opções **discreta** ou **contínua**. Já a escala de categorização é utilizada somente em variáveis qualitativas, suas opções são **binária** ou **policotômica**.

Portanto se a variável for quantitativa ela somente poderá ser ou intervalar ou razão, além disso, discreta ou continua. Se ela for qualitativa, ela poderá ser ordinal ou

Figura 6.2: Escala de mensuração, precisão e categorização das variáveis quantitativa e qualitativa

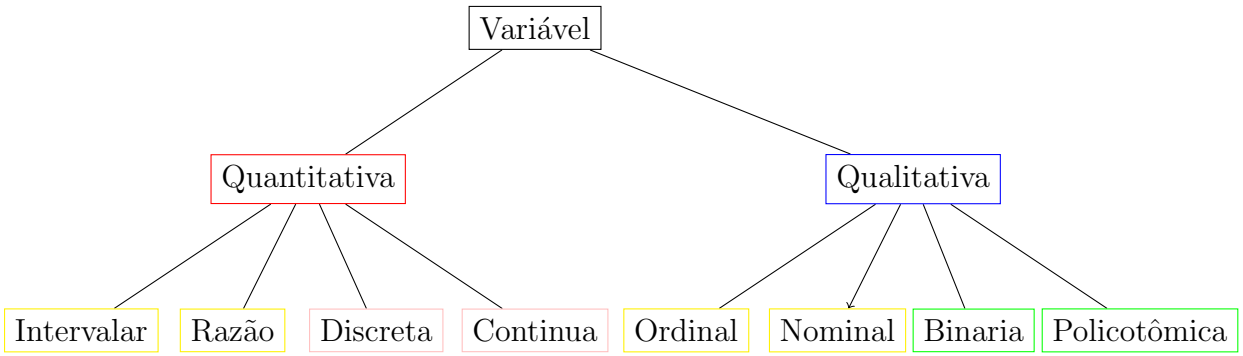


Tabela 6.3: Classificação das variáveis, agora com as escalas.

| **Nome da variável** | **Classificação** |
|---|---|
| Nome | qualitativa, nominal e policotômica |
| Endereço | qualitativa, nominal e policotômica |
| Telefone | qualitativa, nominal e policotômica |
| Estado Civil | qualitativa, nominal e binaria |
| Altura | quantitativa, intervalar e continua |

nominal, também binaria ou policotônica. Em outras palavras ela só terá duas dessas classificações de escala, se estivessemos lidando com cores, bastaria escolher uma opção da cor vermelha e uma da cor azul, se for quantitativa, e uma da cor vermelha e outra da cor verde se qualitativa. A Tabela 6.3 está atualizada com essa classificação das escalas.

### 6.0.1 Quantitativo

Já sabemos que uma variável quantitativa é expressa geralmente (mas nem sempre) como um número, e pode ser quanto a sua escala de mensuração, intervalar ou razão. Além disso, essa variável ainda pode ser classificada em relação a sua escala de precisão, contínuo ou discreto.

Sabendo que a variável é quantitativa podemos utilizar as medidas estatísticas de posição ou localização, tais como, média, mediana, moda, quartis, decis e percentis. Também podemos utilizar as medidas de dispersão, tais como, amplitude, desvio padrão, erro-padrão e coeficiente de variação. Além disso, para uma representação visual dos gráficos podemos utilizar os gráficos do tipo linha, dispersão, histograma, ramo-e-folhas e boxplot, por fim as medidas de forma como assimetria e curtose também podem ser utilizadas[**?**].

#### Escala de precisão: Quantitativa contínua

A variável quantitativa pode possuir um intervalo de domínio real ou inteiro. Lembrando que o conjunto de número reais engloba os números inteiros. Geralmente essa variável é o resultado de uma medida, por exemplo, a altura dos estudantes é um dado do tipo quantitativo contínuo.

As linguagens que suportam fins estatísticos (Python, R, stata) ou as ferramentas estatísticas visuais (tais como: SPSS, JASP ou Jamovi), não são obrigadas a ter um equivalente dessa tipologia teórica. Isso depende das escolhas dos desenvolvedores e do foco da ferramenta.

Python é uma Linguagem de uso diverso, não somente estatístico. Além disso, em programação (Python) o termo variável é utilizado com um significado diferente,

porém nesse capítulo, quando falamos de variável estamos nos referindo a coluna da tabela, a característica medida naquela amostra.

Vamos criar em Python uma tabela (chamado de DataFrame) que tem uma única coluna (variável) do tipo quantitativa continua. Classificamos esta coluna teoricamente em relação a escala de precisão (da contagem) como continua. Veja no código a seguir.

Listing 6.1: Código que cria e exibe uma tabela em Python com uma coluna. A classificação teórica dessa variável é quantitativa continua.

```
>>> import pandas as pd
>>> dados = [1.80, 1.70, 1.50, 1.79]
>>> df = pd.DataFrame(data=dados, columns=['altura'])
>>> df
   altura
0    1.80
1    1.70
2    1.50
3    1.79
>>>
```

O Jamovi (https://www.jamovi.org/) é um software estatístico utilizado para realizar análises. No Jamovi é possível criar visualmente o conjunto de dados. Durante esse processo você informa para cada variável o **tipo de dado** e o **tipo de medida** (Figura 6.3). As opções do tipo de dado no Jamovi são: inteiro, decimal e texto. Já para o tipo de medida: nominal, ordinal, continua e ID.

Na tipologia teórica, uma variável do tipo quantitativa só pode ter uma escala de mensuração entre intervalar e razão, não podendo ser nominal ou ordinal, porque nominal e ordinal estão relacionados a variável qualitativa. Portanto, no Jamovi quando vc escolhe o tipo de dado decimal, automaticamente vc não consegue escolher o tipo de mensuração nominal nem ordinal. Da mesma forma quando vc escolhe o tipo texto (que equivale ao qualitativo) vc não consegue escolher intervalar ou razão, porque são as opções relacionadas a variável quantitativa.

O Jamovi, não tem a distinção entre intervalar e razão, ambas são possíveis no tipo de medida contínua.

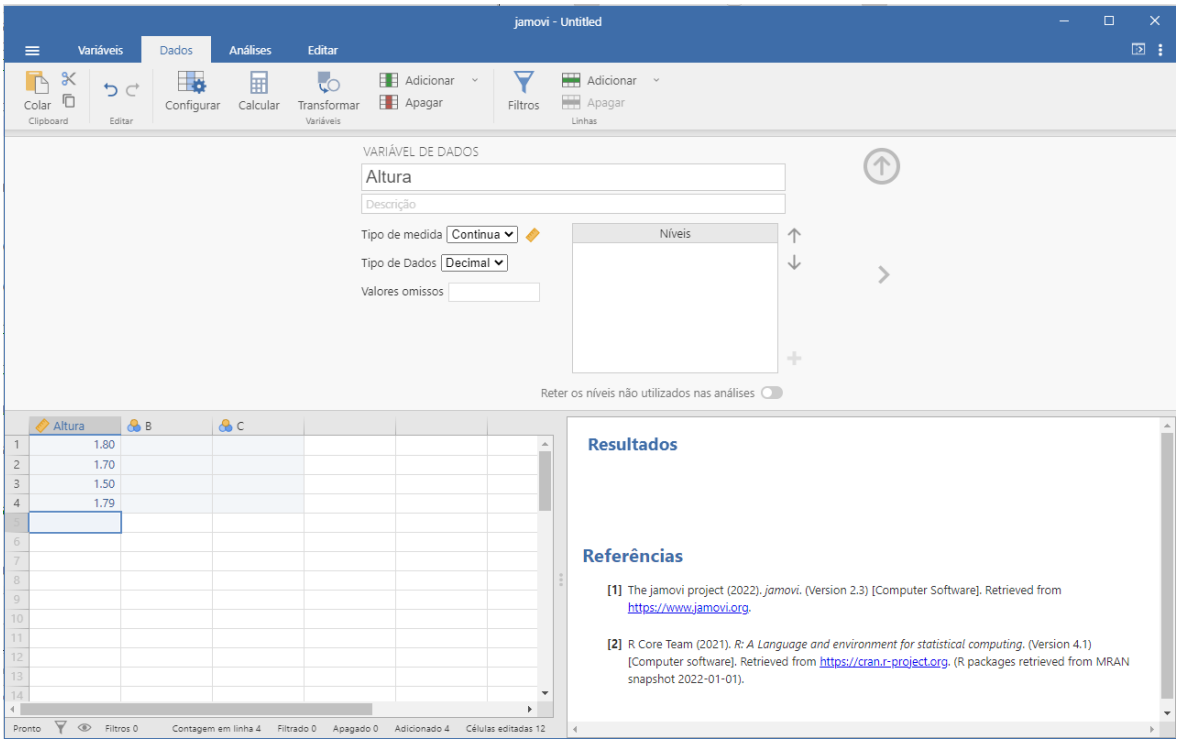



Figura 6.3: Jamovi

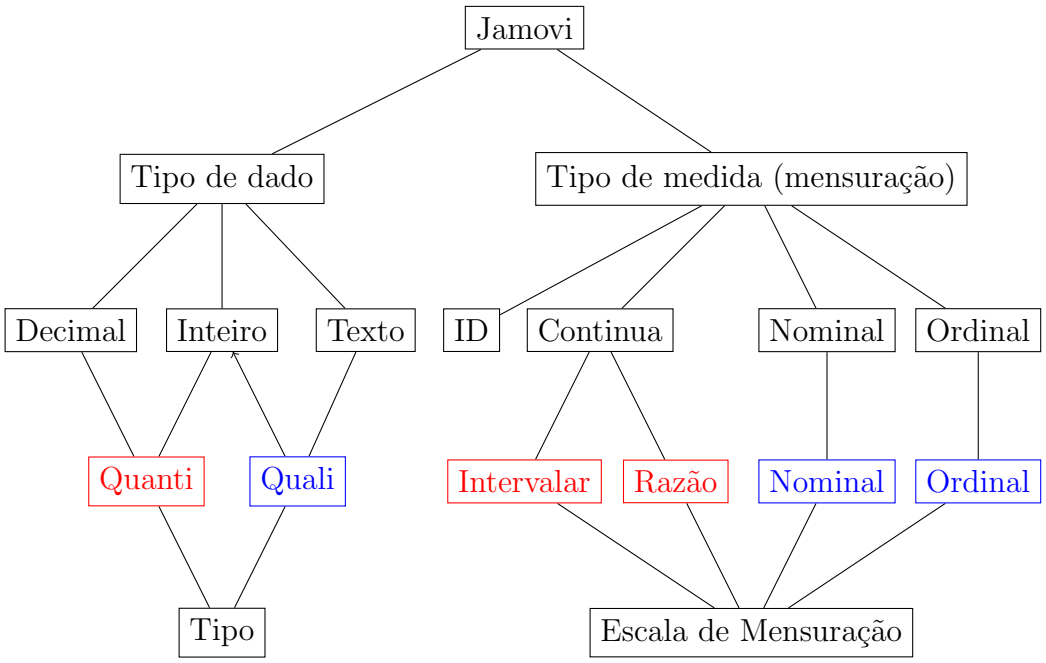



Figura 6.4: Jamovi Tipo de dado e tipo de medida (ou mensuração)

# Capítulo 7

# Árvore de decisão

# Capítulo 8

# Regressão Linear

# Referências Bibliográficas