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

Giseldo Neo

versao alpha 0.1

4 de julho de 2024

Sumário

Capítulo 1

Introdução

1.1 Inteligência artificial

O termo "inteligência" tem várias 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 significado destas palavras.

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 a do filme "2001 uma odisseia no espaço", clássico de Kubric, ou como a do mais recente filme "Ela", com o ator Joaquim Phenix, espero que com sorte, por um garoto interessado em aprender.

Homo-Sapiens vem do latim e significa homem sábio [Wikipedia,]. A importância da sapiência (que é um sinônimo de inteligência) é tamanha que define a nossa própria espécie. Porém, dentro do nosso contexto consideramos que um gato e um cachorro são seres dotados de muita inteligência, uma abelha, então nem se fala, praticamente uma cientista. Portanto, seremos mais contidos e reservados quanto ao termo inteligência.

No entanto, várias questões relacionadas a inteligência também guiam inúmeras pesquisas cientificas, por exemplo: como funciona nossa inteligência? Nossa percepção do ambiente é próxima da realidade objetiva? Ainda não é dessa inteligência que estamos falando. Essas perguntas estão mais próxima da neurociência e da filosofia.

Além disso, "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. Por exemplo, 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 ela já é utilizada em vários domínios do conhecimento e com entusiasmo [wikipedia,].

A nossa "inteligência artificial" está mais relacionada com a capacidade de realizar coisas que seres inteligentes (um gato, um bebê, ou um cientista) realizam, como por exemplo puxar a mão (ou pata) instantaneamente ao tocar em uma superfície quente (inteligência reativa), ou realizar uma prova de anatomia (inteligência cognitiva). Se conseguimos que programas realizem ações realizadas por entidades dotadas de inteligencia, e realizamos isso de forma computacional, estamos próximos da nossa definição desejada de "inteligência artificial".

Russel e Norvig (2020) em um dos livros mais lidos em todas as universidades do mundo tem uma boa definição sobre esse tema: "O campo da inteligência artificial [...] tenta não apenas compreender, mas também construir entidades inteligentes" (tradução nossa) [Norvig and Intelligence, 2002]. Logo, dado o ambicioso desejo de compreender a inteligência, temos o também audacioso objetivo de construir agentes inteligentes dotados dessa inteligência.

A origem do termo "inteligência artificial", nesse contexto, é atribuída a John McCarthy, professor de Matemática da Universidade Dartmouth College[blipblog,], ele organizou uma conferência com duração de oito semanas com mais alguns 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 já demonstrava um bom ensaio sobre as possibilidades de uma máquina inteligente [Turing, 1950].

Foi na década de 1970 que o uso da IA começou a ser mais difundido. Esses primeiros sistemas de IA foram chamados de Sistemas Especialistas (veja um exemplo na Figura 1.1)e dependiam muito dos Homo-Sapiens para transformar o conhecimento tácito (baseado em sua experiência) em explicito (formalizado, documentado), que era então codificado na forma de um software contendo regras em lógica formal. O processo de aquisição do conhecimento por esses especialistas acabou sendo um grande obstáculo na adoção em massa dessa abordagem.

Nestas últimas décadas houve um crescimento exponencial das tecnologias que estão ao redor da IA, tais como, o aumento capacidade de processamento e armazenamento dos computadores e a geração de grandes volumes de dados, além dos avanços científicos e tecnológicos em outras áreas, tais como chips supercondutores e eficiência energética.

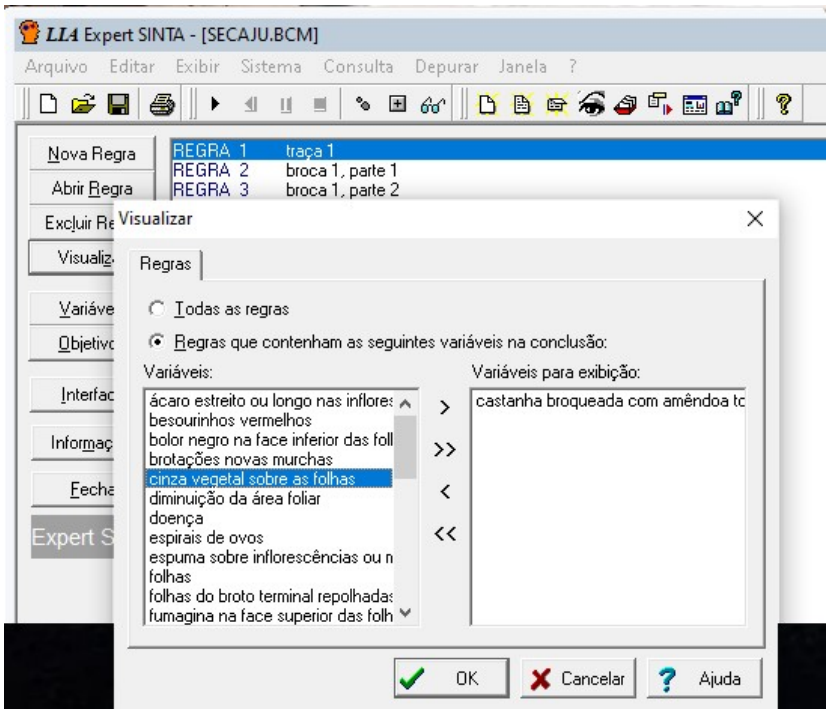


Figura 1.1: ExpertSinta. Uma interface de um Sistema Especialista

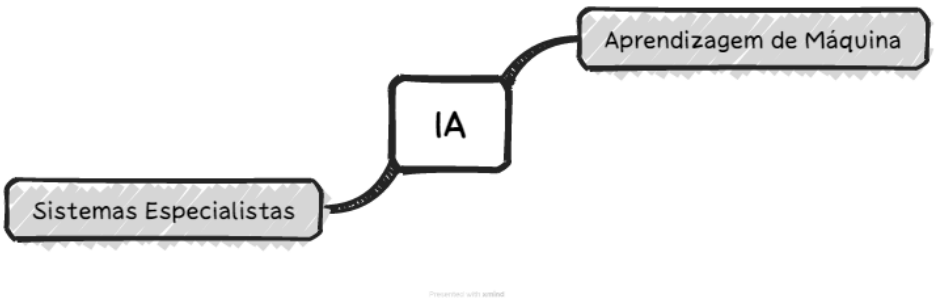


Figura 1.2: AM é uma parte da IA

1.2 Aprendizado de Máquina

O Aprendizado de Máquina (AM) é uma subárea da IA (veja Figura 1.2). que foi 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 **heurísticas** (regra do dedão) que buscam por modelos capazes de representar o conhecimento existente nos dados.

Na Figura 1.3, 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 recomen-

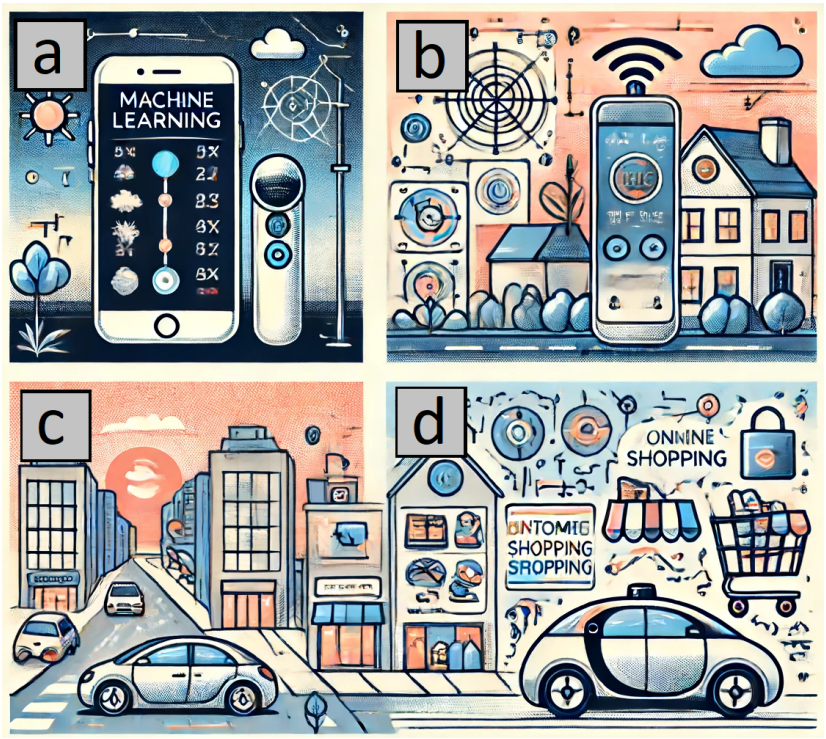


Figura 1.3: Exemplos AM

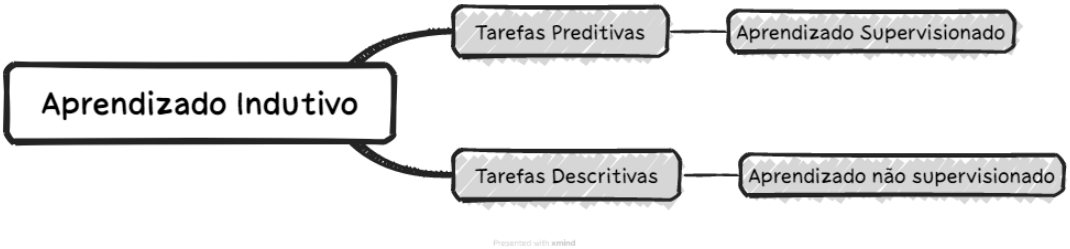


Figura 1.4: Classificação AM

dando produtos a um usuário com base em suas compras anteriores. Essa figura foi criada inclusive com inteligência artificial.

1.2.1 Classificação

As tarefas de aprendizado de máquina podem ser divididas entre tarefas preditivas, que visam inferir o atributo alvo de uma nova entrada a partir da exposição prévia aos dados rotulados durante o treinamento do modelo, e descritivas, que buscam extrair padrões dos atributos preditivos. Por conseguinte, uma vez que pertencem a este paradigma, as tarefas de aprendizado descritivas não possuem atributos alvo. Noutras palavras, tarefas preditivas analisarão os atributos preditivos, comparando-os com os atributos alvo (rótulos), ao passo que tarefas descritivas utilizaram os atributos preditivos entre si para buscar por padrões e correlações.

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 (rótulo) é discreto (enumerável ou finito), ou de regressão, em que o atributo alvo (rótulo) é contínuo (não enumerável ou infinito). Já as descritivas distinguem-se entre agrupamento, que busca por similaridades, associação, que busca por padrões frequentes, e sumarização, que resulta em um resumo do conjunto de dados.

1.2.2 Exemplo de AM

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.

```
1    from sklearn import svm
2    from sklearn.datasets import iris
3    iris = load_iris()
4    X = iris.data
5    y = iris.target
6    model = svm.SVC()
7    model.fit(X, y)
8    model.predict([[2., 2., 2., 2.]])
```

Listing 1.1: Exemplo de código que usa AM

Capítulo 2

Estatística Básica

Um conjunto de dados geralmente é uma estrutura tabular com linhas e colunas, o nome da coluna é o identificador do dado (também chamado de variável) disposto naquela coluna. Cada linha da coluna é chamada de observação (ou registro), e representa uma instância daquele elemento. Por exemplo, uma tabela com dados do cliente é apresentado na Tabela 2.1, a primeira linha (em negrito) é o nome da coluna, são elas: nome, endereço e telefone; cada linha abaixo do nome da coluna representa um cliente. Portanto a linha 1, teria o nome das 3 colunas; já a linha 2, teria um determinado cliente, e a linha 3, outro cliente.

Em resumo, a Tabela 2.1 apresenta 2 observações (ou registros) de clientes. A primeira coluna é uma descrição das informações que existirão naquela coluna, por exemplo "nome" significa que provavelmente todos os dados dessa coluna são referentes ao nome de determinado cliente, já cada linha abaixo da primeira linha são os dados de um cliente em específico.

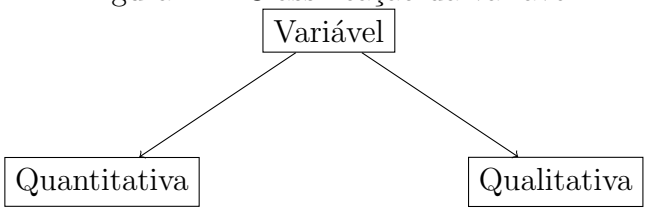
Quando vamos realizar um resumo estatístico (tal como média, mediana ou moda), ou algum gráfico, ou alguma inferência estatística, ou rodar um modelo preditivo, temos que conhecer o tipo teórico, daquele dado (ou variável, ou coluna) para podermos selecionar qual a técnica mais adequada que devemos aplicar. Por exemplo, em nenhuma das colunas da tabela cliente devemos calcular a média. Qual é a média, do nome? ou a média do endereço, ou a média do telefone? Não faz sentido.

Utilizando uma classificação para esse tipo de variável, 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

Tabela 2.1: Tabela cliente.

Nome	**Endereço**	**Telefone**
Giseldo Neo	Rua das alamedas, n 27, Corumbá MS	222 66666
Alex Neo	Avenida Fernandes 325, Macieó, AL	333 6589

Figura 2.1: Classificação da variável



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.

2.1 Variável

Uma variável, em conceitos estatísticos, é uma característica do que foi observado naquele universo (amostra ou população), que foi medida, contada, ou categorizada [Fávero and Belfiore, 2017]. No nosso exemplo do cliente (Tabela 2.1) as variáveis são: nome, endereço e telefone. Elas foram registradas na tabela após um desses processos de mensuração, contagem ou categorização. Por exemplo, medimos a altura de uma pessoa e registramos isso em uma tabela, a altura é então chamada genericamente de variável.

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, pois existem determinadas técnicas para determinados fins. Por isso, vamos entender estas classificações teóricas.

Neste contexto, uma variável pode ser **quantitativa** ou **qualitativa** (Figura 2.1). Além disso, uma variável quantitativa também é chamada de métrica, e a variável qualitativa de não métrica ou categórica [Fávero and Belfiore, 2017].

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 2.2).

Tabela 2.2: 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

Tabela 2.3: Tipo das variáveis da tabela cliente.

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

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 2.3 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 **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 2.3, 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**.

Figura 2.2: Escala da variável

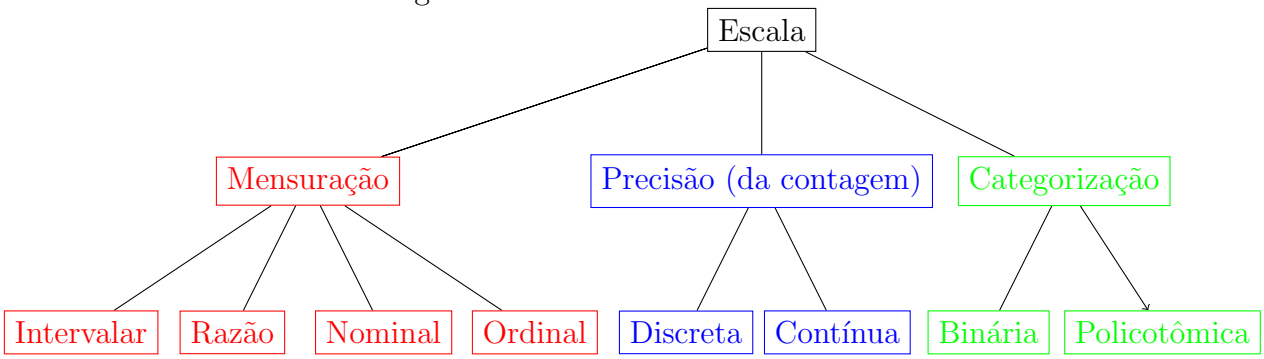


Figura 2.3: Escala de mensuração, precisão e categorização das variáveis quantitativa e qualitativa

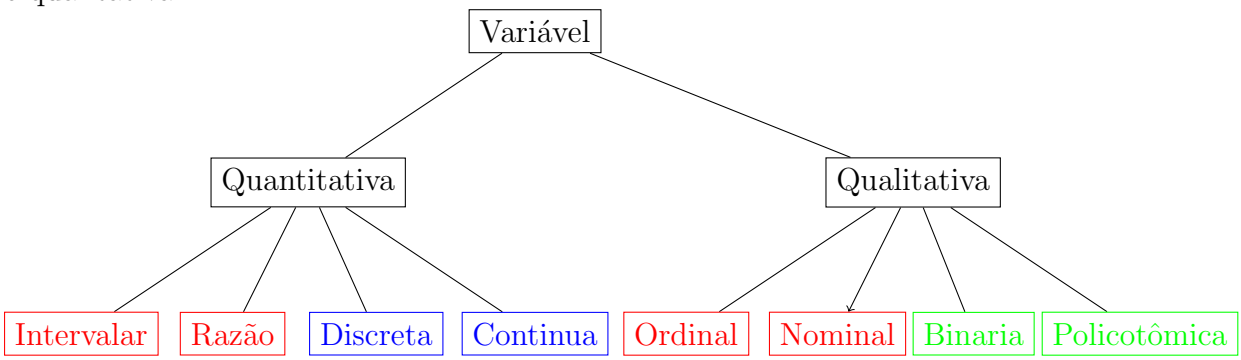


Tabela 2.4: Classificação das variáveis, agora com as escalas.

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

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 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 2.4 está atualizada com essa classificação.

2.1.1 Quantitativa

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[Fávero and Belfiore, 2017].

Escala de precisão: Quantitativa contínua

A variável quantitativa, com escala de precisão contínua, é quando ela possui um intervalo de domínio dos números reais. 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.

Vamos criar em Python uma tabela (chamado de DataFrame) que tem uma única coluna (variável) do tipo quantitativa continua. Porém, as linguagens (Python, R) ou as ferramentas estatísticas visuais (tais como: SPSS, JASP ou Jamovi), não são obrigadas a ter um equivalente dessa tipologia teórica, isso depende do contexto do dado, da interpretação humana do que aquela variável representa. No código a seguir criamos uma tabela com uma única coluna. Classificamos então ela em relação a escala de precisão (da contagem) como continua. Em programação o termo variável é utilizado para um fim diferente, porém nesse capítulo, quando falamos de variável estamos nos referindo a coluna da tabela.

```
>>> 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
>>>
```

Listing 2.1: Código que cria e exibe uma tabela em python com uma variável quantitativa continua.

Referências Bibliográficas

[blipblog,] blipblog. https://www.take.net/blog/tecnologia/inteligenciaartificial/.

[Fávero and Belfiore, 2017] Fávero, L. P. and Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier Brasil.

[Norvig and Intelligence, 2002] Norvig, P. R. and Intelligence, S. A. (2002). A modern approach. *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, 90:33–48.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and intelligence. *Mind*, 49(8):433–460.

[Wikipedia,] Wikipedia. https://pt.wikipedia.org/wiki/Humano.

[wikipedia,] wikipedia. https://pt.wikipedia.org/wiki/TransformadadeFourier.