Analytics e Inteligência Artificial Data Science

Tema da aula
Framework Geral de Modelagem

FIA BUSINESS SCHOOL

lab.data

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseadas em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

BUSINESS SCHOOL lab.data

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil.

Os diretores foram professores de grandes especialistas do mercado.

+10 anos de atuação.

+9.000 alunos formados.

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria;
- Larga experiência de mercado na resolução de *cases;*
- Participação em congressos nacionais e internacionais;
- Professor assistente que acompanha o aluno durante todo o curso.

Estrutura

- 100% das aulas realizadas em laboratórios;
- Computadores para uso individual durante as aulas;
- 5 laboratórios de alta qualidade (investimento +R$2MM);
- 2 unidades próximas à estação de metrô (com estacionamento).

FIA BUSINESS SCHOOL lab.data

PROFA. DRA. ALESSANDRA DE ÁVILA MONTINI

Diretora do LABDATA-FIA, apaixonada por dados e pela arte de lecionar. Tem muito orgulho de ter criado na FIA cinco laboratórios para as aulas de Big Data e Inteligência Artificial. Possui mais de 20 anos de trajetória nas áreas de Data Mining, Big Data, Inteligência Artificial e Analytics. Cientista de dados com carreira realizada na Universidade de São Paulo. Graduada e mestra em Estatística Aplicada pelo IME-USP e doutora pela FEA-USP. Com muita dedicação chegou ao cargo de professora e pesquisadora na FEA-USP, ganhou mais de 30 prêmios de excelência acadêmica pela FEA-USP e mais de 30 prêmios de excelência acadêmica como professora dos cursos de MBA da FIA. Orienta alunos de mestrado e de doutorado na FEA-USP. Parecerista da FAPESP e colunista de grandes portais de tecnologia.

Currículo do Professor

PROF. ÂNGELO CHIODE, MSc

Bacharel, mestre e candidato ao PhD em Estatística (IME-USP), atua como professor de Estatística Aplicada para turmas de especialização, pós-graduação e MBA na FIA. Trabalha como consultor nas áreas de Analytics e Ciência de Dados há 13 anos, apoiando empresas na resolução de desafios de negócio nos contextos de finanças, adquirência, seguros, varejo, tecnologia, aviação, telecomunicações, entretenimento e saúde. Nos últimos 5 anos, tem atuado na gestão corporativa de times de Analytics, conduzindo projetos que envolviam análise estatística, modelagem preditiva e *machine learning*. É especializado em técnicas de visualização de dados e design da informação (Harvard) e foi indicado ao prêmio de Profissional do Ano na categoria Business Intelligence, em 2019, pela Associação Brasileira de Agentes Digitais (ABRADi).

Conteúdo Programático

DISCIPLINAS

- IA E TRANSFORMAÇÃO DIGITAL
- ANALYTICS
- **INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING**
- INTELIGÊNCIA ARTIFICIAL: DEEP LEARNING
- EMPREENDEDORISMO E INOVAÇÃO
- COMPORTAMENTO HUMANO E SOFT SKILLS

TEMAS: ANALYTICS E MACHINE LEARNING

- ANÁLISE EXPLORATÓRIA DE DADOS
- INFERÊNCIA ESTATÍSTICA
- TÉCNICAS DE PROJEÇÃO
- TÉCNICAS DE CLASSIFICAÇÃO
- TÓPICOS DE MODELAGEM
- TÉCNICAS DE SEGMENTAÇÃO
- TÓPICOS DE ANALYTICS
- **MANIPULAÇÃO DE BASE DE DADOS**
- AUTO ML

TEMAS: DEEP LEARNING

- REDES DENSAS
- REDES CONVOLUCIONAIS
- REDES RECORRENTES
- MODELOS GENERATIVOS

FERRAMENTAS

- LINGUAGEM R
- **LINGUAGEM PYTHON**
- DATABRICKS

FIA BUSINESS SCHOOL lab.data

Conteúdo da Aula

1. Introdução e Objetivo
2. Passo a Passo
3. Criação de *Features*
4. Redução de Dimensionalidade
5. Análise de Componentes Principais

Referências Bibliográficas

1. Introdução e Objetivo

Como Construir um Modelo do Zero?

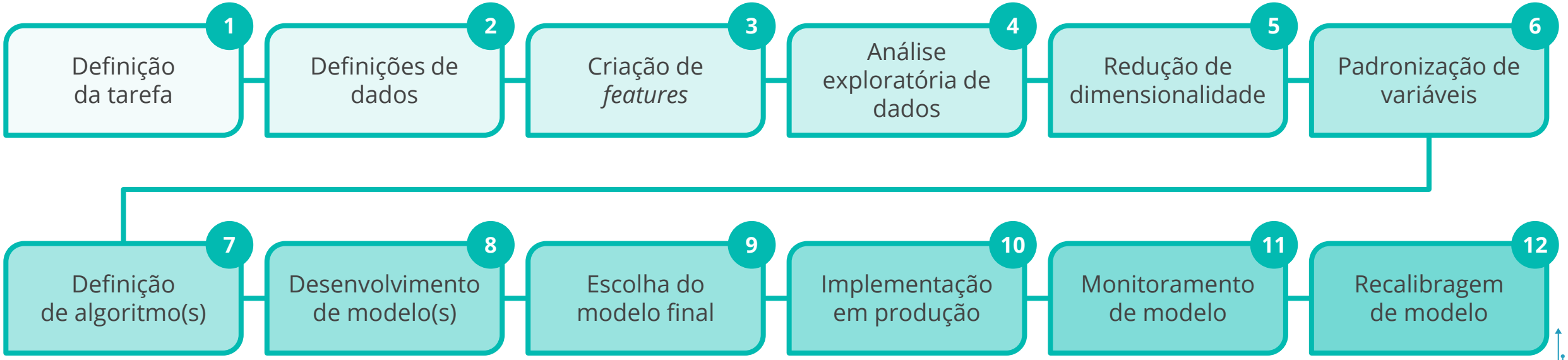
Para construir soluções de modelagem estatística ou *machine learning* que sejam eficazes, é importante seguir algumas boas práticas e tomar certos cuidados com relação aos dados.

Nesta aula, abordaremos o **passo a passo** recomendado para construção de modelos estatísticos e de *machine learning*, recapitulando alguns tópicos já discutidos e aprofundando-nos em algumas das etapas relacionadas à manipulação da base de dados.

2. Passo a Passo

Passo a Passo da Modelagem

O seguinte fluxo abrange os passos para construção de um modelo estatístico ou de *machine learning*. Discutiremos cada um deles a seguir.



FIA BUSINESS SCHOOL lab.data

Passo 1: Definição da Tarefa

1	**Definição da tarefa**
2	Definições de dados
3	Criação de *features*
4	Análise exploratória de dados
5	Padronização de variáveis
6	Redução de dimensionalidade
7	Definição de algoritmo(s)
8	Desenvolvimento de modelo(s)
9	Escolha do modelo final
10	Implementação em produção
11	Monitoramento de modelo
12	Recalibragem de modelo

Uma vez definido o objetivo de negócio, precisamos tangibilizar tecnicamente a natureza da **tarefa de modelagem** que será conduzida.

A tarefa será **supervisionada** caso haja variável resposta/*target* definida; ou **não supervisionada**, caso contrário.

As tarefas supervisionadas mais comuns são:

- **Classificação**
- **Projeção não temporal**
- **Projeção temporal**

As tarefas não supervisionadas mais comuns são:

- **Agrupamento**
- **Redução de dimensionalidade**
  (pode ser, também, uma ferramenta intermediária, em vez da intenção final)

Passo 1: Definição da Tarefa

1	**Definição da tarefa**
2	Definições de dados
3	Criação de *features*
4	Análise exploratória de dados
5	Padronização de variáveis
6	Redução de dimensionalidade
7	Definição de algoritmo(s)
8	Desenvolvimento de modelo(s)
9	Escolha do modelo final
10	Implementação em produção
11	Monitoramento de modelo
12	Recalibragem de modelo

Uma vez definido o objetivo de negócio, precisamos tangibilizar tecnicamente a natureza da **tarefa de modelagem** que será conduzida.

A tarefa será **supervisionada** caso haja variável resposta/*target* definida; ou **não supervisionada**, caso contrário.

As tarefas supervisionadas mais comuns são:

TÉCNICAS JÁ ESTUDADAS

- **Classificação** → Regressão logística
- **Projeção não temporal** → Regressão linear
- **Projeção temporal** → Modelos (S)ARIMA, regressão linear temporal

As tarefas não supervisionadas mais comuns são:

- **Agrupamento** → Clusterização hierárquica, k-médias, k-medoides
- **Redução de dimensionalidade**

(pode ser, também, uma ferramenta intermediária, em vez da intenção final)

FIA BUSINESS SCHOOL lab.data

Passo 2: Definições de Dados

Nesta etapa, define-se os **critérios gerais** a respeito dos dados que serão utilizados para a modelagem, levando em conta o objetivo de negócio.

As principais definições realizadas nessa etapa são:

- **Visão da base**: Qual é a granularidade que se deseja modelar?

  Exemplos: cliente; transação; produto; matriz; filial; região.

- **Filtros**: Quais critérios definem as observações que devem ser incluídas ou excluídas da modelagem?

  Exemplos: incluir apenas clientes que compraram nos últimos 12 meses; incluir apenas transações com status igual a "concluída"; excluir produtos com venda descontinuada.

- **Variável resposta**: Qual a variável resposta *(target)* do modelo, caso ele seja supervisionado?

  Exemplos: índice de satisfação de um cliente; faturamento mensal da empresa; indicação de se um prospect adquire ou não um produto oferecido.

Passo 2: Definições de Dados

Nesta etapa, define-se os **critérios gerais** a respeito dos dados que serão utilizados para a modelagem, levando em conta o objetivo de negócio.

As principais definições realizadas nessa etapa são:

- **Períodos**: Quais safras vão ser consideradas para construção e validação *out-of-time* do modelo? E quais os períodos de histórico e previsão?

  Exemplo (supervisionado):
  safras: construção em jan/24 (M0) e validação em abr/24 (M0)
  histórico dos últimos 12 meses (M-0 a M-11)
  previsão nos próximos 2 meses (M+1 a M+2)

  Exemplo (não supervisionado):
  safra: jan/24 (M0)
  histórico dos últimos 6 meses (M-0 a M-5)

Não há **período de previsão** nem **validação** em tarefas não supervisionadas.

FIA BUSINESS SCHOOL lab.data

Passo 3: Criação de *Features*

Após as definições principais, é necessário definir quais são as **demais variáveis** *(features)* que devem ser consideradas no modelo.

Exemplos:

- *quantidade de reclamações nos últimos 12 meses;*
- *quantidade mensal de vendas da empresa nos últimos 36 meses;*
- *renda declarada do cliente.*

Nem sempre as variáveis ideais para um modelo estão disponíveis para utilização de forma imediata. No contexto corporativo, é comum que as informações estejam pulverizadas em diferentes origens de dados e em diferentes visões, de forma que precisamos **definir** e **calcular as *features*** que melhor representem o problema de interesse.

Em modelos supervisionados, estamos interessados, também, que as *features* sejam calculadas de forma a maximizar a **explicabilidade** da variável resposta.

Vamos discutir este tópico com mais detalhe a seguir, na **Seção 3**.

Passo 4: Análise Exploratória de Dados

A primeira etapa a ser realizada com os dados em mãos é sempre a **análise exploratória**, a fim de gerar conhecimento inicial a respeito do contexto e identificar/sanar possíveis inconsistências.

Os principais pontos de atenção nesta etapa são:

- **Unicidade**: Verificar se não há repetições indevidas de registros na base. Se houver, elas devem ser investigadas e excluídas.
- **Preenchimento**: Contabilizar a quantidade de valores ausentes *(missings)* em cada variável. Se possível, adotar algum destes tratamentos:
  - **Sanar** os valores ausentes, recuperando a informação original.
  - **Substituir** os valores ausentes por algum valor pertinente (ex.: zero)*.
  - **Categorizar** as variáveis em faixas de valores, e uma categoria "*missing*".
  - **Excluir** as observações por completo, apenas se os valores ausentes ocorrerem em grande quantidade de variáveis e/ou em variáveis fundamentais para a modelagem (ex.: variável resposta, filtros).

** Obs.: A **imputação** de missing values em variáveis quantitativas a partir da média, mediana, moda etc., apesar de popular, não é efetiva em muitas situações. Ela distorce a distribuição da variável e não costuma refletir o real processo subjacente por meio do qual os valores ausentes surgiram.*

FIA BUSINESS SCHOOL lab.data

Passo 4: Análise Exploratória de Dados

A primeira etapa a ser realizada com os dados em mãos é sempre a **análise exploratória**, a fim de gerar conhecimento inicial a respeito do contexto e identificar/sanar possíveis inconsistências.

Os principais pontos de atenção nesta etapa são:

- ***Outliers***: Examinar se há valores atípicos *(outliers)* nas variáveis quantitativas. Se necessário, adotar algum destes tratamentos:
  - **Substituir** os valores atípicos das caudas superior e inferior por algum percentil extremo; em geral, usa-se os percentis 99/1, ou 99,9/0,01.
  - **Categorizar** as variáveis em faixas de valores.
  - **Excluir** observações pontualmente, caso os *outliers* estejam presentes em variáveis fundamentais para a modelagem (ex.: variável resposta, filtros).

Importante: A presença de *missing values* ou *outliers* não exerce influência em algoritmos baseados em **árvores**, que veremos mais adiante no curso.

FIA BUSINESS SCHOOL lab.data

Passo 5: Padronização de Variáveis

Em alguns tipos de algoritmos, faz-se necessária a **padronização** de **variáveis quantitativas**, a fim de que a relevância delas no modelo não seja influenciada por questões de escala.

A padronização é importante quando utilizamos:

- Algoritmos que envolvem estimação de parâmetros, tais como a regressão (linear ou logística), apenas quando realizada via **gradiente descendente**.
- Algoritmos que envolvem **cálculo de distâncias**, tais como algoritmos de segmentação (hierárquico, *k*-médias) e de redução de dimensionalidade (ACP).
- Demais algoritmos de *machine learning* sensíveis à **escala dos dados**, como *support vector machines* (SVM) e redes neurais.

Relembrando, as formas usuais de padronização de variáveis quantitativas são:

Método *z-score*

$$x_i\ padronizado = \frac{x_i - \text{média}\ (X)}{\text{d. p.}\ (X)}$$

Método *range*

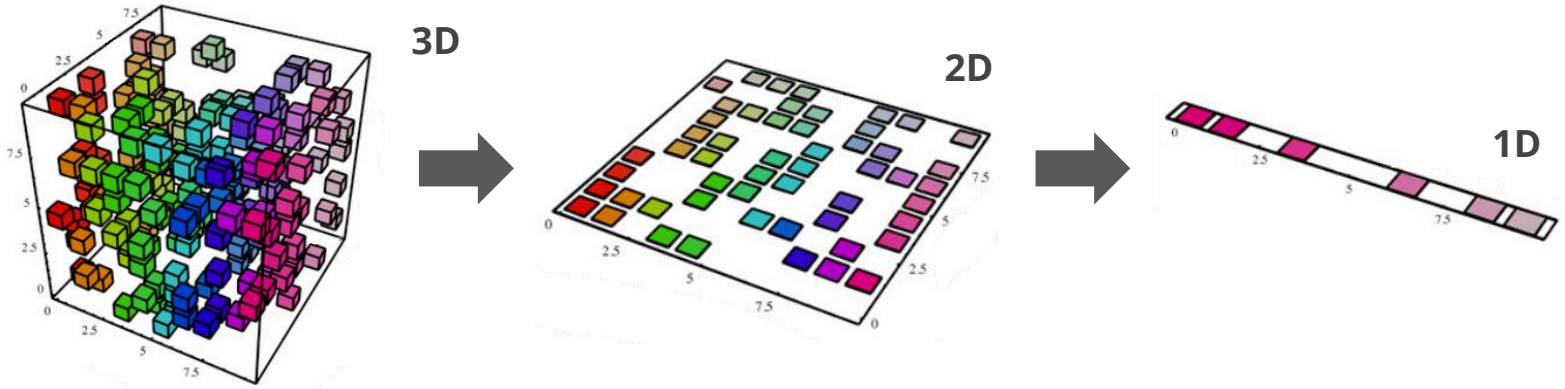
$$x_i\ padronizado = \frac{x_i - \text{mín}\ (X)}{\text{máx}\ (X) - \text{mín}\ (X)}$$

FIA BUSINESS SCHOOL lab.data

Passo 6: Redução de Dimensionalidade

Em algumas situações, a quantidade de variáveis explicativas disponíveis pode ser **grande**, o que torna o trabalho de modelagem mais oneroso do ponto de vista analítico e computacional.

Nesses casos, convém utilizar uma técnica preliminar de **redução de dimensionalidade**, que visa representar as informações contidas em uma base de dados de forma mais parcimoniosa, por meio de um conjunto menor de variáveis.



Vamos discutir este tópico com mais detalhe a seguir, nas **Seções 4 e 5**.

Passo 7: Definição de Algoritmo(s)

Finalizados os passos de preparação de dados, temos que definir **qual(is) algoritmo(s)** será(ão) considerado(s) na modelagem, em vista da natureza da tarefa definida no passo 1.

Existem algumas centenas de algoritmos de *machine learning* disponíveis em linguagens de programação como Python e R. A cada ano, novos algoritmos são desenvolvidos. Nas próximas aulas, estudaremos alguns dos mais consolidados e utilizados atualmente.

Sempre que possível, recomenda-se testar diferentes tipos de algoritmos para um mesmo problema, a fim de comparar os resultados e adotar o modelo mais adequado.

Passo 8: Desenvolvimento de Modelo(s)

Nesta etapa, desenvolve-se um ou mais modelos candidatos para o objetivo de interesse, com base no seguinte racional.

- Para modelos **supervisionados**:
  - Caso a quantidade de observações (n) seja pequena (em geral, $n \leq 100$), treinar o modelo com **todas** as observações da amostra. É inviável realizar validação quando há poucas observações.
  - Caso contrário, convém separar conjuntos distintos para **treino** e para **validação** do modelo, para garantir que o modelo tenha boa capacidade de extrapolação. Isto pode ser feito por meio de diferentes racionais:
    - ✓ **Validação simples** *(holdout validation)* ---------- JÁ ESTUDADO
    - ✓ **Validação cruzada** *(k-fold cross validation)*
    - ✓ **Validação cruzada aninhada** *(nested k-fold cross validation)*
  - Em modelos de classificação, tal separação pode ser feita de forma **estratificada**, assegurando que as proporções de observações de cada categoria sejam preservadas em cada parte de treino/validação.
  - Vamos discutir este tópico com detalhe mais adiante em nosso curso.

FIA BUSINESS SCHOOL lab.data

Passo 8: Desenvolvimento de Modelo(s)

Nesta etapa, desenvolve-se um ou mais modelos candidatos para o objetivo de interesse, com base no seguinte racional.

- Para modelos **supervisionados**:
  - Adicionalmente, pode haver uma segunda etapa de validação, que é a ***out-of-time***. Aplica-se o modelo em dados de outra(s) safra(s) temporais, calculados segundo os mesmos critérios da base de desenvolvimento, a fim de avaliar a **estabilidade** do modelo ao longo do tempo.

Passo 8: Desenvolvimento de Modelo(s)

Nesta etapa, desenvolve-se um ou mais modelos candidatos para o objetivo de interesse, com base no seguinte racional.

- Para modelos **não supervisionados**:
  - ➢ Utiliza-se **todas** as observações da amostra. Não é usual validar algoritmos em outras bases de dados quando não há variável resposta.

Passo 8: Desenvolvimento de Modelo(s)

Para treino de um algoritmo de *machine learning*, também é necessário definir algumas características gerais que regem a forma como ele será desenvolvido. Essas características são conhecidas como **hiperparâmetros**.

Exemplo: Pensando no algoritmo como uma receita culinária, na qual os ingredientes são as *features*, alguns hiperparâmetros possíveis seriam:

- Temperatura do forno
- Velocidade da batedeira
- Tempo de cozimento

Note que, a depender dos valores adotados para os hiperparâmetros, o resultado final do algoritmo pode ser diferente, mesmo com as mesmas *features* de partida.

Durante o processo de desenvolvimento, é comum testar diferentes **combinações de hiperparâmetros**, a fim de identificar quais valores proporcionam um modelo mais adequado. Ex.: velocidade da batedeira média + forno a 200°? ou velocidade da batedeira alta + forno a 180º?

Falaremos mais a respeito dos hiperparâmetros pertinentes durante as aulas específicas de cada tipo de algoritmo.

Passo 9: Escolha do Modelo Final

Ao final do desenvolvimento, analisa-se os resultados obtidos e toma-se a decisão de qual será o modelo final.

- Para modelos **supervisionados**:
  - Sob a ótica do aprendizado estatístico, prioriza-se a **interpretação** para escolha do modelo final: avaliação das variáveis mais fortes do modelo, valores dos seus coeficientes (em regressão) ou quebras utilizadas (em árvores) etc., adesão às premissas teóricas, bem como o **desempenho**.
  - Sob a ótica do aprendizado de máquina, prioriza-se fundamentalmente o **desempenho** e a **performance computacional**.

Passo 9: Escolha do Modelo Final

Ao final do desenvolvimento, analisa-se os resultados obtidos e toma-se a decisão de qual será o modelo final.

- Para modelos **não supervisionados**:
  - A **interpretação** é um dos fatores fundamentais, tanto no aprendizado estatístico quanto de máquina, devido à ausência de variável resposta.
  - No caso de segmentação, caso se preze pela otimização **estrutural** dos *clusters*, o WSS pode ser utilizado como métrica para comparação e escolha do cenário final.

$$WSS = \sum_{k=1}^{K} \left(\sum_{i=1}^{n_k} d(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_k)^2 \right)$$

Quanto menor o WSS, maior a homogeneidade dentro dos *clusters*

onde:

- K é a quantidade de *clusters*
- n_k é a quantidade de observações dentro do *cluster* k, $k = 1, \dots, K$
- $\boldsymbol{x}_i$ é o vetor de coordenadas referentes a cada observação
- $\overline{\boldsymbol{x}}_k$ é o vetor de coordenadas referentes ao centroide ou medoide do *cluster* k
- $d(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_k)$ é uma medida de distância entre $\boldsymbol{x}_i$ e $\overline{\boldsymbol{x}}_k$

FIA BUSINESS SCHOOL lab.data

Passo 10: Implementação em Produção

No contexto corporativo, a fase de **implementação** de um modelo em ambiente produtivo garante que ele possa funcionar de forma robusta e eficiente mediante demanda de novas classificações, projeções ou agrupamentos.

Essa tarefa costuma ser realizada por profissionais especializados e com *skills* técnicos mistos entre *machine learning* e engenharia de *software*, tais como engenheiros de *machine learning*, engenheiros de *analytics* ou mesmo cientistas de dados.

As etapas principais são:

- **Integração**: Integrar o modelo treinado com os sistemas de dados existentes, incluindo implementação em ambientes de nuvem (AWS, GCP, Azure).
- **Automatização**: Criar *pipelines* para ingestão automática de novos dados e geração de *outputs* para novas observações.
- **Manutenção**: Planejar atualizações sistêmicas e garantir que seja possível lidar com aumento no volume ou complexidade de dados ao longo do tempo.

Passo 11: Monitoramento de Modelo

Uma vez que o modelo está implementado em ambiente produtivo, é importante **monitorar a sua estabilidade** de forma contínua.

Algumas questões nas quais estamos interessados nesta etapa são:

- O comportamento das **variáveis de entrada** (variáveis explicativas e resposta, se aplicável) permanece estável em relação à base de desenvolvimento?
- As métricas de **desempenho** (caso seja supervisionado) permanecem estáveis em relação à base de desenvolvimento?
- A **distribuição** (%) e as **características** de cada *cluster* (caso se trate de uma segmentação) permanecem estáveis em relação à base de desenvolvimento?

Tais questões podem ser refletidas por meio de **indicadores** (KPIs) a respeito do modelo, que podem ser analisados de forma evolutiva mediante técnicas de análise exploratória (tabelas, gráficos).

É comum elaborar ***dashboards* de visualização de dados** que facilitem o acompanhamento e a identificação rápida de eventuais anomalias que possam prejudicar a tomada de decisão baseada no modelo.

Passo 12: Recalibragem de Modelo

Passado certo tempo da implantação do modelo, caso se identifique no monitoramento que o comportamento dos dados está mudando e/ou o desempenho do modelo está caindo, o ideal é **recalibrar o modelo**.

A recalibragem consiste, essencialmente, em reconstruir o modelo com novos dados, a fim de **capturar os padrões mais recentes** e melhorar a qualidade das próximas classificações, projeções ou agrupamentos.

Neste momento, pode-se trocar ou não o tipo de técnica (algoritmo) utilizada.

É possível que as recalibragens do modelo sejam realizadas de forma manual ou automática, a depender dos seguintes fatores:

- **Recursos computacionais:** Disponibilidade de infraestrutura para suportar recalibragens automáticas recorrentes.
- **Complexidade**: Dificuldade associada à atualização do modelo.
- **Disponibilidade de dados**: Quantidade e frequência de novos dados disponíveis para atualizar o modelo.
- **Volatilidade**: Frequência com a qual os padrões de comportamento da base modelada sofrem alterações (refletida por meio do monitoramento).

FIA BUSINESS SCHOOL lab.data

3. Criação de *Features*

Passo a Passo de Modelagem

O seguinte fluxo abrange os passos para construção de um modelo estatístico ou de *machine learning*. Discutiremos cada um deles a seguir.

1. Definição da tarefa
2. Definições de dados
3. **Criação de *features***
4. Análise exploratória de dados
5. Redução de dimensionalidade
6. Padronização de variáveis
7. Definição de algoritmo(s)
8. Desenvolvimento de modelo(s)
9. Escolha do modelo final
10. Implementação em produção
11. Monitoramento de modelo
12. Recalibragem de modelo

Objetivo

Retornando à nossa analogia de receita culinária:

> Quão bom tende a ser o resultado final da receita se ela for executada por um *chef* renomado, com uma cozinha bem equipada e com os melhores utensílios e eletrodomésticos?

Espera-se que tais fatores contribuam para um bom resultado da receita. Porém:

> Quão bom tende a ser o resultado final da receita se ela for executada por um *chef* renomado, com uma cozinha bem equipada e com os melhores utensílios e eletrodomésticos, **mas com os ingredientes errados**?

Evidentemente, de nada adianta ter um bom profissional e boas ferramentas, se os insumos para a condução da receita não forem adequados.

O mesmo vale para a modelagem estatística e de *machine learning*, onde o profissional é o **cientista de dados**; os instrumentos são os **algoritmos**; e o insumo são os **dados**, que devem ter qualidade e **refletir bem o problema de interesse**.

→ *Análise exploratória*

Objetivo

O **senso crítico** e o **conhecimento de negócio** são essenciais para definir quais *features* devem ser englobadas em um modelo não supervisionado, tal como a clusterização, ou testadas em um modelo supervisionado, tal como a regressão.

No caso de modelos **supervisionados**, especificamente, o desempenho está diretamente associado ao poder das informações que foram consideradas como *input*, independentemente do tipo de algoritmo utilizado.

Vejamos um exemplo, a seguir.

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

1 Definição da tarefa

Para resolver o problema, qual tarefa precisamos solucionar?

- Classificação
- Projeção não temporal
- Projeção temporal
- Agrupamento

FIA BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

1 Definição da tarefa

Para resolver o problema, qual tarefa precisamos solucionar?

- **Classificação** → Queremos **classificar** os clientes em duas categorias:
  - os que **compram** uma passagem internacional entre abril e junho;
  - os que **não compram** uma passagem internacional entre abril e junho.
- Projeção não temporal
- Projeção temporal
- Agrupamento

FIA BUSINESS SCHOOL lab.data

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

FIA BUSINESS SCHOOL lab.data

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

➢ Visão **cliente comprador**

Poderíamos também pensar na visão **cliente viajante**, ou seja, cada passageiro que voou nos últimos 24 meses. Porém, como estamos interessados em avaliar propensão à compra, é mais coerente consolidar a base na visão dos compradores.

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

SUGESTÕES

- Incluir apenas clientes **ativos**, ou seja, aqueles que já haviam comprado uma viagem aérea (nacional ou internacional) nos últimos 24 meses (M0 a M-23).
- Incluir apenas clientes com pelo menos **18 anos de idade**.

FIA BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

SUGESTÕES

- Excluir clientes da categoria **diamante** do programa de milhas.
- Excluir clientes que viajaram apenas com tarifas **corporativas**.
- Excluir clientes **estrangeiros** (sem residência fixa no BR).

FIA BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

➢ Temos uma variável resposta **binária**:

- **1 (um)** se o cliente faz **compra** de ao menos uma passagem aérea internacional entre abril e junho, independentemente da data agendada para o voo.
- **0 (zero)** caso contrário.

FIA BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

- O ideal é considerar ao menos **duas safras**: uma para construção do modelo e outra para validação *out-of-time.*
- Por exemplo:
  - Construção na safra de M0 = **31/mar/23**
  - Validação na safra de M0 = **31/mar/24**

FIA BUSINESS SCHOOL lab.data

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados

Realizemos as principais definições a respeito da base de dados:

- Qual a **visão** da base?
- Quais os **possíveis filtros** de inclusão e exclusão?
- Qual a **variável resposta**?
- Quantas e quais **safras** devem ser consideradas?
- Quais os **períodos** de histórico e de previsão?

➢ Para cada safra:

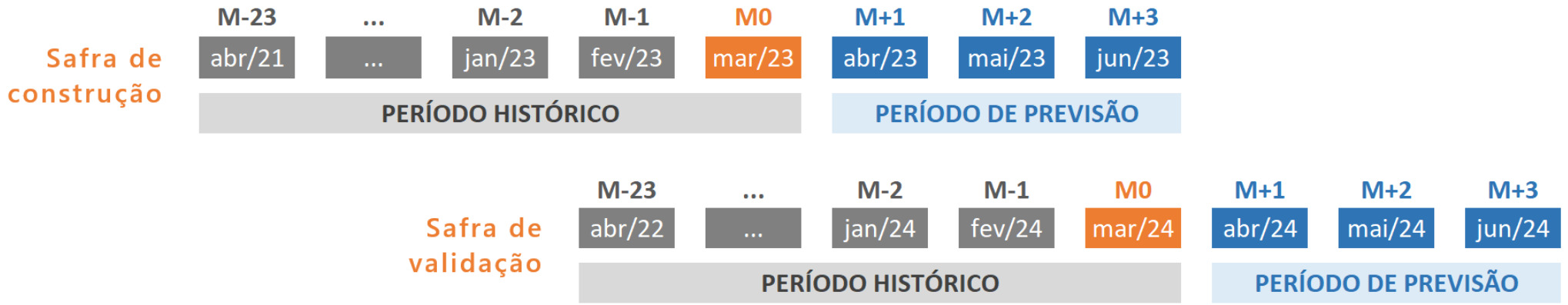
- Histórico: **M0 a M-23**
- Previsão: **M+1 a M+3**

FIA BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

2 Definições de dados



FIA BUSINESS SCHOOL lab.data

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

3 Criação de *features*

Quais são as possíveis *features* a serem testadas no modelo?

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

3 Criação de *features*

Para facilitar o trabalho de idealização de *features*, além de conhecer sobre o negócio, é importante saber quais informações estão **disponíveis** nas camadas de dados mais elementares, também chamadas comumente de “bases brutas”.

Suponha que **oito origens de dados** estão disponíveis para o nosso problema, conforme *slide* a seguir. A partir delas, proponha possíveis *features* a serem testadas no modelo.

FIA BUSINESS SCHOOL lab.data

Case: Companhia Aérea

Uma companhia aérea deseja identificar quais clientes são mais propensos a comprar uma passagem internacional entre os meses de abril e junho, com base em informações cadastrais e transacionais do seu histórico como viajantes nos últimos 24 meses.

DADOS CADASTRAIS

Visão: Cada cliente que comprou

DADOS DE CAMPANHAS

Visão: Cada abordagem realizada (e-mail, SMS, app)

DADOS DE SAC

Visão: Cada contato realizado ou recebido pela central

DADOS DE MILHAGEM

Visão: Cada operação realizada (acúmulo ou resgate)

DADOS DE TRANSAÇÕES

Visão: Cada compra realizada (on-line, agências etc.)

DADOS DE PASSAGEIROS

Visão: Cada passageiro que voou

DADOS DE CHECK-IN

Visão: Cada check-in realizado (on-line, totem, balcão etc.)

DADOS DE SERVIÇOS

Visão: Cada serviço comprado (assento, wi-fi a bordo etc.)

FIA BUSINESS SCHOOL lab.data

Dicas para Criação de Variáveis

A seguir, lista-se algumas dicas que podem auxiliar na elaboração de *features* em situações diversas, ainda que a sua definição exata dependa do conhecimento de cada contexto específico.

1. Utilizar **agregações**, tais como média, soma, mínimo/máximo, contagens simples e contagens acumuladas.
2. Variar as **janelas temporais** de cálculo histórico.

   Pode gerar colinearidade, o que prejudica a interpretação, sob a ótica do aprendizado estatístico. Neste caso, requer uma análise subsequente para escolha da melhor janela temporal para cada variável. Porém, a colinearidade não prejudica o desempenho, que é o foco principal sob a ótica de machine learning.
3. Criar variáveis de **proporção** (percentual) e **razão**.
4. Criar variáveis que denotam **variações** de determinados indicadores entre janelas temporais.
5. Criar variáveis que denotam **tempo desde a ocorrência** de determinados eventos (recência).
6. Criar réplicas das variáveis para diferentes **segmentos** e **categorias**.
7. Criar variáveis que denotem aspectos de **dispersão** (ex.: desvio padrão), se fizer sentido.
8. Criar variáveis indicadoras da ocorrência de **eventos** ou presença de **características específicas** *(flags)*.

Layout de Modelagem

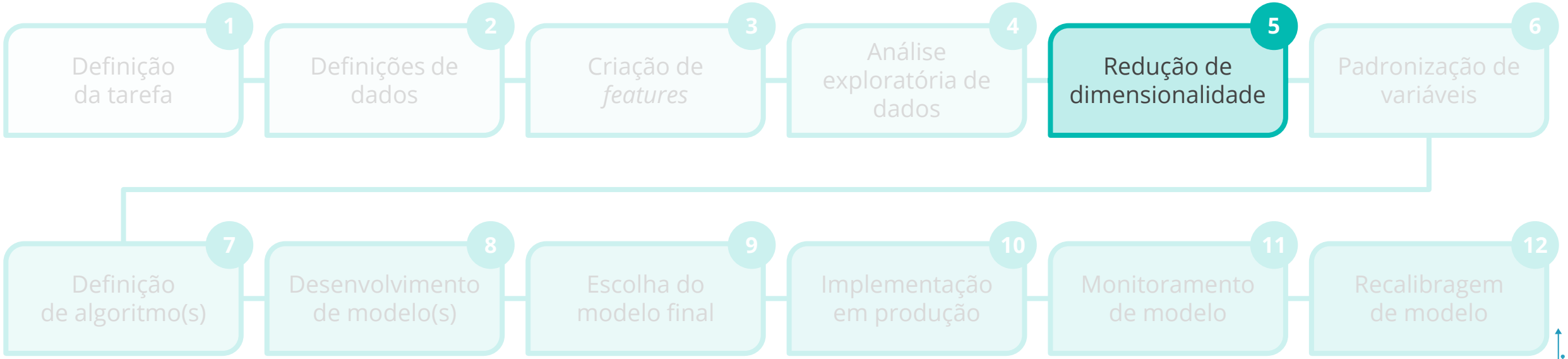
A base de dados final utilizada para a modelagem costuma ser chamada de **base analítica**, proveniente do inglês *analytical base table* (ABT).

Convém criar um documento que sumariza todas as especificações realizadas acerca da base analítica para construção do modelo, incluindo a lista completa de variáveis envolvidas. Este documento costuma ser chamado de ***layout* de modelagem**, ou ***layout* de extração para modelagem**.

4. Redução de Dimensionalidade

Passo a Passo de Modelagem

O seguinte fluxo abrange os passos para construção de um modelo estatístico ou de *machine learning*. Discutiremos cada um deles a seguir.



FIA BUSINESS SCHOOL lab.data

Objetivo

Com frequência, lidamos com bases de dados com grande quantidade de variáveis concomitantemente **associadas** entre si. Ou seja, carregam uma certa porção de informações redundantes.

Quando havíamos construído modelos de **regressão** de forma artesanal, prezando pela interpretação das variáveis, tentamos contornar problemas de colinearidade por meio da retirada de algumas variáveis e cálculo da estatística VIF.

Porém, quando a base de dados possui maior volume de variáveis e/ou não estamos preocupados com a **interpretabilidade** (sob a ótica de *machine learning*), podemos tratar as redundâncias de forma mais eficiente por meio de técnicas de redução de dimensionalidade.

Objetivo

As técnicas de **redução de dimensionalidade** têm como objetivo realizar composições/transformações entre variáveis potencialmente correlacionadas em uma base de dados, de forma que se obtenha um conjunto **menor** de variáveis **independentes** entre si, e que preservem uma porção relevante da informação presente nos dados originais.

***Exemplo**: Partindo de 30 variáveis com diferentes graus de associação, obter um novo conjunto de 5 variáveis transformadas, independentes entre si, que retêm 90% da informação presente nas 30 variáveis originais.*

Técnica Utilizada

A seguir, vamos estudar a técnica de **análise de componentes principais (ACP)**, a mais utilizada em redução de dimensionalidade. Ela é baseada no conceito de vetores e em princípios de álgebra linear.

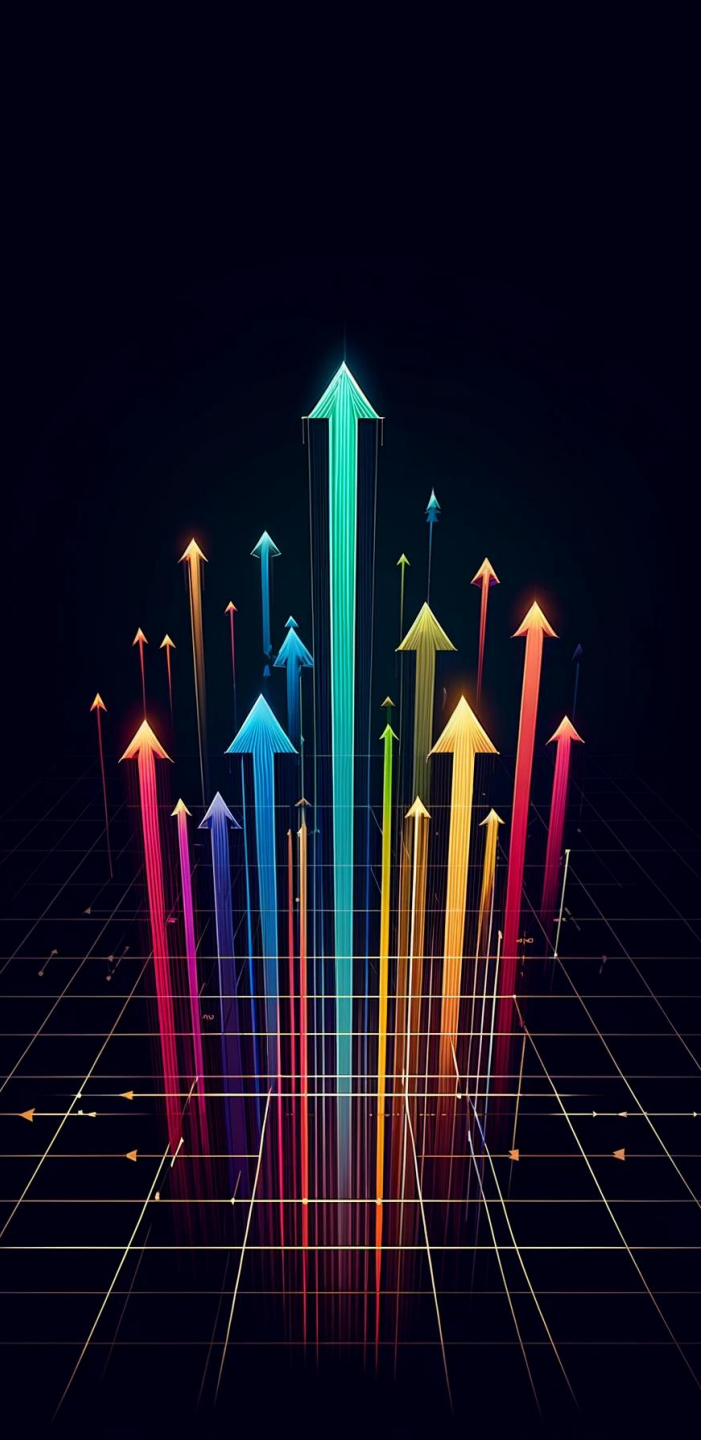
A ACP é especialmente útil quando as variáveis são **quantitativas** e as relações entre elas são aproximadamente **lineares**. Entretanto, pode ser utilizada também para variáveis qualitativas (se codificadas como *dummies*) e/ou quando há relações não lineares, ainda que com menor efetividade de redução.

Exemplo:

i. Em um conjunto de 20 variáveis quantitativas com relações lineares entre si, é possível que 2 novos componentes já sejam suficientes para reter grande parte da informação original.

ii. Já em um conjunto de 20 variáveis mistas (quantitativas e qualitativas), com diferentes tipos de relações entre si, é possível que precisemos de mais componentes para garantir o mesmo grau de retenção de informação obtido com 2 componentes no caso (i).

5. Análise de Componentes Principais

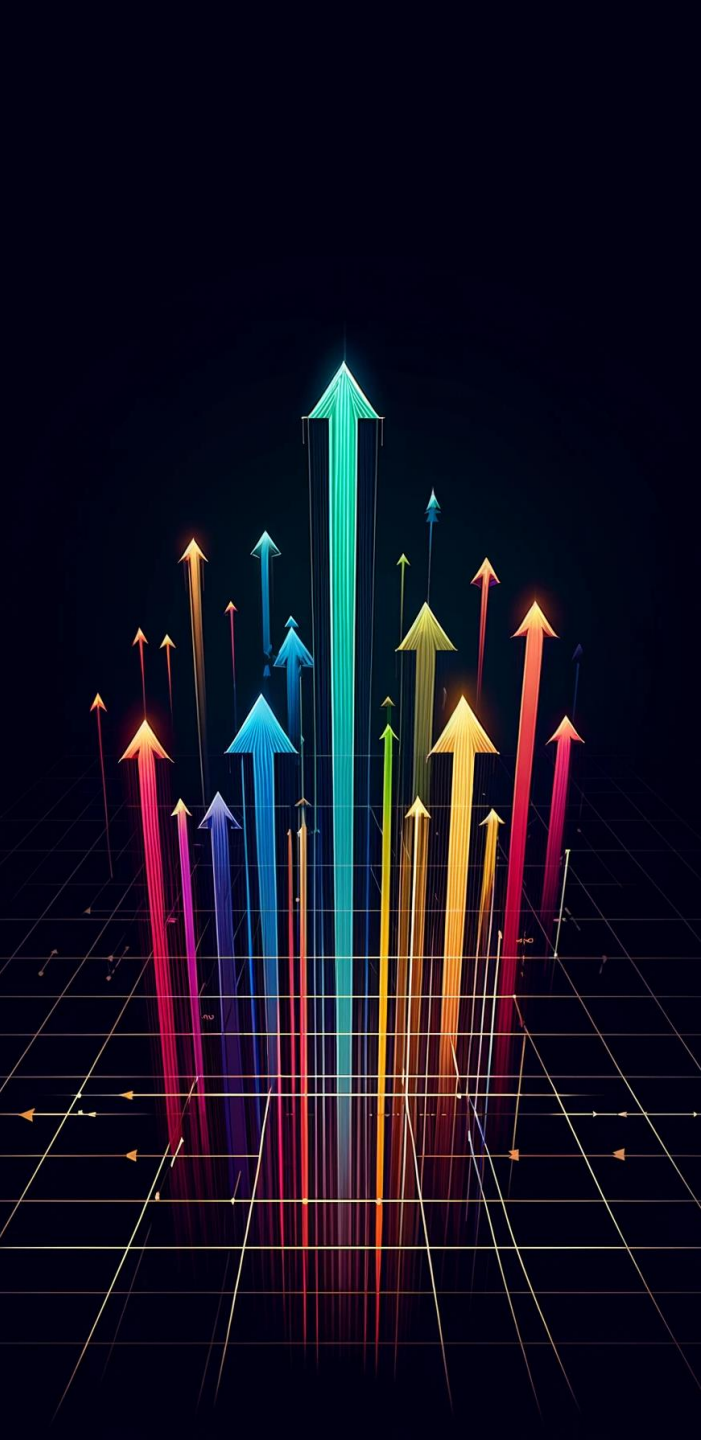


Definição de Vetor

Para entender a estrutura da análise de componentes principais, precisamos relembrar o conceito de **vetor**.

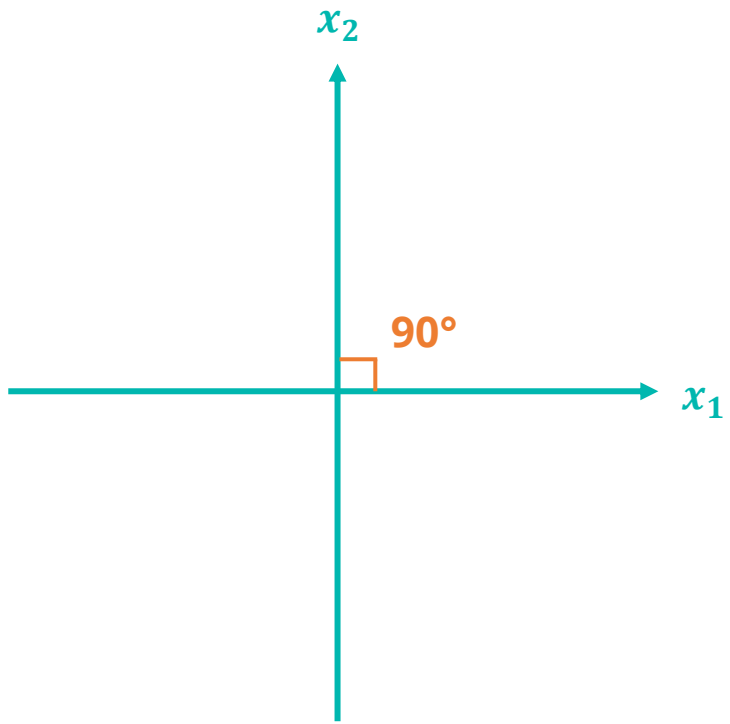
Na Física, um **vetor** é um segmento de reta que representa grandezas físicas que envolvem magnitude, direção e sentido, tais como força e velocidade. Geralmente, vetores são representados graficamente como setas.

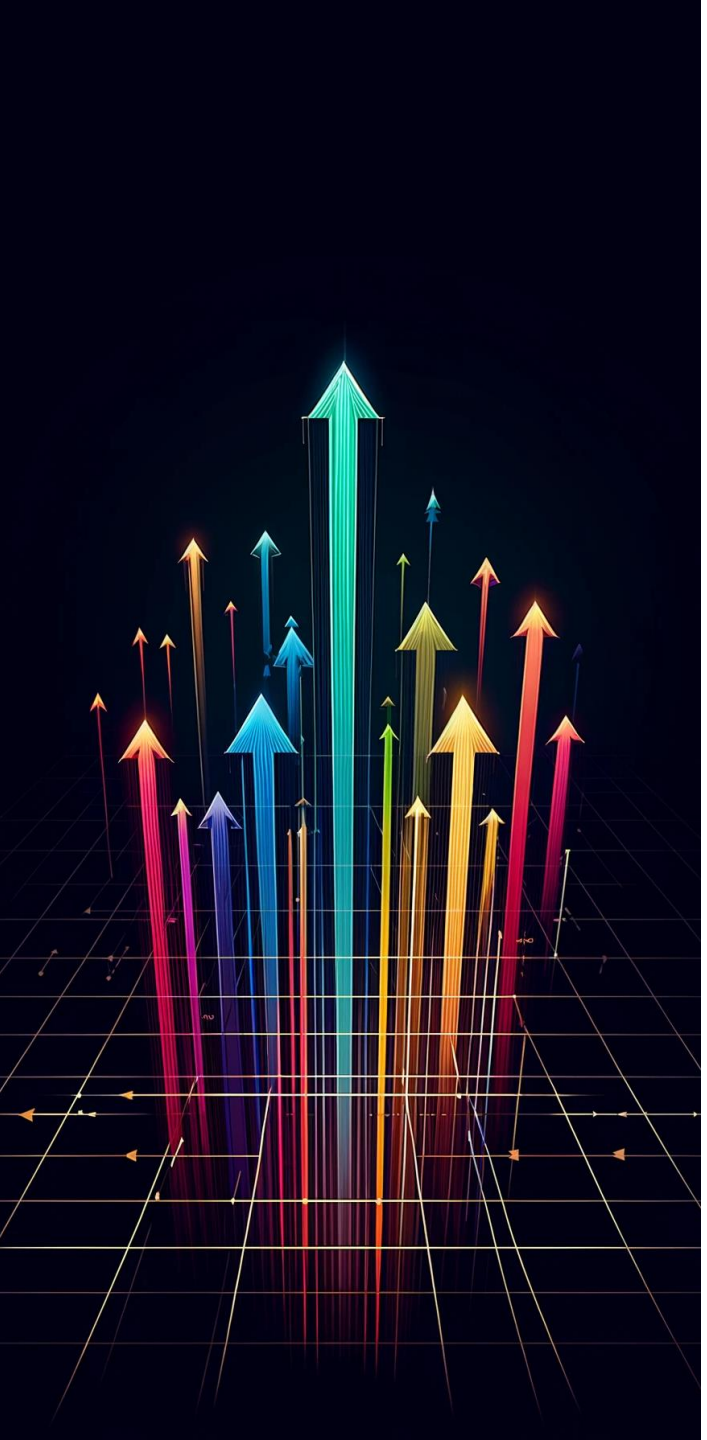
Para o uso de vetores na análise de componentes principais, o aspecto de magnitude será irrelevante, ou seja, importará apenas a **direção** e o **sentido** do vetor.



Vetores Perpendiculares

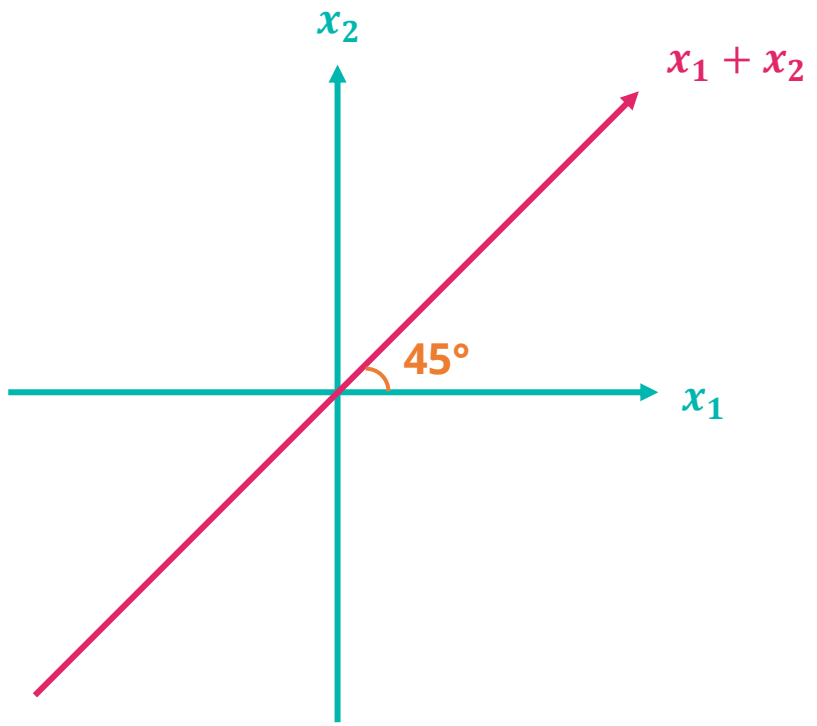
Ao representar um gráfico quantitativo em duas dimensões, os dois **eixos** podem ser compreendidos como **vetores perpendiculares**, isto é, que possuem um ângulo de 90° entre eles.

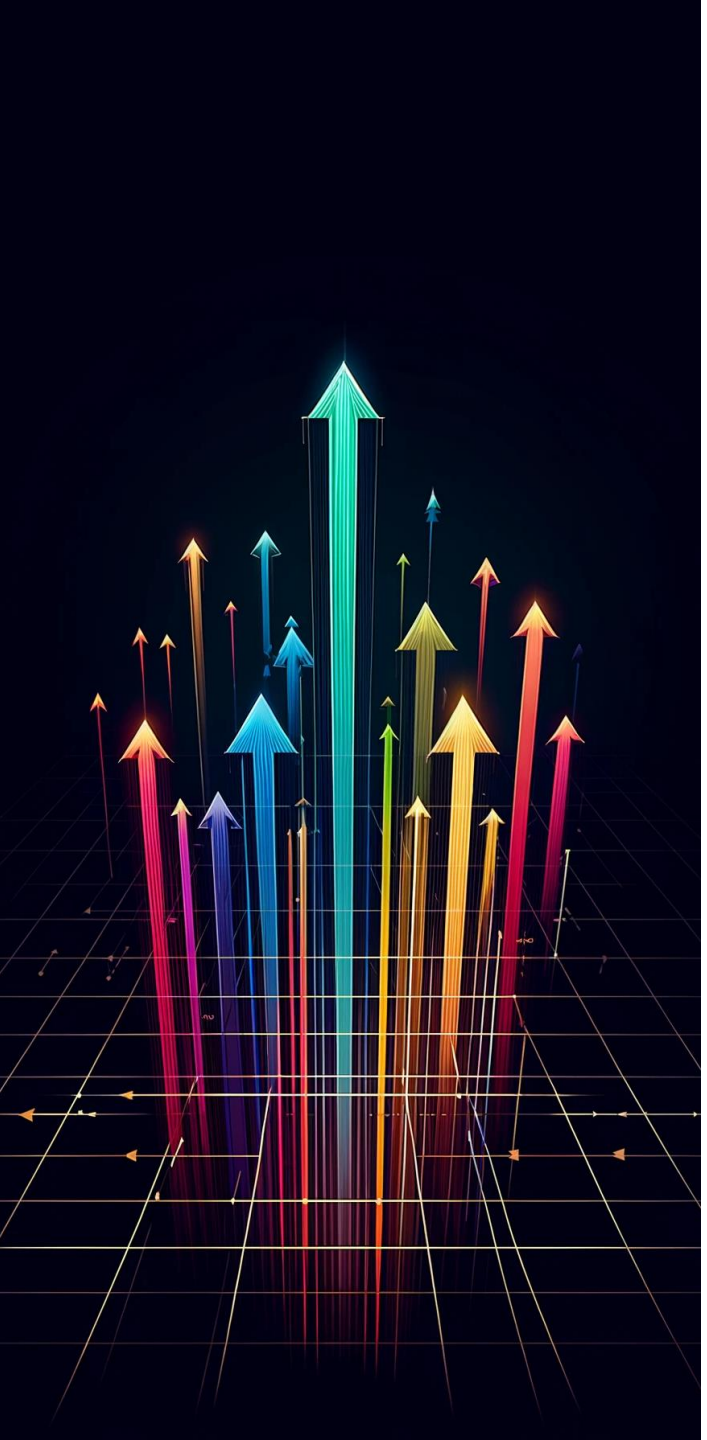




Vetor Resultante

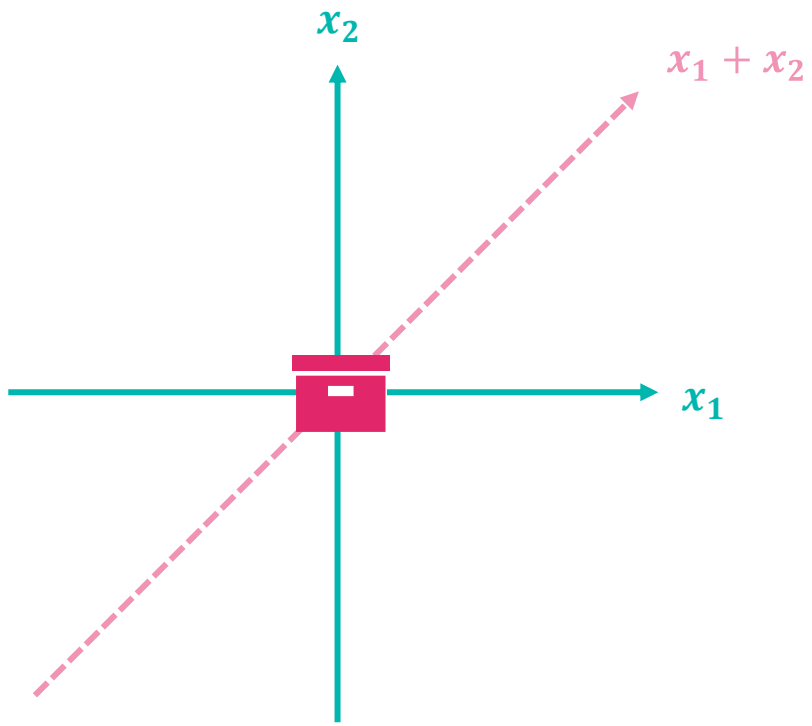
O **vetor resultante** a partir de dois vetores perpendiculares x_1 e x_2, também chamado de **vetor soma**, corresponde ao vetor com ângulo de 45° que passa pelo meio dos vetores x_1 e x_2.

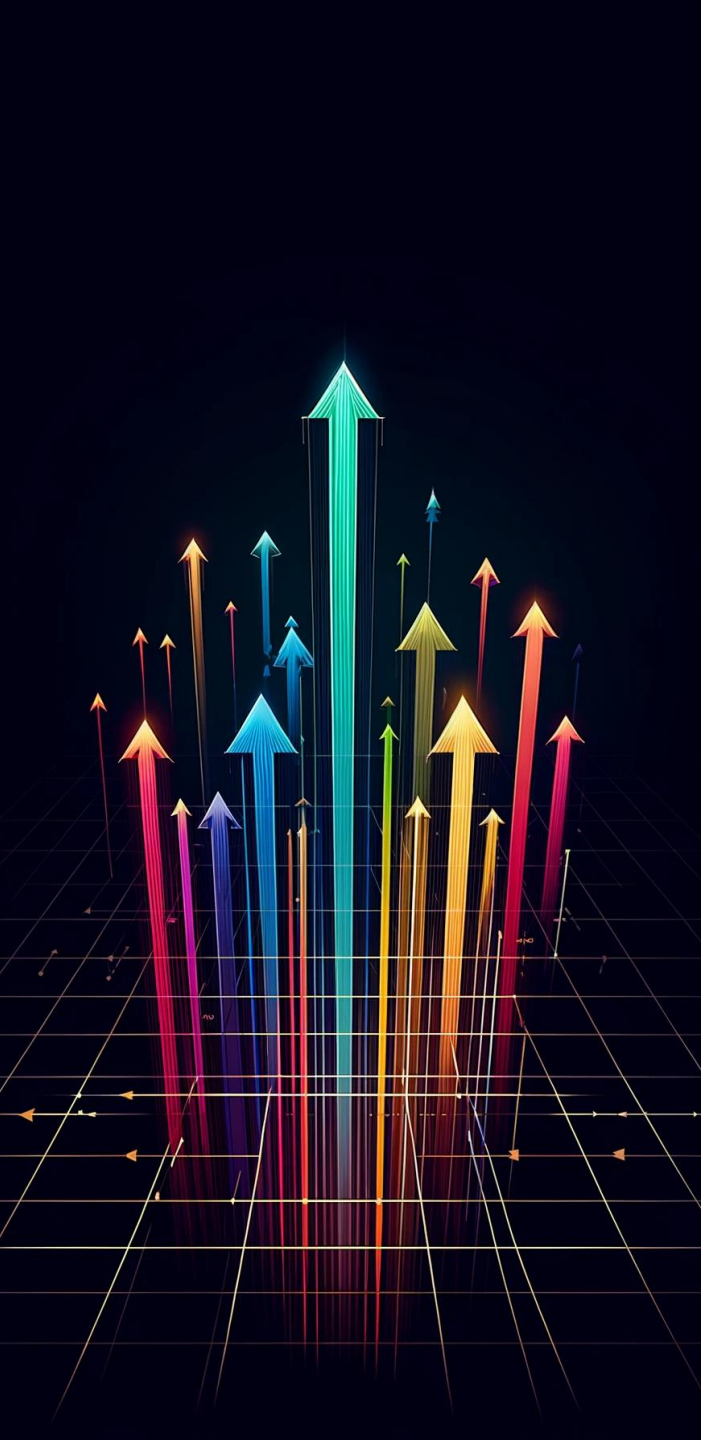




Vetor Resultante

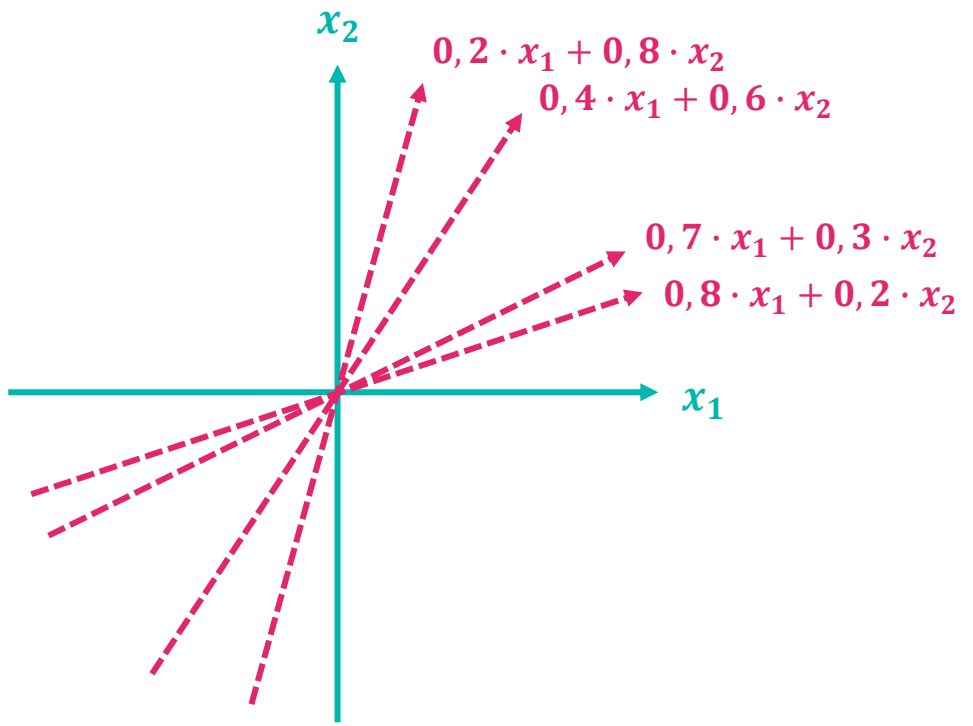
Exemplo: Pensando na grandeza de **força**, imagine que um objeto situado ao centro do gráfico é "empurrado" por duas forças independentes e de magnitude idêntica, uma na direção de x_1 (leste) e outra na direção de x_2 (norte). O objeto se deslocará na direção do vetor $x_1 + x_2$, como resultado da combinação das forças.

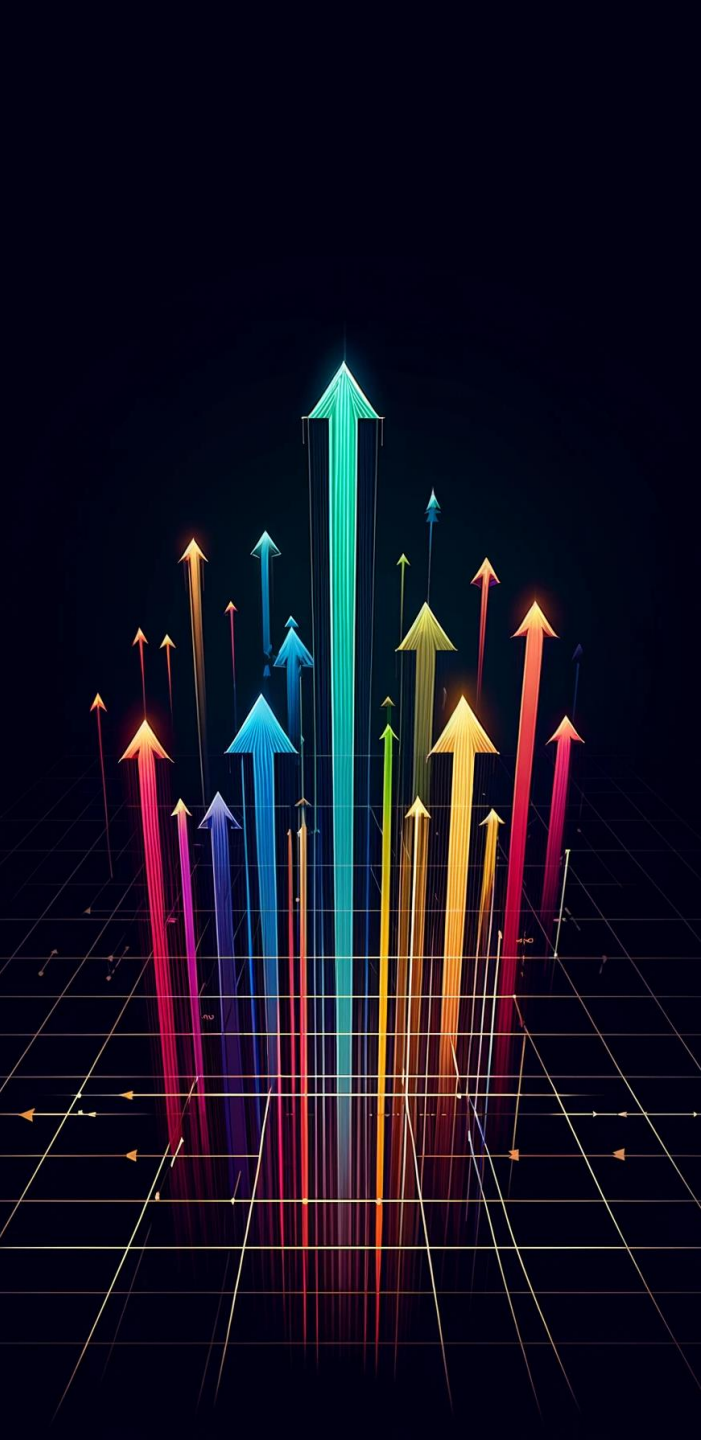




Vetor Resultante

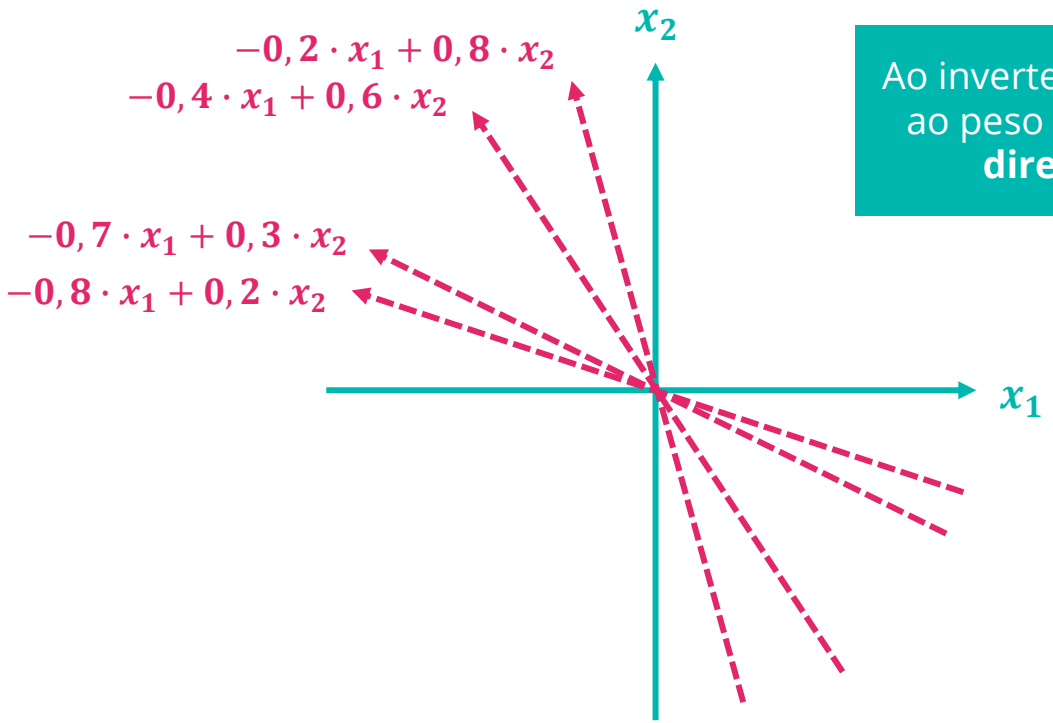
De forma análoga, qualquer outra **direção** em duas dimensões pode ser obtida a partir de uma **soma ponderada** dos vetores perpendiculares x_1 e x_2, sendo que os pesos da ponderação definem a direção do vetor resultante.





Vetor Resultante

De forma análoga, qualquer outra **direção** em duas dimensões pode ser obtida a partir de uma **soma ponderada** dos vetores perpendiculares x_1 e x_2, sendo que os pesos da ponderação definem a direção do vetor resultante.

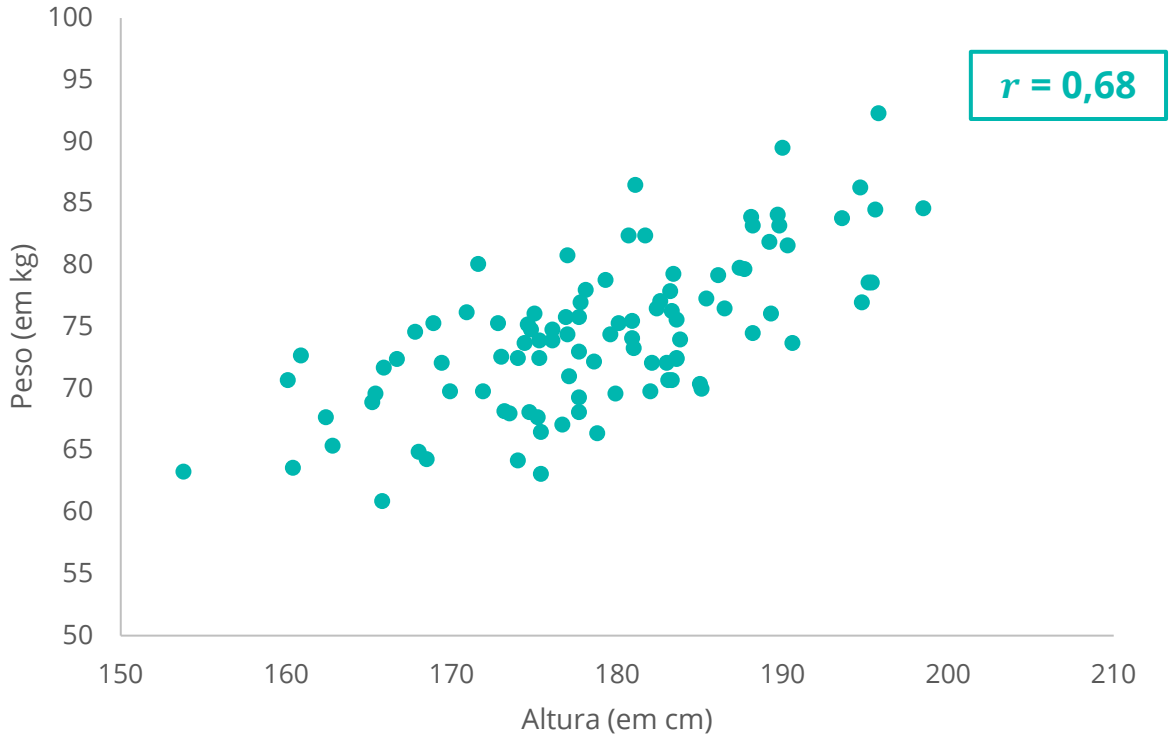


Ao inverter o **sinal** associado ao peso de x_1, inverte-se a **direção** do vetor.

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

NADADOR	ALTURA	PESO
001	180,9	74,1
002	170,9	76,2
003	175,2	67,7
004	190,3	81,6
005	173,2	68,2
006	180,7	82,4
007	190,6	73,7
008	160,1	70,7
009	169,4	72,1
010	188,2	83,2
...	...	...



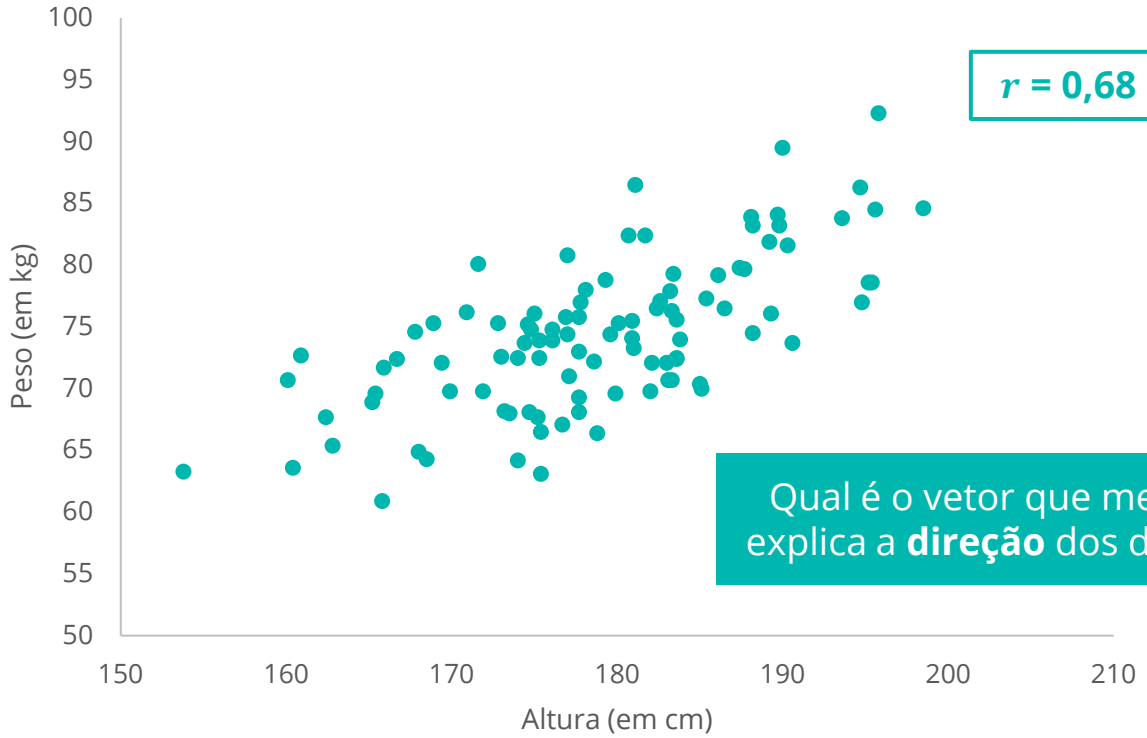
Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

NADADOR	ALTURA	PESO
001	180,9	74,1
002	170,9	76,2
003	175,2	67,7
004	190,3	81,6
005	173,2	68,2
006	180,7	82,4
007	190,6	73,7
008	160,1	70,7
009	169,4	72,1
010	188,2	83,2
...	...	...



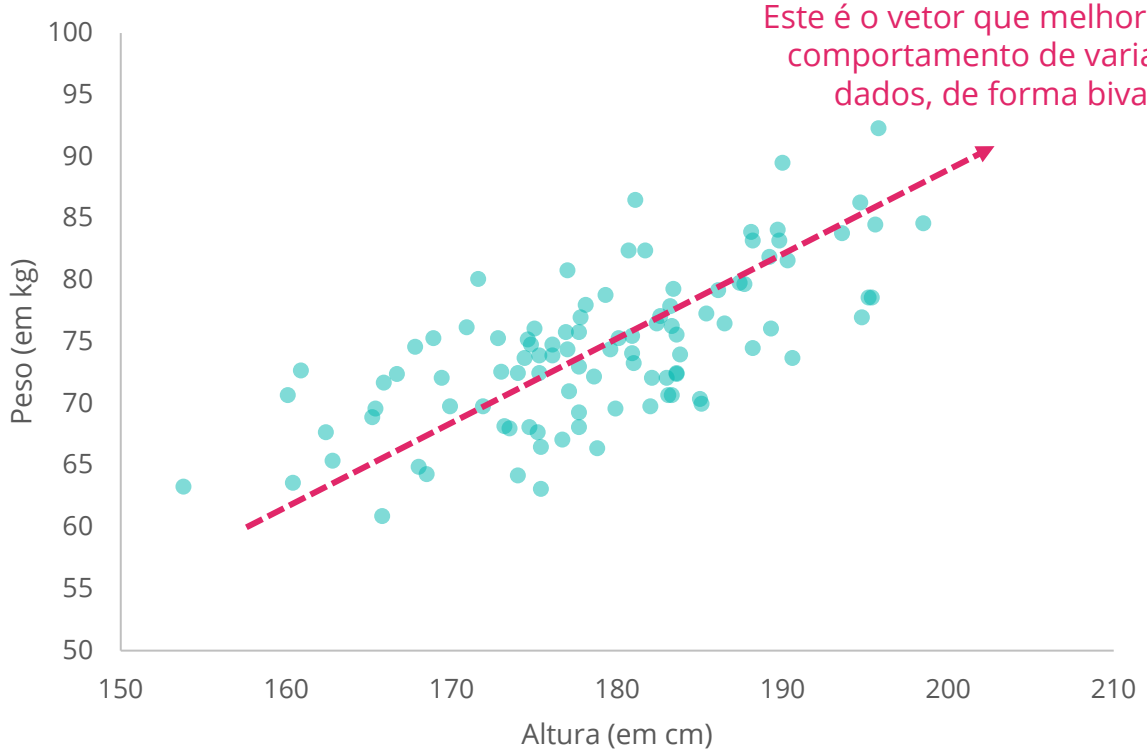
Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

NADADOR	ALTURA	PESO
001	180,9	74,1
002	170,9	76,2
003	175,2	67,7
004	190,3	81,6
005	173,2	68,2
006	180,7	82,4
007	190,6	73,7
008	160,1	70,7
009	169,4	72,1
010	188,2	83,2
...	...	...



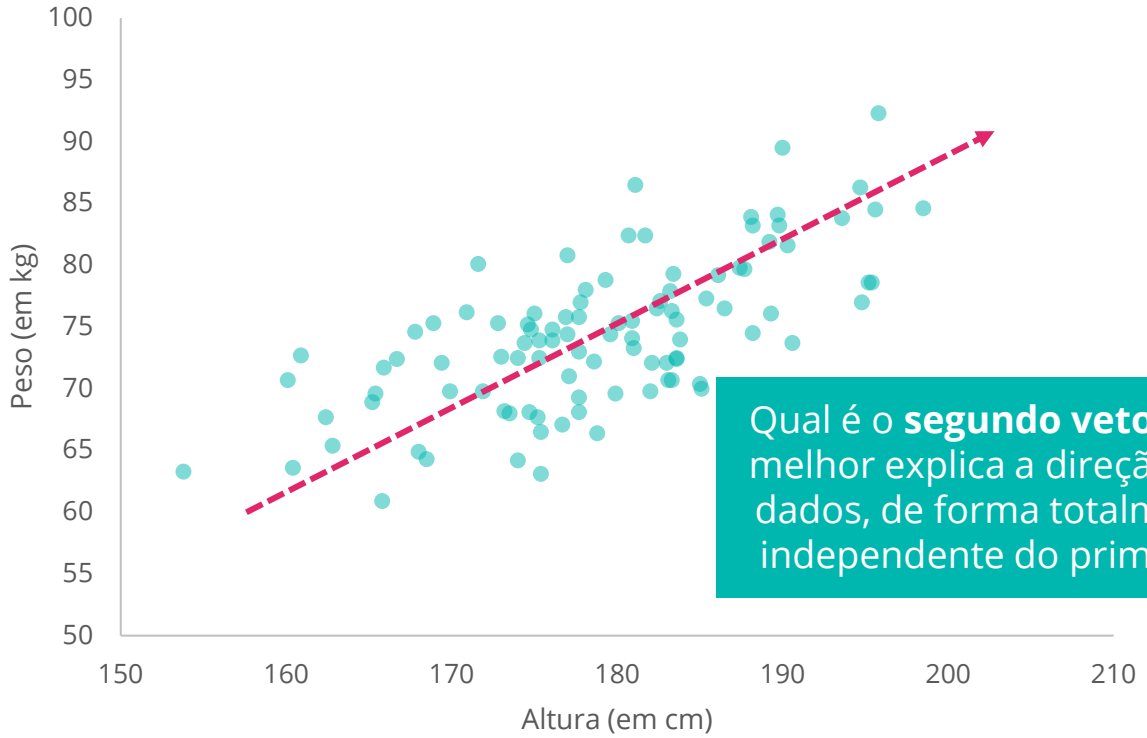
Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

NADADOR	ALTURA	PESO
001	180,9	74,1
002	170,9	76,2
003	175,2	67,7
004	190,3	81,6
005	173,2	68,2
006	180,7	82,4
007	190,6	73,7
008	160,1	70,7
009	169,4	72,1
010	188,2	83,2
...	...	...



Qual é o **segundo vetor** que melhor explica a direção dos dados, de forma totalmente independente do primeiro?

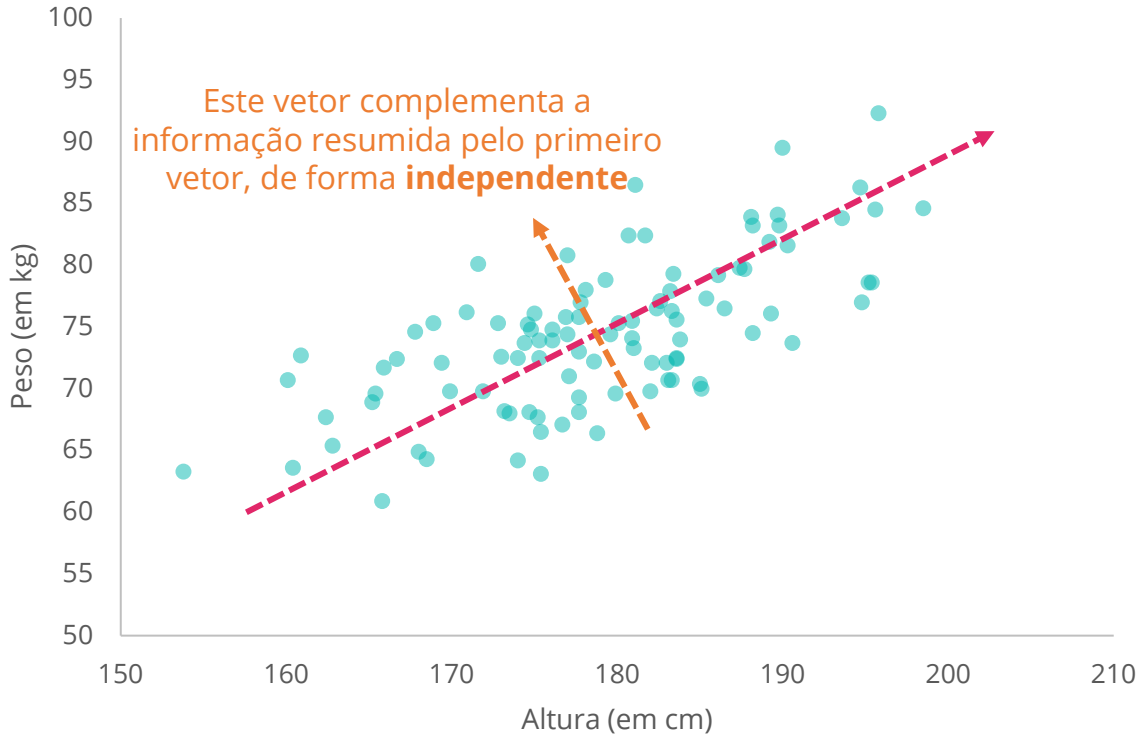
Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

NADADOR	ALTURA	PESO
001	180,9	74,1
002	170,9	76,2
003	175,2	67,7
004	190,3	81,6
005	173,2	68,2
006	180,7	82,4
007	190,6	73,7
008	160,1	70,7
009	169,4	72,1
010	188,2	83,2
...	...	...



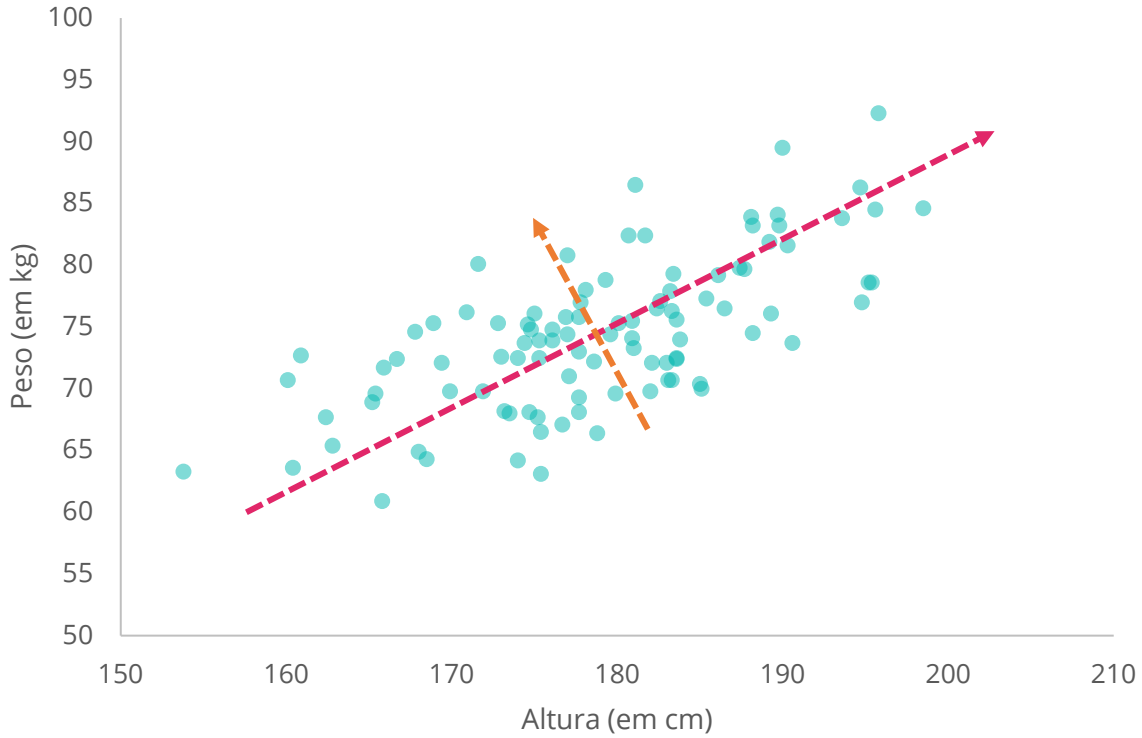
Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

A noção de "independência" entre os dois vetores provém do fato de eles serem **perpendiculares** (também chamados de **ortogonais**), ou seja, apontam para direções opostas.



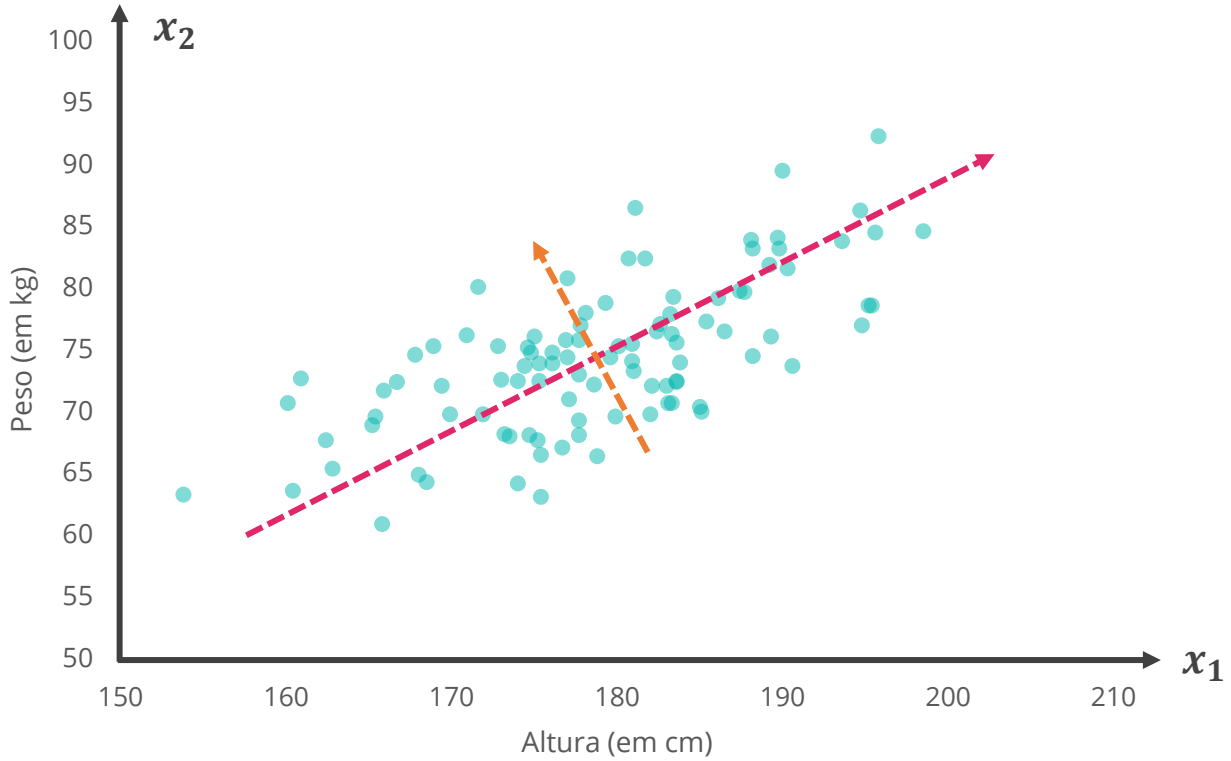
Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

A ideia também pode ser compreendida como uma **rotação** dos eixos x_1 e x_2, de forma que os novos vetores encontrados passem a ser os eixos do gráfico.



Arquivo: Antropometria.txt

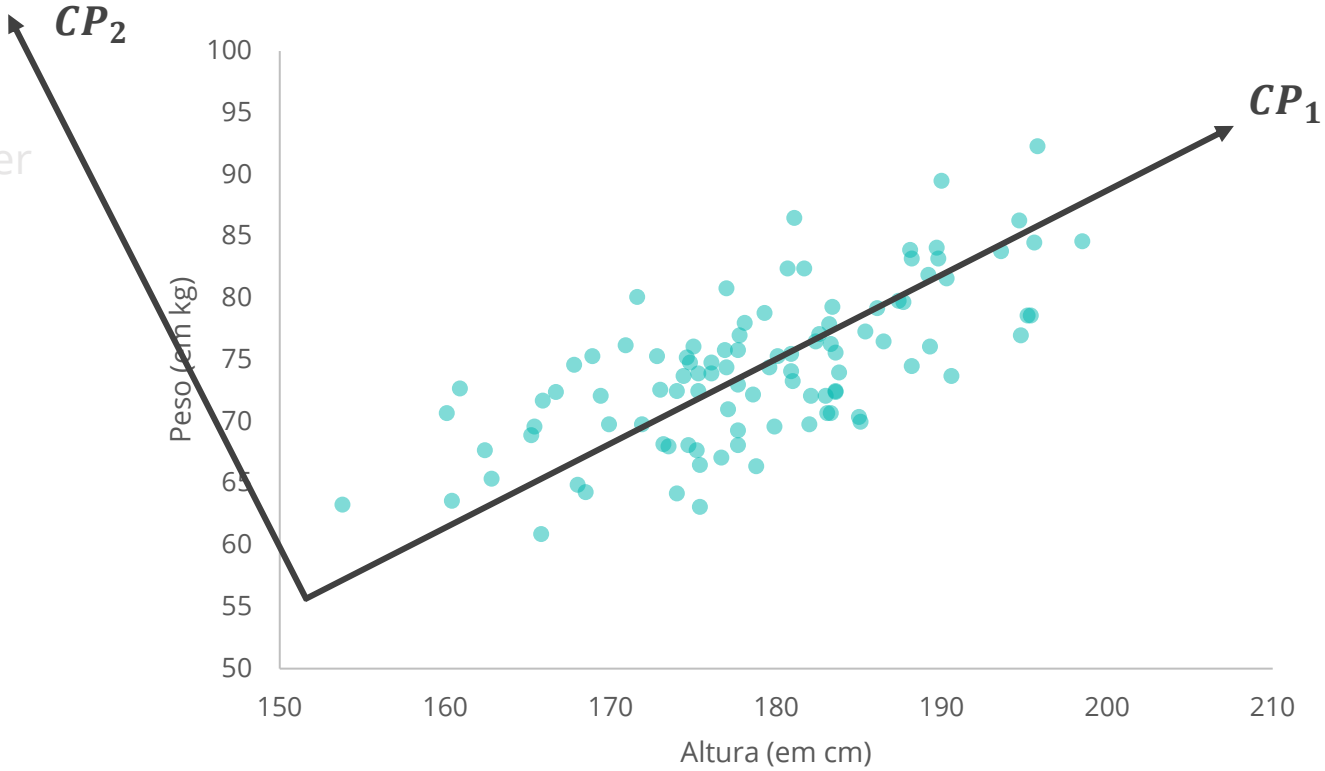
Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

A ideia também pode ser compreendida como uma **rotação** dos eixos x_1 e x_2, de forma que os novos vetores encontrados passem a ser os eixos do gráfico.

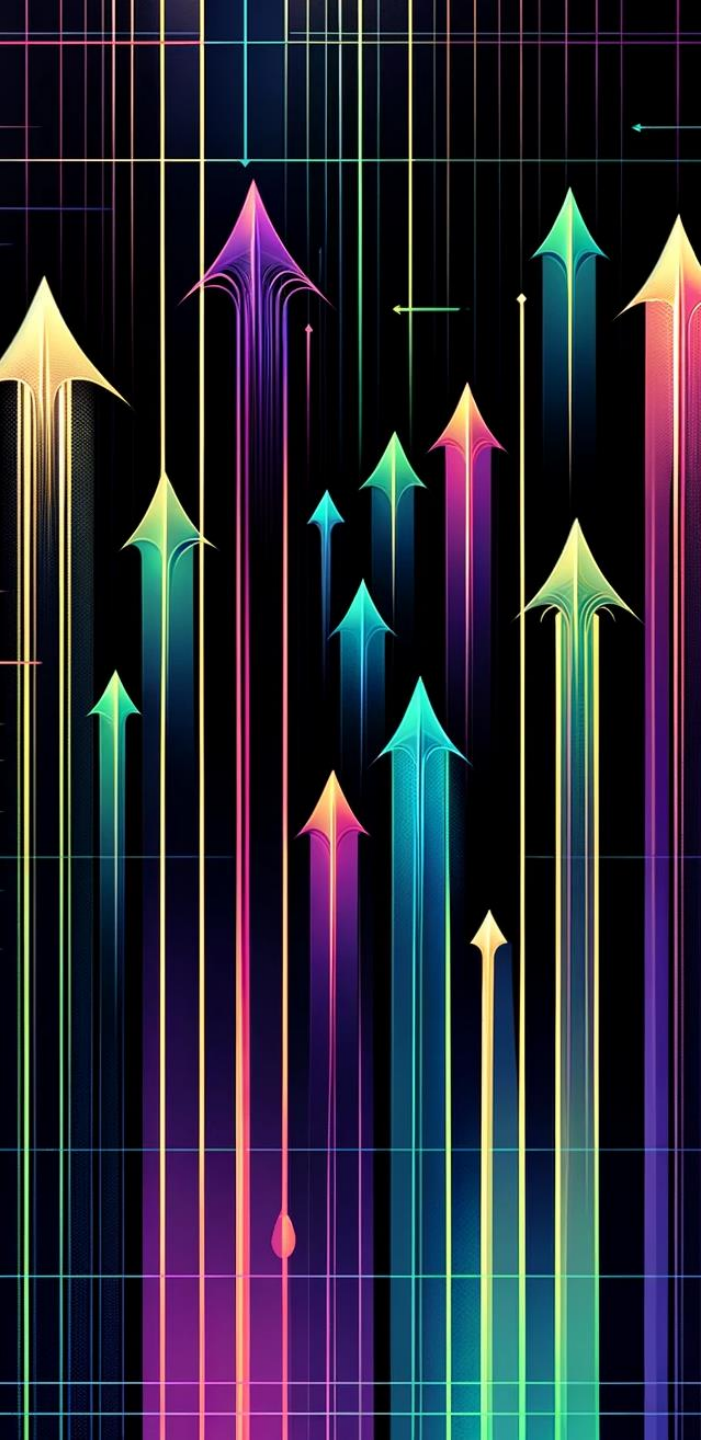
Fica claro que, nesse novo sistema de coordenadas, a altura e o peso passam a ser variáveis **independentes**.

Os novos eixos são ditos **componentes principais** associados aos dados originais.



Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data



Definição dos Componentes

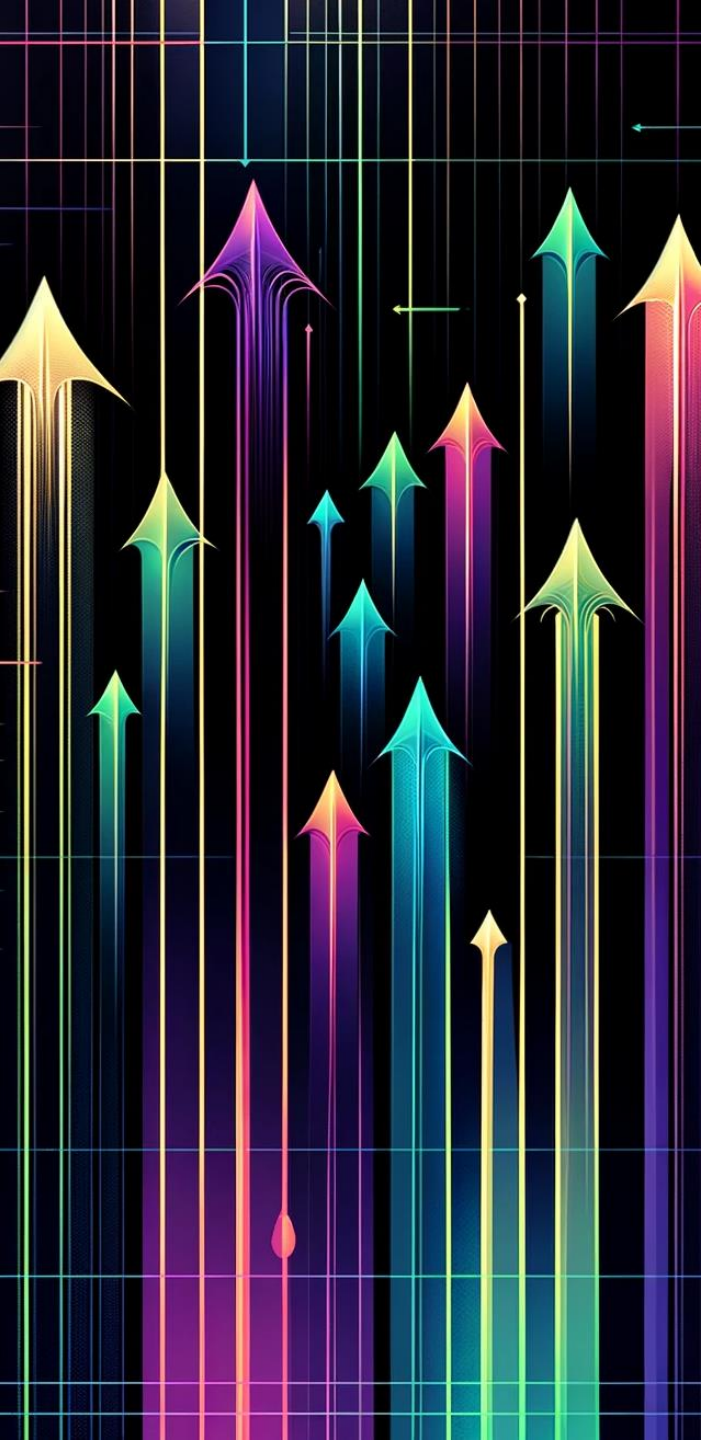
Os k **componentes principais** (CP's) de um conjunto de k variáveis quantitativas são vetores ortogonais resultantes da soma ponderada dos eixos iniciais. Em outras palavras, são **combinações lineares independentes** das variáveis originais, após padronização (para isolar o efeito de diferentes escalas).

$$\begin{cases} CP_1 = a_{11} \cdot Z_1 + a_{12} \cdot Z_2 + \ldots + a_{1k} \cdot Z_k \\ CP_2 = a_{21} \cdot Z_1 + a_{22} \cdot Z_2 + \ldots + a_{2k} \cdot Z_k \\ \qquad\qquad \ldots \\ CP_k = a_{k1} \cdot Z_1 + a_{k2} \cdot Z_2 + \ldots + a_{kk} \cdot Z_k \end{cases}$$

Os valores a_{ij} são chamados **coeficientes** (ou "pesos") associados a cada eixo original para obtenção dos novos eixos.

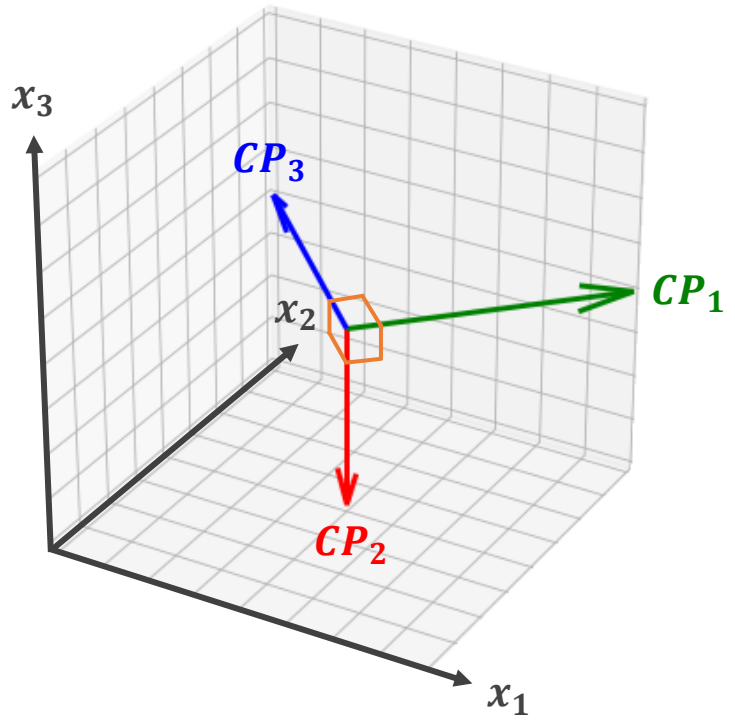
No *case* anterior de antropometria, em que tínhamos apenas **duas variáveis**, as expressões acima se reduzem a:

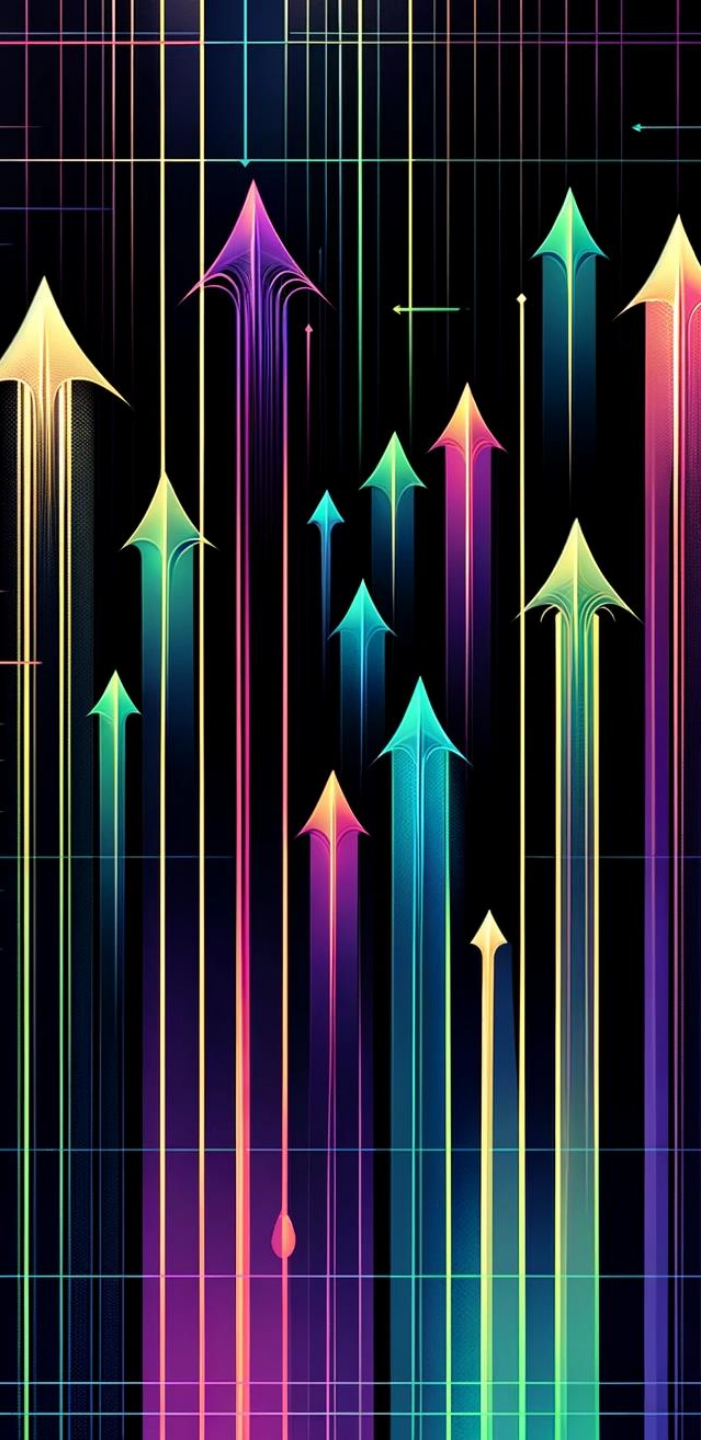
$$\begin{cases} CP_1 = a_{11} \cdot Z_1 + a_{12} \cdot Z_2 \\ CP_2 = a_{21} \cdot Z_1 + a_{22} \cdot Z_2 \end{cases}$$



Visualização Tridimensional

No caso de **três variáveis quantitativas**, ainda é possível visualizar graficamente os três componentes principais como uma **rotação tridimensional** do plano cartesiano original.





Obtenção dos Componentes

Os cálculos envolvidos na obtenção dos **componentes principais** envolvem conceitos mais complexos de álgebra linear e computação, que vão além do escopo do nosso curso*.

Em resumo, o processo de obtenção dos componentes envolve os seguintes passos:

1. **Padronização** das variáveis.
2. Cálculo do **coeficiente de correlação linear** entre todos os pares de variáveis.
3. Cálculo dos **autovetores** e **autovalores** associados à matriz de correlações.
    - Os autovetores correspondem aos vetores que melhor representam as **direções de variação** dos dados no espaço.
    - Os autovalores correspondem à **porção de variabilidade** dos dados originais que é representada (ou "retida") por cada autovetor.

** Referências adicionais:*

- *https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf*
- *https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf*

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

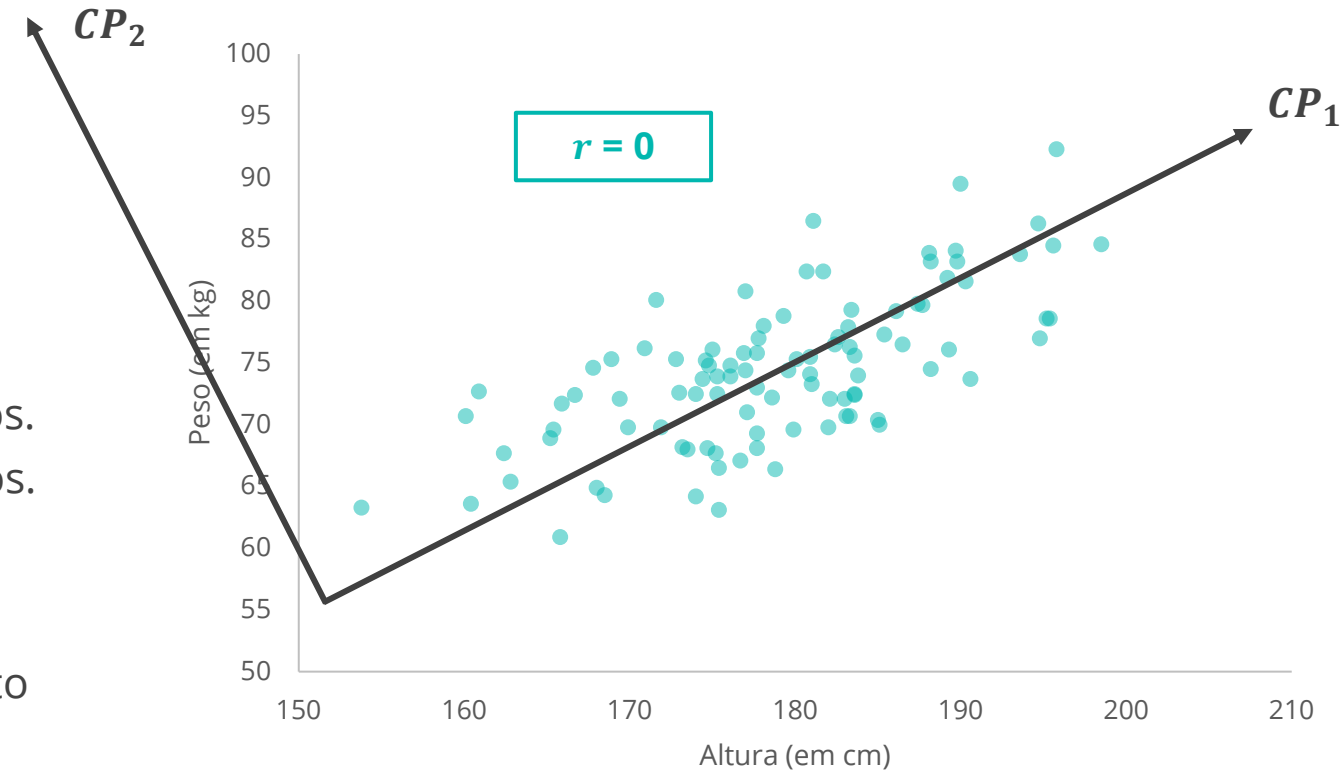
Por meio do Python, obtemos os seguintes componentes principais:

$$\begin{cases} CP_1 = 0{,}707 \cdot Z_1 + 0{,}707 \cdot Z_2 \\ CP_2 = 0{,}707 \cdot Z_1 - 0{,}707 \cdot Z_2 \end{cases}$$

Onde:

- CP_1 captura **84%** da variabilidade dos dados.
- CP_2 captura **16%** da variabilidade dos dados.
- CP_1 e CP_2 são **não correlacionados**.

Ou seja, um **único componente** é suficiente para explicar a maior parte do comportamento bivariado de peso e altura dos nadadores.



Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Antropometria

Um pesquisador da área de saúde esportiva levantou dados acerca da altura (em cm) e do peso (em kg) de 100 nadadores profissionais. Seria possível unificar esses dois indicadores em um único, que represente o porte geral do atleta?

Aplicando os novos componentes na base de dados, o componente 1 fornece uma espécie de *score* que resume boa parte do porte físico dos nadadores (84%) e permite **ordená-los**.

A imagem ao lado exibe os primeiros 10 nadadores com **maiores *scores*** obtidos a partir do primeiro componente principal.

Este procedimento se torna mais interessante quando temos **muitas variáveis**, pois não é possível examinar as múltiplas correlações entre elas de forma simples.

	NADADOR	ALTURA	PESO	Comp_1	Comp_2
65	66	195.8	92.3	3.396740	0.761430
78	79	198.5	84.6	2.711883	-0.346027
64	65	190.0	89.5	2.616959	0.889460
72	73	194.7	86.3	2.612350	0.149211
92	93	195.6	84.5	2.473292	-0.130713
18	19	193.6	83.8	2.235304	-0.055663
87	88	189.7	84.1	1.965008	0.284464
42	43	189.8	83.2	1.868088	0.171892
41	42	188.1	83.9	1.816516	0.386402
28	29	195.4	78.6	1.770974	-0.801728

Arquivo: Antropometria.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

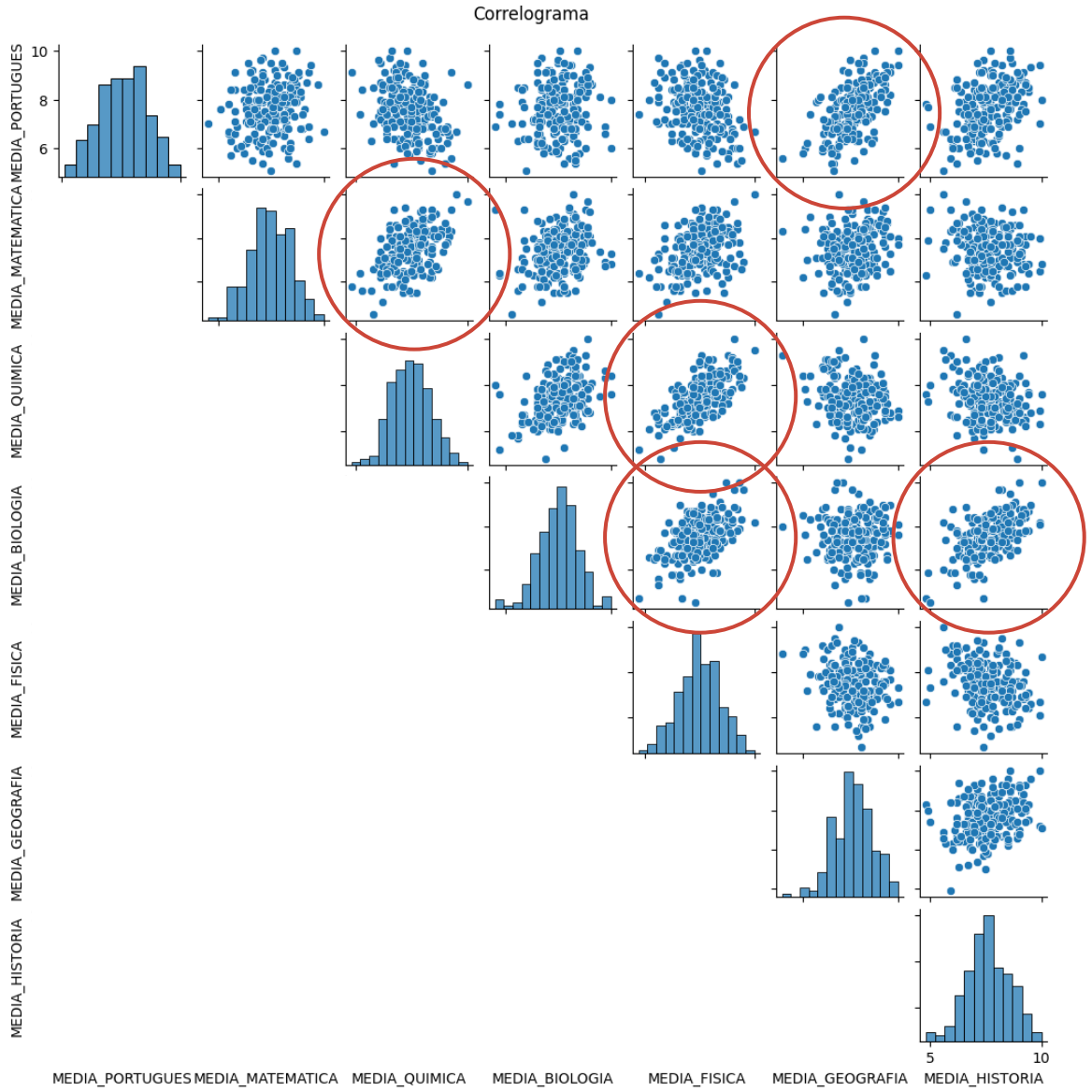
ALUNO	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA
001	7,3	8,2	8,3	8,2	8,8	8,5	8,0
002	6,3	6,8	7,8	6,5	6,2	7,8	7,0
003	10	8,7	7,3	8,4	7,8	8,8	7,7
004	7,8	7,9	7,7	6,9	7,2	8,4	7,4
005	6,0	8,4	7,8	7,6	9,0	5,1	6,3
006	8,8	7,2	7,0	8,7	6,9	8,5	9,1
007	6,3	5,6	5,7	5,6	6,6	7,2	7,0
008	7,8	6,8	6,6	7,7	6,8	8,4	9,2
009	6,1	6,9	7,8	6,5	7,9	6,3	6,0
010	5,8	6,2	8,4	8,5	8,8	5,0	7,5
...	...	...	...	...	...	...	...

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Alguns pares de variáveis aparentam ter correlações **moderadas** entre si.



Arquivo: Desempenho_Pre_Vestibular.txt

Case: Desempenho Pré-Vestibular

Alguns pares de variáveis aparentam ter correlações **moderadas** entre si.

	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA
MEDIA_PORTUGUES	1.000000	0.155060	-0.315112	0.100238	-0.282595	0.592658	0.385540
MEDIA_MATEMATICA		1.000000	0.492824	0.250311	0.405768	0.175195	-0.137465
MEDIA_QUIMICA			1.000000	0.411225	0.683395	-0.208881	-0.234284
MEDIA_BIOLOGIA				1.000000	0.518281	0.013473	0.565787
MEDIA_FISICA					1.000000	-0.168895	-0.272502
MEDIA_GEOGRAFIA						1.000000	0.290116
MEDIA_HISTORIA							1.000000

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

Obtenção dos componentes

Função ***fit_transform*** do pacote **sklearn** do Python

	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA
0	-0.320215	0.319906	0.549129	0.269683	0.544302	-0.261484	-0.237091
1	-0.456814	-0.311665	-0.109316	-0.508985	-0.133438	-0.421452	-0.477224
2	0.289421	0.511947	0.062369	-0.454857	0.009754	0.421417	-0.515436
3	0.136077	0.589912	0.014235	-0.030529	-0.477727	-0.608321	0.184591
4	0.659039	-0.179799	-0.294032	0.126616	0.385539	-0.435331	-0.304467
5	-0.382999	0.396673	-0.768663	0.118091	0.296067	0.051501	0.029419
6	-0.074037	-0.027608	-0.071529	0.656050	-0.470441	0.119131	-0.568095

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

Obtenção dos componentes

Função ***fit_transform*** do pacote **sklearn** do Python

	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA
0	-0.320215	0.319906	0.549129	0.269683	0.544302	-0.261484	-0.237091
1	-0.456814	-0.311665	-0.109316	-0.508985	-0.133438	-0.421452	-0.477224
2	0.289421	0.511947	0.062369	-0.454857	0.009754	0.421417	-0.515436
3	0.136077	0.589912	0.014235	-0.030529	-0.477727	-0.608321	0.184591
4	0.659039	-0.179799	-0.294032	0.126616	0.385539	-0.435331	-0.304467
5	-0.382999	0.396673	-0.768663	0.118091	0.296067	0.051501	0.029419
6	-0.074037	-0.027608	-0.071529	0.656050	-0.470441	0.119131	-0.568095

PRIMEIRO COMPONENTE PRINCIPAL

- Ajuda a segregar alunos fortes (ou fracos) em disciplinas de **humanidades** e **linguagem** (sinal negativo) *versus* disciplinas de **ciências exatas** e **naturais** (sinal positivo).
- **Química** e **Física** exercem os maiores pesos neste componente (0,549 e 0,544).

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

Obtenção dos componentes

Função ***fit_transform*** do pacote **sklearn** do Python

	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA
0	-0.320215	0.319906	0.549129	0.269683	0.544302	-0.261484	-0.237091
1	-0.456814	-0.311665	-0.109316	-0.508985	-0.133438	-0.421452	-0.477224
2	0.289421	0.511947	0.062369	-0.454857	0.009754	0.421417	-0.515436
3	0.136077	0.589912	0.014235	-0.030529	-0.477727	-0.608321	0.184591
4	0.659039	-0.179799	-0.294032	0.126616	0.385539	-0.435331	-0.304467
5	-0.382999	0.396673	-0.768663	0.118091	0.296067	0.051501	0.029419
6	-0.074037	-0.027608	-0.071529	0.656050	-0.470441	0.119131	-0.568095

SEGUNDO COMPONENTE PRINCIPAL

- Ajuda a segregar alunos **mais fortes** do que a média (ou **menos fortes** do que a média) em todas as disciplinas.
- Em particular, disciplinas que exigem maior compreensão textual ou contextualização exercem maiores pesos, nesta ordem: **Biologia**, **História**, **Português** e **Geografia**.

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

Obtenção dos componentes

Função ***fit_transform*** do pacote **sklearn** do Python

	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA
0	-0.320215	0.319906	0.549129	0.269683	0.544302	-0.261484	-0.237091
1	-0.456814	-0.311665	-0.109316	-0.508985	-0.133438	-0.421452	-0.477224
2	0.289421	0.511947	0.062369	-0.454857	0.009754	0.421417	-0.515436
3	0.136077	0.589912	0.014235	-0.030529	-0.477727	-0.608321	0.184591
4	0.659039	-0.179799	-0.294032	0.126616	0.385539	-0.435331	-0.304467
5	-0.382999	0.396673	-0.768663	0.118091	0.296067	0.051501	0.029419
6	-0.074037	-0.027608	-0.071529	0.656050	-0.470441	0.119131	-0.568095

TERCEIRO COMPONENTE PRINCIPAL

- Ajuda a segregar alunos fortes (ou fracos) em **História** e **Biologia** juntas (sinal negativo), bem como em **Matemática** e **Geografia** juntas (sinal positivo).
- É muito pouco influenciado pelo desempenho dos alunos em **Química** e **Física** (~0,06 e ~0,01).

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

Variabilidade retida pelos componentes

Atributo ***explained_variance_ratio_*** da classe **PCA** do pacote **sklearn** do Python

	Var. Retida
0	0.364403
1	0.287010
2	0.173664
3	0.068918
4	0.054661
5	0.040747
6	0.010597

- Os três primeiros componentes concentram cerca de **82%** do comportamento das sete variáveis originais.

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

Após aplicar os novos componentes na base, destaca-se os 5 alunos com **menor *score*** e os 5 alunos com **maior *score*** no **1º** componente principal, que ajuda a segregar alunos fortes em humanas *versus* exatas/biológicas.

ALUNO	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6	Comp_7
149	7.4	6.4	5.2	4.7	4.7	8.1	7.4	-4.163354	2.544013	0.874335	0.484623	-0.852915	0.373330	-0.199464
183	9.1	5.8	4.8	6.9	5.7	7.4	8.9	-4.112649	0.251750	-1.129524	0.462632	1.020733	0.448768	-0.186856
17	8.5	6.6	5.8	5.6	5.6	8.5	7.4	-3.387168	1.134416	1.070063	0.035163	-0.067143	-0.064511	-0.106711
77	9.7	7.1	6.5	6.7	5.6	8.2	8.6	-3.055430	-0.704867	0.428988	0.878291	0.277685	-0.762701	-0.227467
27	8.0	6.2	6.1	6.6	5.2	7.9	7.7	-3.036951	1.062902	-0.155715	0.318114	-0.279926	-0.325645	0.571309

ALUNO	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6	Comp_7
67	6.7	10.0	9.5	8.2	10.0	6.9	5.6	4.751301	-0.108406	1.682989	0.212427	0.311107	0.319690	0.102473
78	5.6	8.3	9.3	8.0	8.8	3.9	5.9	4.357322	2.324608	-0.850805	1.502634	0.659135	-0.332803	0.180395
117	6.0	8.4	8.8	8.8	9.2	6.0	6.0	3.849389	0.766841	-0.287650	0.141072	0.406492	0.333798	0.737358
152	5.4	8.7	9.0	7.8	9.2	6.5	6.2	3.795204	1.150484	0.302283	0.017662	-0.511953	0.409171	-0.001855
139	8.6	9.7	10.0	8.9	9.1	8.2	6.6	3.478273	-2.198575	1.758012	0.135797	0.273446	-1.047867	0.444301

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data

Case: Desempenho Pré-Vestibular

Um colégio de cursos preparatórios para o vestibular *(cursinho)* deseja compreender quais foram os principais pilares de desempenho dos alunos no ano anterior, com base em suas notas médias em sete disciplinas distintas.

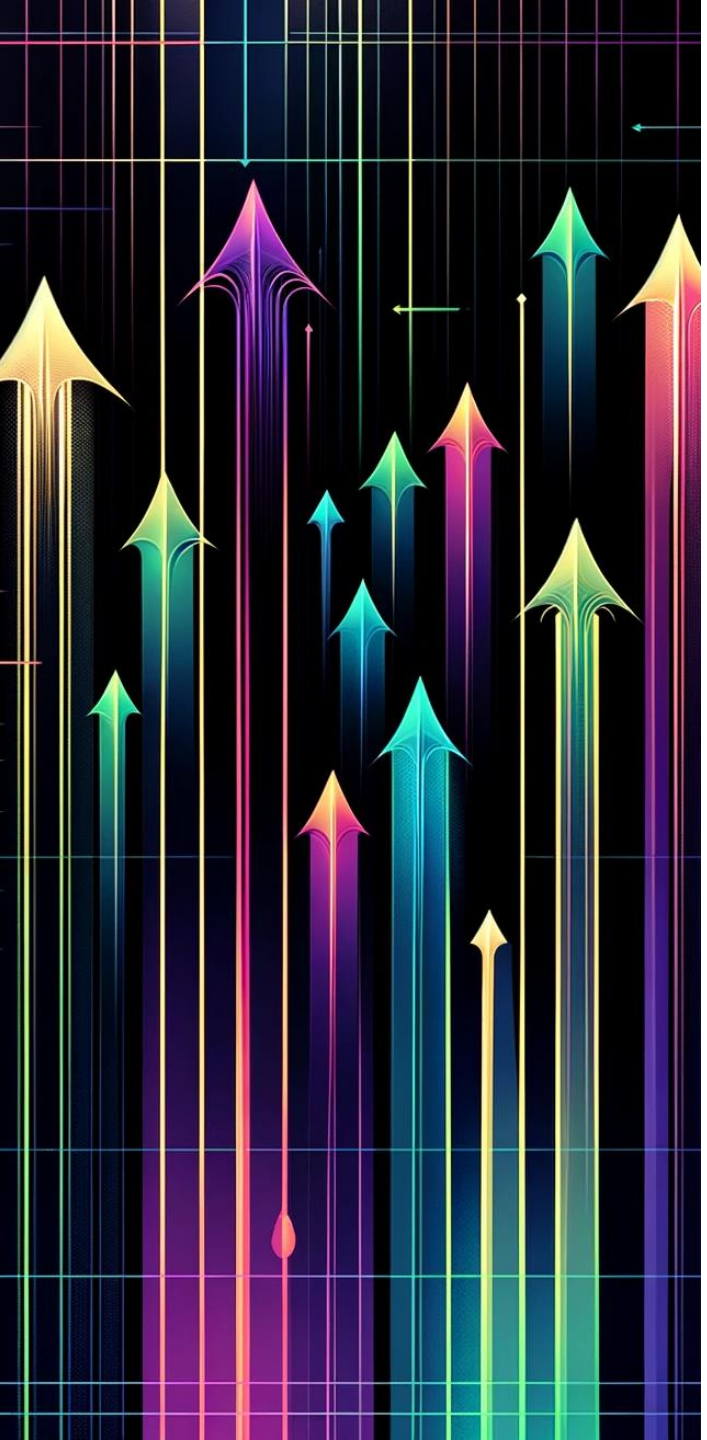
Examinemos, agora, os 5 alunos com **menor *score*** e os 5 alunos com **maior *score*** no **2º** componente principal, que ajuda a segregar alunos bons em todas as disciplinas, com maior ênfase em Biologia, História, Português e Geografia.

ALUNO	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6	Comp_7
104	9.1	9.3	9.3	8.8	7.6	8.8	9.2	1.070396	-3.457009	0.548810	0.845867	-0.792926	-1.158626	-0.284420
91	9.3	8.1	6.8	9.0	6.6	9.6	9.5	-1.752717	-3.295071	-0.137210	0.184915	-0.423154	0.232463	0.498776
31	9.4	8.0	6.9	8.1	6.7	10.0	9.9	-2.154982	-3.206697	0.238067	-0.045993	-0.750706	0.017179	-0.379373
112	8.7	9.2	7.9	9.0	7.6	9.2	8.7	0.360419	-3.102836	0.616582	0.369353	-0.549284	0.208627	0.339111
196	8.9	8.5	6.9	8.9	8.5	9.3	8.4	-0.019353	-2.811367	0.491614	-0.614434	0.449751	0.993598	0.092213

ALUNO	MEDIA_PORTUGUES	MEDIA_MATEMATICA	MEDIA_QUIMICA	MEDIA_BIOLOGIA	MEDIA_FISICA	MEDIA_GEOGRAFIA	MEDIA_HISTORIA	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6	Comp_7
34	7.0	4.5	5.8	5.3	5.6	6.7	6.2	-2.941454	3.967299	-0.422393	-0.574042	0.463198	-0.525568	0.337614
200	5.1	7.5	8.4	6.0	7.6	6.6	5.6	1.794748	3.198262	0.759755	-0.059869	-1.036267	-0.184293	0.003826
7	6.3	5.6	5.7	5.6	6.6	7.2	7.0	-2.082936	3.036201	-0.398801	-0.668332	-0.167653	0.672470	-0.331774
51	7.8	6.3	7.6	4.7	6.7	8.3	4.8	-1.079185	3.002117	2.580582	-1.110708	0.160610	-1.361064	0.113473
127	6.2	7.0	7.3	6.5	7.2	5.2	6.7	0.604072	2.835147	-0.679016	0.997624	0.290341	0.067208	-0.225023

Arquivo: Desempenho_Pre_Vestibular.txt

FIA BUSINESS SCHOOL lab.data



Benefícios para a Modelagem

Quando a redução de dimensionalidade é utilizada como etapa preliminar para a construção de um modelo, ela pode contribuir de forma relevante para a sua simplificação. Vide exemplo, a seguir.

Case: Aparelhos Celulares

Uma loja de eletrônicos adquiriu 55 modelos de aparelhos de celular para revender aos consumidores. O seu interesse está em compreender quais especificações técnicas desses produtos influenciam o seu preço de custo, ou seja, o preço cobrado pelos fabricantes.

Variável	Descrição
ID_MODELO	Código identificador do modelo de aparelho
TAMANHO_TELA	Tamanho da tela, em polegadas
RESOLUCAO_ALTURA	Resolução da altura do aparelho, em pixels
RESOLUCAO_LARGURA	Resolução da largura do aparelho, em pixels
MEMORIA_RAM	Memória RAM do aparelho, em gigabytes (GB)
ARMAZENAMENTO	Capacidade de armazenamento do aparelho, em gigabytes (GB)
BATERIA	Capacidade de bateria do aparelho, em miliampère-hora (mAh)
RESOLUCAO_CAM_PRINCIPAL	Resolução da câmera principal, em megapixels (MP)
RESOLUCAO_CAM_FRONTAL	Resolução da câmera frontal, em megapixels (MP)
PESO_INVERSO	Inverso do peso do aparelho, em gramas^{-1} (1/g)
PROCESSADOR	Frequência do processador do aparelho, em gigahertz (GHz)
NUCLEOS	Quantidade de núcleos do processador do aparelho
PRECO	Preço de custo do produto, em reais

Arquivo: Aparelhos_Celulares.txt

FIA BUSINESS SCHOOL lab.data

Case: Aparelhos Celulares

Uma loja de eletrônicos adquiriu 55 modelos de aparelhos de celular para revender aos consumidores. O seu interesse está em compreender quais especificações técnicas desses produtos influenciam o seu preço de custo, ou seja, o preço cobrado pelos fabricantes.

a) Faça uma breve análise exploratória da base de dados.
b) Realize uma análise bidimensional das correlações entre os pares de variáveis. Quais as principais conclusões?
c) Padronize as variáveis da base de dados, utilizando o método *z*-score.
d) Obtenha os componentes principais. Quantos componentes são necessários para reter 70%, 80% ou 90% do comportamento das variáveis originais?
e) Descreva os quatro primeiros componentes principais no que diz respeito às variáveis que mais os influenciam, bem como o sentido dessa influência.
f) Construa um modelo de regressão logística para predizer o preço de custo dos aparelhos a partir de suas especificações, utilizando as variáveis originais. Avalie a presença de colinearidade a partir da estatística VIF, bem como o grau de variabilidade explicada por meio do R^2 ajustado.
g) Construa um novo modelo de regressão logística, agora considerando como variáveis explicativas os dois primeiros componentes principais. Compare o VIF e o R^2 ajustado deste novo modelo com os obtidos no modelo do item (f).

Arquivo: Aparelhos_Celulares.txt

Referências Bibliográficas

FRAMEWORK GERAL DE MODELAGEM

- James, G., Witten, D., Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in Python.* Springer, 2023.
- Kuhn, M., Johnson, K. *Applied Predictive Modeling*. Springer, 2013.
- Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

lab.data

http://labdata.fia.com.br
Instagram: @labdatafia
Facebook: @LabdataFIA