Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Overview

Framework de Análise de Dados

Overview

Preditiva.ai

Para que essa “máquina” chamada de **Análise de Dados** funcione, **precisamos de pelo menos três “engrenagens”**. São elas:

- ✓ Matemática fundamental
- ✓ Técnicas de Otimização
- ✓ Estatística Descritiva
- ✓ Probabilidades
- ✓ Inferência (ex: Testes de Hipótese...)
- ✓ Modelagem Estatística
- ✓ *Machine Learning*
- ✓ Entre outros...

- ✓ Planilhas e Tipos de Arquivos
- ✓ Linguagens de programação
- ✓ Bancos de dados
- ✓ Visualização de Dados
- ✓ Infra estrutura (DWs, Data Lakes etc)
- ✓ ETL / Pipelines de Dados
- ✓ Entre outros...

- ✓ Entendimento dos processos internos
- ✓ Alinhamento com os objetivos da empresa
- ✓ Entrosamento com o time e principais *stakeholders*
- ✓ Análise da concorrência
- ✓ Entre outros...

Framework de Análise de Dados

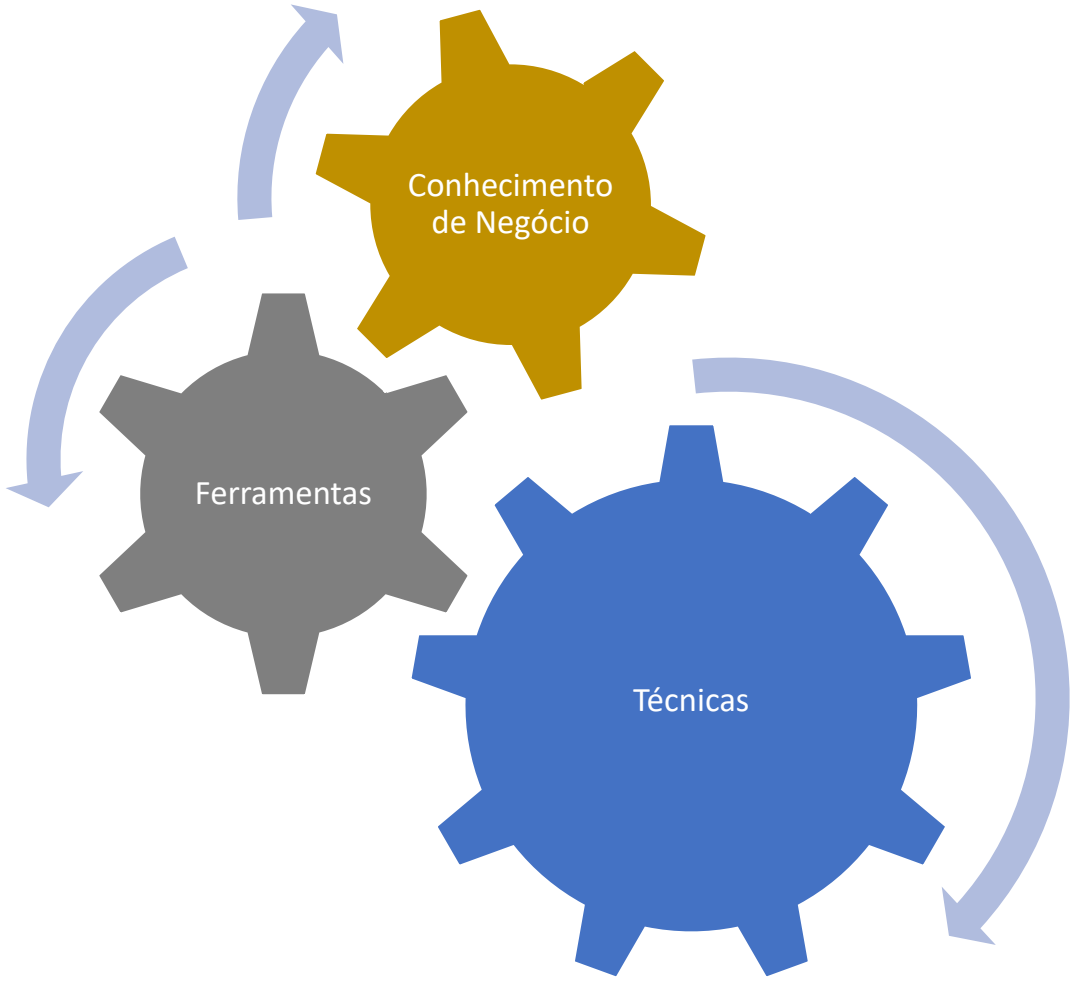
Overview

Preditiva.ai

No entanto, essas “engrenagens” funcionando sozinhas **não** entregam o real valor para as áreas das empresas como deveriam.

Felizmente, existe uma forma eficiente para que as engrenagens funcionem em sinergia:

Aplicando-se uma
Metodologia de Projetos de Dados.



Framework de Análise de Dados

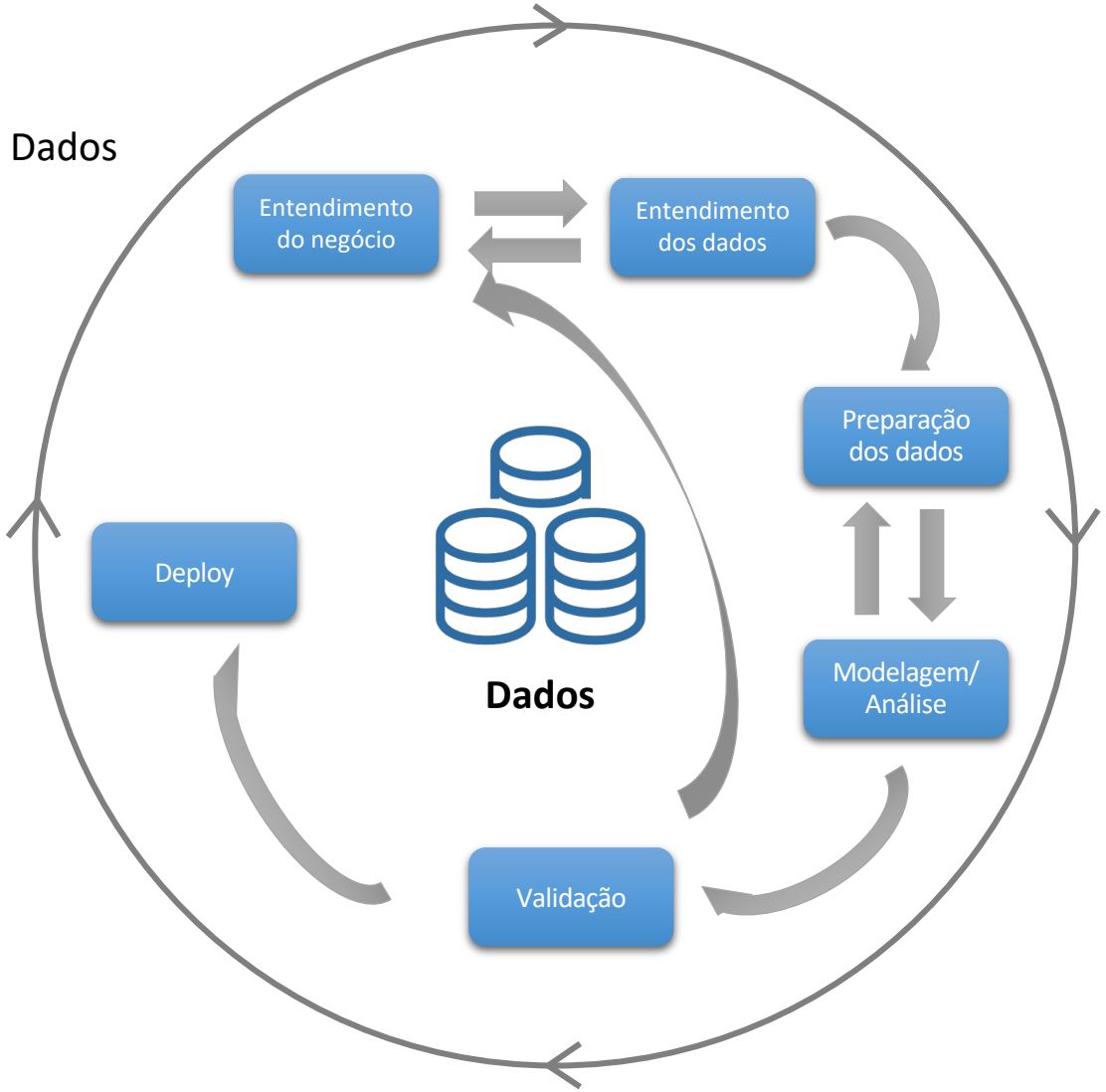
A metodologia CRISP-DM

Preditiva.ai

Uma metodologia muito utilizada por profissionais de Dados é a chamada **CRISP-DM**. Sua abreviação vem de ***Cross Industry Standard Process for Data Mining***.

A metodologia foi criada unindo-se as melhores práticas no processo de análise de dados.

O **CRISP-DM tem 6 grandes passos** conforme mostrado na imagem ao lado. Veremos com mais detalhes ao longo das próximas aulas.



A metodologia CRISP-DM

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 1: Entendimento do Negócio

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ **Objetivo**

O que deve ser feito ?

Antes de começar qualquer projeto de dados temos que ter muita clareza do que deve ser resolvido. Queremos aumentar as vendas? Descobrir as causa do Churn de clientes? Entender que tipo de cliente é mais propenso à compra? O objetivo do trabalho deve ser muito claro!

Como realizar ?

Muitos analistas somente entendem o objetivo geral e já saem construindo *queries* e criando *dashboards*. Grande erro! Nesta etapa, sempre tenha as respostas para seguintes perguntas:

- ❑ Qual o objetivo deste trabalho?
- ❑ O que queremos conhecer? O que queremos mudar na área com esse projeto?
- ❑ Já existe algo realizado ou em andamento feito por alguém? Quais os resultados?
- ❑ Supondo que consigamos atingir o objetivo, o que vem depois? Como este trabalho será usado?

Dicas da Preditiva:

- Faça **várias perguntas sobre o negócio para seu cliente**. Marque quantas reuniões forem necessárias, mas o entendimento do processo é fundamental antes de começar.
- Sempre **ajude o seu cliente a priorizar os projetos**. Isso envolve questioná-lo sobre o potencial de resultado e de como esse resultado será usado na prática.
- Se o uso do resultado de seu **projeto de análise de dados não estiver claro**, **é uma boa ideia abandoná-lo** e deixar no roadmap para uma próxima oportunidade.

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ Objetivo
- ❑ **Premissas**

O que deve ser feito ?

Uma premissa é a assunção de alguma verdade. Exemplo: Vamos assumir que essa amostra extraída dos dados seja aleatória e que não temos vieses de seleção.

Nesta etapa precisamos pensar em todas as premissas que seu trabalho irá assumir.

Como realizar ?

Na entrevista com seu cliente já se pode entender algumas delas. Outras premissas são descobertas na etapa de Entendimento dos Dados, pois nem sempre o que o cliente conhece é o que de fato acontece nos dados extraídos para análise.

Dicas da Preditiva:

- Após a etapa de Análise Exploratória dos Dados muitas vezes temos mais clareza das premissas que devem ser assumidas. Sempre **volte para a sua "Lista de Premissas" a medida que vai conhecendo melhor sua base** de dados.
- As **premissas devem ser informadas para seu cliente desde o início**. Isso é muito importante para que o cliente tenha clareza se o resultado da análise será realmente útil para ele. Infelizmente, vários trabalhos de análise são "invalidados" em uma reunião com o gestor ou cliente da análise pelo simples fato da premissa não ter sido informada antes do projeto iniciar. Exemplo: "Nossa, eu não sabia que você tinha pego dados do sistema XPTO. Esse sistema está cheio de problemas. Como vamos confiar em todas as análises que fez?"

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ Objetivo
- ❑ Premissas
- ❑ **Riscos envolvidos**

O que deve ser feito ?

Todo processo tem risco. Não devemos evitar o risco, mas sim controla-lo. Portanto, nesta etapa precisamos ter clareza de quais riscos nosso projeto de dados está exposto e de como mitigá-lo (diminuí-lo), se possível.

Como realizar ?

Ao entender bem as premissas e objetivos, os riscos tendem a aparecer naturalmente. Exemplo de riscos comuns em projetos de dados:

- Os dados da análise não estão estruturados em um repositório de dados validado, como o DW ou Data Lake. Desta forma corremos o risco de juntar os dados de forma incorreta devido ao trabalho manual. Além disso, o trabalho levará mais tempo.
- As áreas de negócio não têm familiaridade com interpretação de dados ou uso de ferramentas analíticas, fazendo com que o resultado do projeto de análise possa ser mal utilizado, produzindo resultados ruins.
- Conformidade com a **LGPD**: Todos nossos projetos de dados devem estar em conformidade com os requisitos da lei. Sempre verifique isso com as áreas de Governança e Privacidade da empresa.
- O sistema ou infraestrutura de dados anda congestionada. Com isso a coleta dos dados pode demorar mais do que o previsto.

Dicas da Preditiva:

- Para entender bem os riscos envolvidos é interessante realizar uma **análise SWOT*** do processo. Uma das etapas deste tipo de análise é justamente pensar riscos internos e externos que podem ser pontos de atenção para o projeto de análise de dados.

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ Objetivo
- ❑ Premissas
- ❑ Riscos envolvidos
- ❑ **Custo x Benefício**

O que deve ser feito ?

Todo projeto tem um custo. Nem que esse custo seja o seu tempo ou da equipe. Nesta etapa da metodologia devemos ter o máximo de clareza de qual o custo que estamos lidando. Porém, o custo muitas vezes é um investimento de um benefício muito maior. Portanto, devemos levantar os custos e benefícios esperados do projeto de dados para avaliar se vale a pena continuar ou se deixamos esse projeto para uma próxima oportunidade.

Como realizar ?

Para alcançar o objetivo do projeto precisamos coletar os dados necessários. Desta forma, algumas perguntas são bem naturais nesta etapa:

- Onde esses dados estão? Estão disponíveis e atualizados?
- Se não, qual o custo para começar a coletá-los?
- São dados que precisam ser adquiridos de um fornecedor externo? Se sim, qual o custo?
- Qual o benefício esperado de seu uso?

Dicas da Preditiva:

- Construa uma planilha para documentar essa pesquisa. Depois priorize os dados de acordo com a avaliação de custo x benefício.

Dados	Custo Tangível	Custo Intangível	Benefício
Dívidas no mercado	12 centavos por CPF	Construção de API para pegar os dados	Melhoria considerável do modelo de crédito
Uso do aplicativo	Nenhum, dado próprio	Construção de ETL para carregamento	Não muito claro. É preciso realizar uma análise exploratória.
...	...	...	...

Framework de Análise de Dados

Análise de Custo vs Benefício

Preditiva.ai

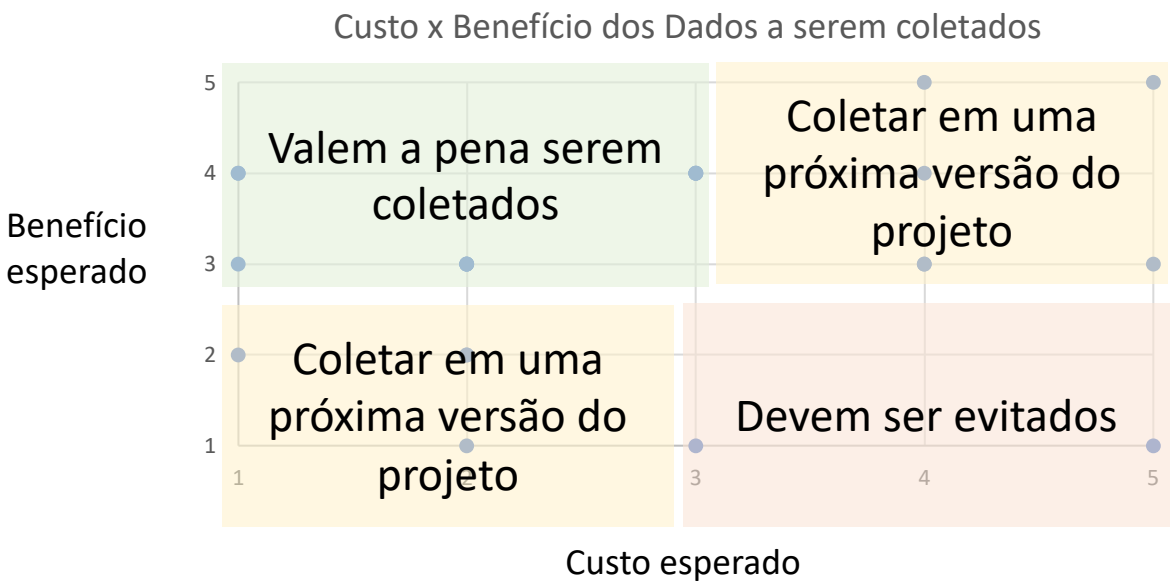
Etapa 1: Criação de planilha com o levantamento dos dados disponíveis.

Dados	Custo Tangível	Custo Intangível	Benefício	Nota para Custo (1 a 5)	Nota para benefício (1 a 5)
Dívidas no mercado	12 centavos por CPF	Construção de API para pegar os dados	Melhoria considerável do modelo de crédito	4	5
Uso do aplicativo	Nenhum, dado próprio	Construção de ETL para carregamento	Não muito claro. É preciso realizar uma análise exploratória.	3	2
...	...	...	...	...	...

Dicas:

Para atribuir a nota sempre verifique com as áreas clientes do projeto. São eles que conhecem o negócio ou processo melhor do que você.

Etapa 2: Criação de uma matriz de custo vs benefício.



Dicas:

Lembre-se que o objetivo de um projeto de dados é maximizar o resultado com o menor custo.

Portanto, deixe para depois a análise dos dados que fogem dessa regra.

Se após essa análise você tiver poucos dados no quadrante verde, é importante decidir se o projeto vale o esforço do time.

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ Objetivo
- ❑ Premissas
- ❑ Riscos envolvidos
- ❑ Custo x Benefício
- ❑ **Critérios de sucesso**

O que deve ser feito ?

Muitas vezes quando analisamos uma base dados temos a impressão que podemos investigar infinitas possibilidades. Isso é verdade! Por isso precisamos ser objetivos e focar no que realmente importa. Segundo o CRISP-DM, a melhor forma de fazer isso é definir um claro **Critério de Sucesso** para o projeto. Ele funciona como um tipo de "critério de parada". Ou seja, quando chegarmos a esse objetivo podemos estar satisfeitos com o projeto e encerrá-lo.

Como realizar ?

Bons critérios de sucesso são criados levando em consideração as métricas do negócio ou as próprias métricas obtidas do modelo estatístico desenvolvido.

Exemplos de critérios de sucesso:

- Diminuição esperada de 10% na taxa de churn de clientes **(Métrica de Negócio);**
- Melhoria do KS do modelo de crédito para um patamar de pelo menos 30% **(Métrica de Modelos);**
- Mitigar todos os riscos levantados pelo time de Controles Internos **(Métrica Regulatória);**

Dicas da Preditiva:

- Busque referências nas áreas de negócio de bons indicadores e/ou trabalhos passados. O que funcionou bem e o que não funcionou? Se o trabalho anterior melhorou 5% do resultado, talvez uma expectativa de aumento de 50% no seu projeto seja irrealista.

- Se o seu projeto tem como objetivo melhorar um produto de dados anterior (ex: um modelo estatístico que ruim), leia a documentação (se existir) e verifique o que funcionou anteriormente que vale a pena continuar no seu projeto.

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ Objetivo
- ❑ Premissas
- ❑ Riscos envolvidos
- ❑ Custo x Benefício
- ❑ Critérios de sucesso
- ❑ **Planejamento do projeto**

O que deve ser feito ?

Após coletar todas essas informações você precisar criar um cronograma contendo cada etapa do projeto. O CRISP-DM tem 5 etapas adicionais após a etapa de conhecimento do negócio.

Como realizar ?

O cronograma deve conter os tempos em dias estimados para cada etapa e tarefa do projeto.
Insira também as reuniões esperadas com o cliente para reporte do andamento do projeto.

Dicas da Preditiva:

- O cronograma deve ser aprovado pelo cliente**. Evite que essa aprovação seja feita de forma verbal**. Peça para o cliente confirmar por e-mail e só inicie as próximas etapas após o “de acordo” formal do cliente.
- Marque as **reuniões com o cliente logo após a aprovação do cronograma**, assim a agenda já fica bloqueada. O ideal é realizar uma reunião com o cliente semanalmente. Se não for possível, agende ao menos quinzenalmente.
- O tempo de conclusão de projetos depente de fatores como: Facilidade de acesso aos dados, disponibilidade do cliente, critérios de sucesso claros, tipo de risco envolvido (ex: indisponibilidade sistêmica), sofisticação da técnica de estatística utilizada, entre outros. Portanto**, sempre dê uma estimativa otimista** (supondo que tudo funcione) **e uma pessimista** (supondo o pior cenário).

Framework de Análise de Dados

Exemplo de cronograma de projeto

Preditiva.ai

Etapa	Tarefa	Responsável	Estimativa em dias		Mês 1				Mês 2				Mês 3	
			Pessimista	Otimista	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Entendimento do negócio	Objetivo, premissas e demais análises iniciais	Analista de Dados	10	7										
	Aprovação do projeto	Cliente	2	1										
Entendimento dos dados	Coleta dos dados	Analista de Dados	10	4										
Entendimento dos dados	Análise exploratória inicial e sanity check	Analista de Dados	14	10										
Preparação dos dados	Ajustes da Base (padronização e limpeza)	Analista de Dados	7	5										
Modelagem	Desenvolvimento do estudo ou modelo	Analista de Dados	14	10										
Validação	Apresentação do trabalho e Aprovação	Analista / Cliente	2	1										
Deploy	Implantação e acompanhamento	Analista de Dados	10	6										
	Finalização da documentação	Analista de Dados	5	3										
	Total em dias		74	47										
	Total em meses		**3,2**	**2,0**										

Legenda

- Concluído
- Em andamento
- A iniciar
- Reuniões de reporte

Dicas:

- As três etapas **inicias geralmente são as mais demoradas** (cerca de 60% do projeto). Não sub-estime o prazo dessas etapas.
- O cronograma acima foi planejado considerando os prazos pessimistas. Fica a seu critério criar outra versão com os prazos otimistas.
- **Sempre informe seu cliente de desvios em relação aos prazos combinados**. O que pega mal não é o atraso, mas a falta de aviso para o cliente que está esperando uma informação. **Não fique com receito em dar noticiais ruins!**

Framework de Análise de Dados

Etapa 1: Entendimento do negócio

Preditiva.ai

Entendimento do negócio

- ❑ Objetivo
- ❑ Premissas
- ❑ Riscos envolvidos
- ❑ Custo x Benefício
- ❑ Critérios de sucesso
- ❑ Planejamento do projeto
- ❑ **Início da documentação**

O que deve ser feito ?

Um projeto sem documentação tem grandes chances de não durar por muito tempo. O motivo é que a documentação é essencial para continuidade em caso de mudanças nos times e escopo das áreas de negócio. Imagine você ter que fazer uma nova versão de um Dashboard. Como saberá como o atual funciona sem documentação? Nesta etapa devemos consolidar todo o conhecimento obtido em um documento de fácil acesso para você, time e cliente do projeto.

Como realizar ?

Uma documentação envolve clareza na escrita e organização de seus tópicos. Seguem abaixo os principais tópicos a serem descritos em um documento de projetos de dados:

- Todos os critérios e conhecimentos obtidos na etapa de “Entendimento de Negócio”;
  - Incluindo as premissas, riscos mapeados, custos x benefícios e critérios de sucesso.
- Cronograma do projeto;
- Análise exploratória e transformações de dados realizadas na base (filtros, correções, padronizações, tratamento de valores faltantes etc);
- Estudo/modelo desenvolvido e suas conclusões;
- Plano de implantação e acompanhamento;

Dicas da Preditiva:

- A documentação é parte das melhores práticas da metodologia de Gestão do conhecimento. Vale a pena conhecer mais.
- Crie um glossário dos termos de negócio e inclua na documentação.
- Sempre mantenha um atualizado histórico de versões do documento.
- Sempre documente as aprovações e alinhamentos com o cliente (guarde as ATA’s de reunião na pasta do projeto). Acredite, isso pode te evitar muita dor de cabeça!

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 2: Entendimento dos Dados

Framework de Análise de Dados

Etapa 2: Entendimento dos dados

Preditiva.ai

Entendimento dos dados

- ❑ **Descrição dos dados e coleta**

O que deve ser feito ?

Chegamos na etapa da busca dos dados. Nesta hora precisamos ter acesso ao dicionário de dados (a.k.a. Metadados) do repositório de dados de sua empresa. A partir daí, utilizamos a ferramenta de manipulação de dados disponível, que em geral é o SQL, para a devida consulta, manipulação e extração dos dados necessários para sua análise.

Como realizar ?

Para ter acesso ao dicionário de dados, fale com o time de TI (Analistas de Bancos de Dados, Arquitetos de Dados ou Engenheiro de Dados). É muito comum você ter acesso a documento ERD (Diagrama de Relacionamento de Entidades) do banco de dados.

Para a coleta, realize as consultas usando o SQL ou qualquer ferramenta disponível. Algumas empresas ou processos podem não ter os dados já estruturados. Neste caso, você será o responsável por estruturar. Isso é muito comum.

Dicas da Preditiva:

- **Evite fazer um SELECT * FROM de seu banco** de dados para extrair a base completa para sua ferramenta de análise (Excel, Power BI, Python etc). Em vez disso, tente extrair apenas a informação necessária, pois será mais performático desta forma.

Framework de Análise de Dados

Etapa 2: Entendimento dos dados

Preditiva.ai

Entendimento dos dados

- ❑ Descrição dos dados e coleta
- ❑ **Análise Exploratória e Sanity Check**

O que deve ser feito ?

Nesta etapa você estará aplicando técnicas de estatística descritiva para entender cada variável de sua base e suas relações entre as outras variáveis. Além disso, ao resumir os dados você tem condições de verificar sua qualidade (Sanity Check). Se algo estranho for encontrado (valores fora de um limite razoável, *outliers* etc), você sempre pode questionar as áreas que liberaram a informação para verificação.

Como realizar ?

Seguem abaixo as principais técnicas de Análise Exploratória de Dados em projetos:

- ❑ Tabelas de Frequência: Frequência Absoluta, Relativa e Acumulada.
- ❑ Visualização de Dados: Box-Plots, Histogramas, Gráficos de Linhas ou Colunas.
- ❑ Medidas Resumo: Média, Mediana, Quartis, Desvio Padrão.
- ❑ Análise Bidimensional: Correlação de Pearson, Coeficiente de Determinação e *Information Value* (IV).

Dicas da Preditiva:

- Faça a análise de forma univariada (variável a variável) primeiro e bivariada/bidimensional posteriormente. Na bivariada, **foque na variável que interessa no primeiro momento**. Evite ficar calculando cada combinação de variáveis dois a dois sem um motivo claro.
- Uma ótima forma de analisar uma variável numérica é criar um BoxPlot e Histograma. Rapidamente você já conhecerá as concentrações e possíveis outliers para investigar.
- Para Sanity Check (“Teste de Qualidade”), em variáveis numéricas, verifique a amplitude da variável calculando o mínimo e seu máximo. Para variáveis qualitativas, faça contagens para entender as categorias da variável que estão disponíveis. Muitos problemas aparecem dessa forma.

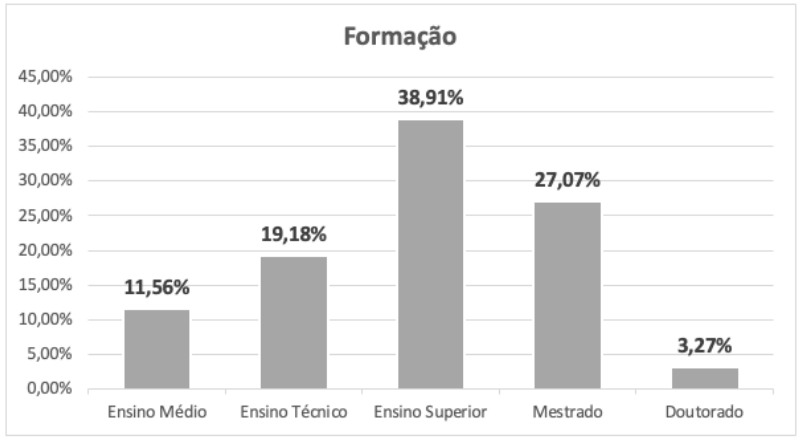
Framework de Análise de Dados

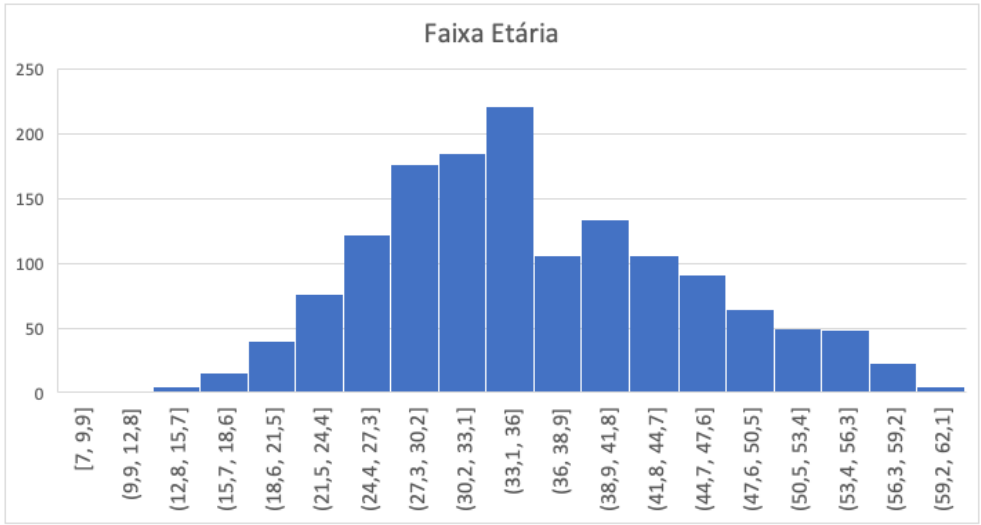
Exemplo de Análise Exploratória

Preditiva.ai

Número de Funcionários que deixaram a empresa

Deixou empresa	No Func.	% Func.
Não	1233	83,88%
Sim	237	16,12%
Total Geral	1470	100,00%

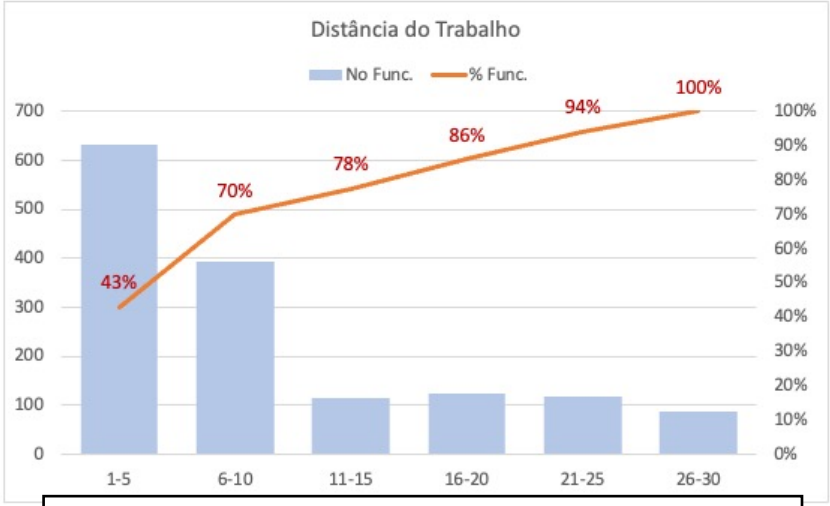




Dicas: Existem muitas formas de se mostrar a análise exploratória para seu cliente. Pode-se usar gráficos, tabelas ou um combinado dos dois. Independente da solução mostrada, é muito importante que se escolha bons gráficos, com cores com pouca saturação, e que tanto as tabelas quanto os gráficos tenham rótulos e legendas bem claras.

Sempre coloque também descritivos rápidos abaixo do gráfico destacando insights que você acredita serem relevantes.

Se a sua base tiver muitas variáveis, mostre apenas as mais importantes para o negócio (verifique com o ponto focal / cliente quais são) ou as mais importantes segundo algum critério de seleção de variáveis (como as mais relacionadas com a variável do interesse, por exemplo).



70% dos colaboradores moram até 10 km da empresa.

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 3: Preparação dos dados

Framework de Análise de Dados

Etapa 3: Preparação dos dados

Preditiva.ai

Preparação dos dados

- ❑ **Seleção das variáveis**

O que deve ser feito ?

Nem toda variável que você tem disponível no Data Lake (repositório de dados) é importante para seu problema. É comum coletarmos vários dados que pensávamos serem importantes na etapa de conhecimento de negócio que se mostraram pouco relevantes após a análise exploratória inicial. Nesta etapa fazemos uma seleção das variáveis que realmente importam para responder a pergunta de negócio.

Como realizar ?

Se o seu problema for do tipo bidimensional, quando quer entender os fatores que influenciam uma determinada variável de seu interesse (*target*), podemos aplicar técnicas de associação de variáveis para "ranquear" as variáveis mais relacionadas com o seu target. Desta forma, focamos nas variáveis que realmente vão trazer valor para o problema a ser resolvido.

Dicas da Preditiva:

- Existem técnicas simples e sofisticadas para escolher as variáveis mais importantes. Um das mais simples e eficazes é a técnica de **Information Value (IV).** Com ela, você pode produzir um ranking de IV's do maior para o menor. As variáveis com IV muito fraco são ótimas candidatas para descarte. **Veremos um exemplo a seguir.**

- Já as técnicas mais sofisticadas envolvem o cálculo do p-valor em modelos lineares como a Regressão Linear Múltipla e Logística (Ex: técnica Stepwise) ou ainda distribuições de probabilidade (ex: Algoritmo Boruta).

- Após selecionar as variáveis mais importantes, pode ser uma boa ideia já falar com o time de **Engenharia de Dados** para construir uma *query* automatizada para que você utilize esses dados futuramente quando o seu projeto for implantado. Assim você ganha tempo na etapa de implantação. Cada time de Engenharia tem seus processos e prazos. Consulte!

Framework de Análise de Dados

Exemplo de seleção de variáveis

Preditiva.ai

Para verificar quais fatores estão mais relacionados com o Turnover, vamos utilizar a técnica do **Information Value (IV).** Essa medida é responsável por mensurar o "**poder de separação/discriminação**" que uma variável possui sobre nossa variável target (no caso, o turnover).

Se uma variável tem poder forte, isso significa que uma ou mais categorias da variável tem um alto ou baixo nível de turnover, sendo útil estudá-la com mais profundidade.

Variável	IV	Grau de Discriminação do Turnover
Salário	0,42	Forte
Faz_hora_extras?	0,40	Forte
Tempo_de_empresa	0,34	Forte
Tempo_de_carreira	0,34	Forte
Idade	0,31	Forte
Anos_no_mesmo_cargo	0,27	Médio
Anos_com_o_mesmo_chefe	0,26	Médio
Estado_Civil	0,22	Médio
Frequência de Viagens	0,13	Médio
Qte_Empresas_Trabalhadas	0,10	Médio
E-Sat	0,10	Médio
Distância_do_trabalho	0,08	Fraco
Equilibrio_de_Vida	0,06	Fraco
Qte_ações_da_empresa	0,06	Fraco
Horas_de_treinamento	0,05	Fraco
Perc_de_aumento	0,05	Fraco
Anos_desde_a_ultima_promocao	0,02	Muito Fraco
Formação	0,01	Muito Fraco
Gênero	0,00	Muito Fraco

Framework de Análise de Dados

Etapa 3: Preparação dos dados

Preditiva.ai

Preparação dos dados

- ❑ Seleção das variáveis
- ❑ **Limpeza e Formatação dos dados**

O que deve ser feito ?

Agora que selecionamos as variáveis que usaremos para atingir o objetivo precisamos prepará-las para a aplicação das técnicas estatísticas.

Como realizar ?

Este processo também chamado de *Data Wrangling* ou *Data Cleaning* envolve basicamente:

- Tratamento de dados faltantes (*missing values*);
- Padronização de *case-sensitive*;
- Categorização de variáveis numéricas;
- União de tabelas (*joins*);
- *Feature engineering* (extração de caraterísticas);
- Conversão de tipos de dados;
- Entre outros.

Dicas da Preditiva:

- Um erro comum acontece quando as pessoas removem as variáveis ou observações da base que contenham dados faltantes por achar que não conseguirão extrair informações relevantes.
- **Não é aconselhável remover esses dados sem antes entender o motivo dos dados faltantes.** Algum sistema não tem uma validação de campo obrigatório no formulário? O time de Engenharia de Dados pecou em alguma etapa do pipeline? Dependendo da origem do problema, você pode ajustar facilmente extraindo os dados novamente ou ainda usando técnicas de imputação de dados.
- **Use e abuse de *Feature Engineering***. Muitas variáveis com informação relevante surgem de cálculos ou lógicas aplicadas nas outras variáveis “brutas” da base. Ex: idades de datas de nascimento, sexo do cliente através do nome, média mensal ou anual de faturamento, taxa de crescimento etc.

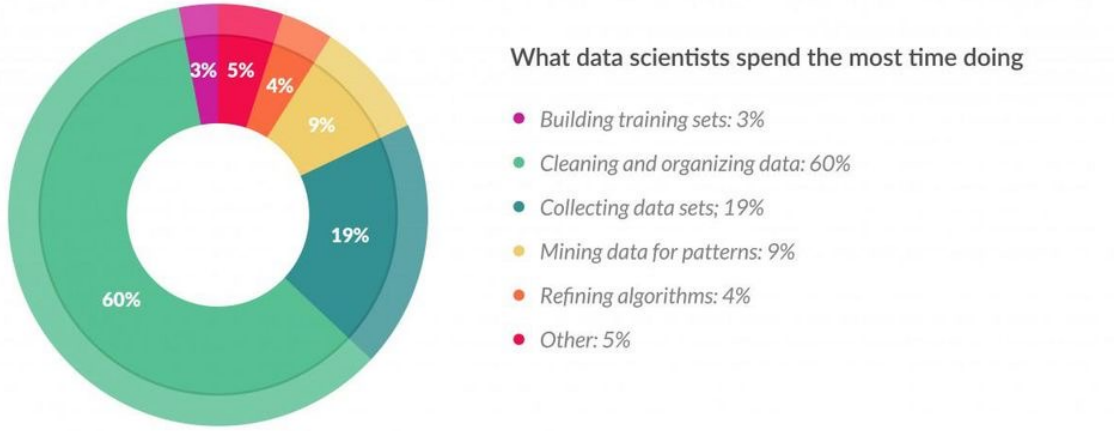
Framework de Análise de Dados

Exemplo

Preditiva.ai

Base “bruta” extraída do repositório de dados

Nome (A)	Data de Nascimento (B)	Renda (C)	Estado (D)
Sr Marcelo	20/01/1985	2000 mil	São Paulo
Sra Luana	14/06/1995	3000 mil	sao paulo
Sr João	11/01/1991	1200 mil	RJ
Sra Valquiria	23/07/1981	-	Rio de Janeiro



Fonte: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

Base “tratada” após tarefas de tratamento

Variável criada a partir de (A)	Variável criada a partir de (E)		Variável criada a partir de (B)	Variável criada a partir de (F)			Variável criada a partir de (D)
Pronome (E)	**Sexo**	**Data de Nascimento**	**Idade (F)**	**Faixa Etária**	**Renda**	**Estado**	**Estado_Ajustado**
Sr	Masculino	20/01/1985	36	33-38	R$ 2.000	São Paulo	SP
Sra	Feminino	14/06/1995	26	23-28	R$ 3.000	sao paulo	SP
Sr	Masculino	11/01/1991	30	28-33	R$ 1.200	RJ	RJ
Sra	Feminino	23/07/1981	40	38-43	**R$ 2.067**	Rio de Janeiro	RJ
Feature Engineering			Feature Engineering + Categorização		Ajuste de tipo de dados + imputação de missing		Padronização + Ajuste de Case Sensitive

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 4: Desenvolvendo o estudo ou modelo

Framework de Análise de Dados

Etapa 4: Desenvolvendo o estudo ou modelo

Preditiva.ai

Desenvolvimento

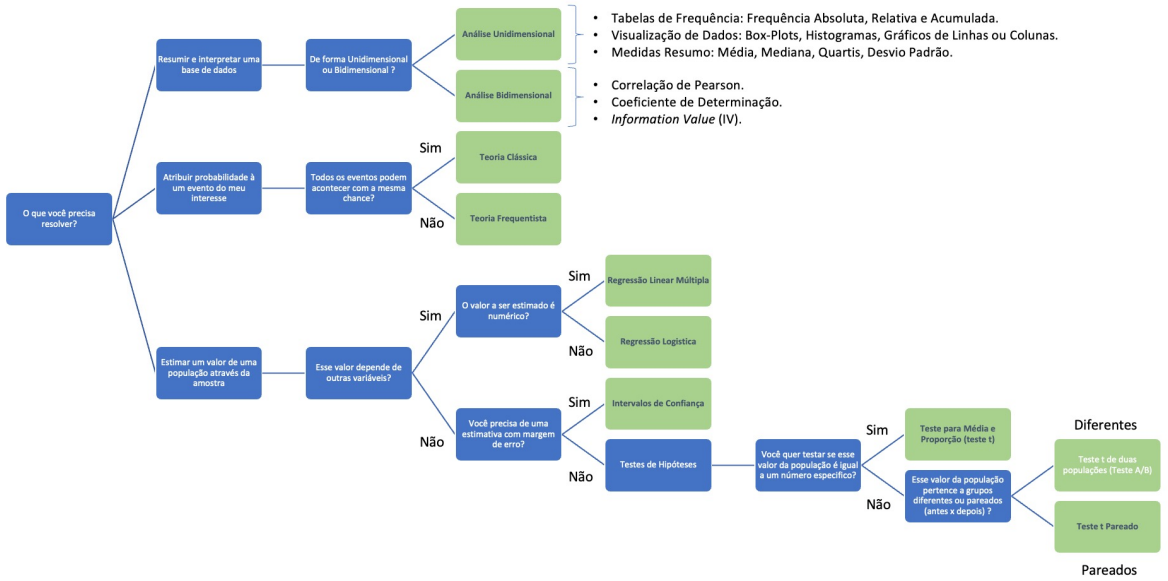
- ❑ **Escolha da técnica estatística que responde o problema**

O que deve ser feito ?

Aqui que muita gente perde a oportunidade de extrair bons *insights*. Por não ter um bom reportório de técnicas conhecidas, as pessoas tendem a utilizar uma ou duas técnicas para todo tipo de problema de negócio. É preciso ter clareza do tipo de "produto de dados" que deve ser entregue e escolher a técnica mais adequada para cada contexto.

Como realizar ?

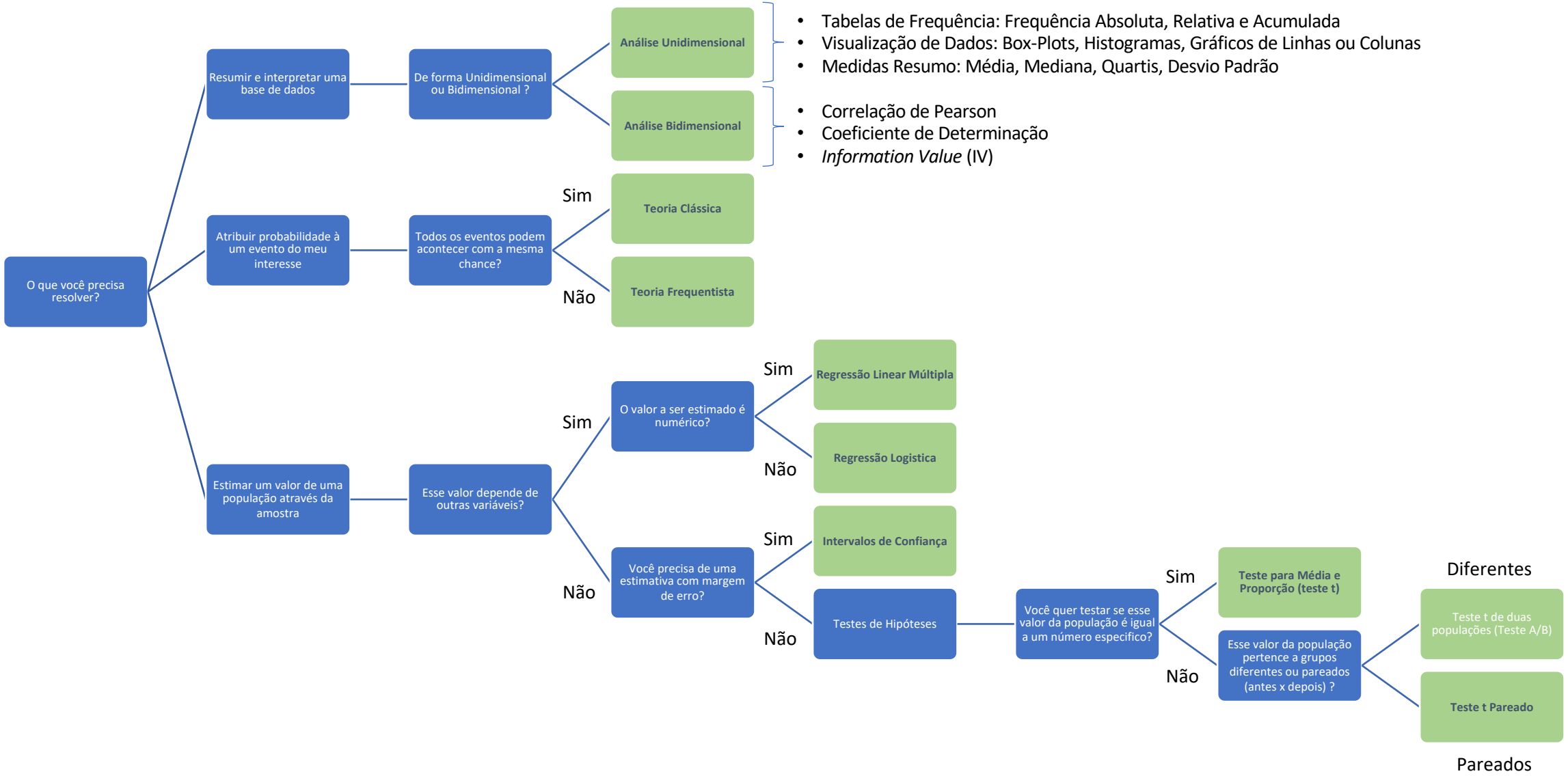
Nas páginas a seguir, fornecemos um overview das principais técnicas de estatística e de Machine Learning. A ideia é entender o tipo de problema de negócio e escolher a técnica mais indicada para cada contexto e oportunidade.



Framework de Análise de Dados

Resumo das principais técnicas estatísticas

Preditiva.ai



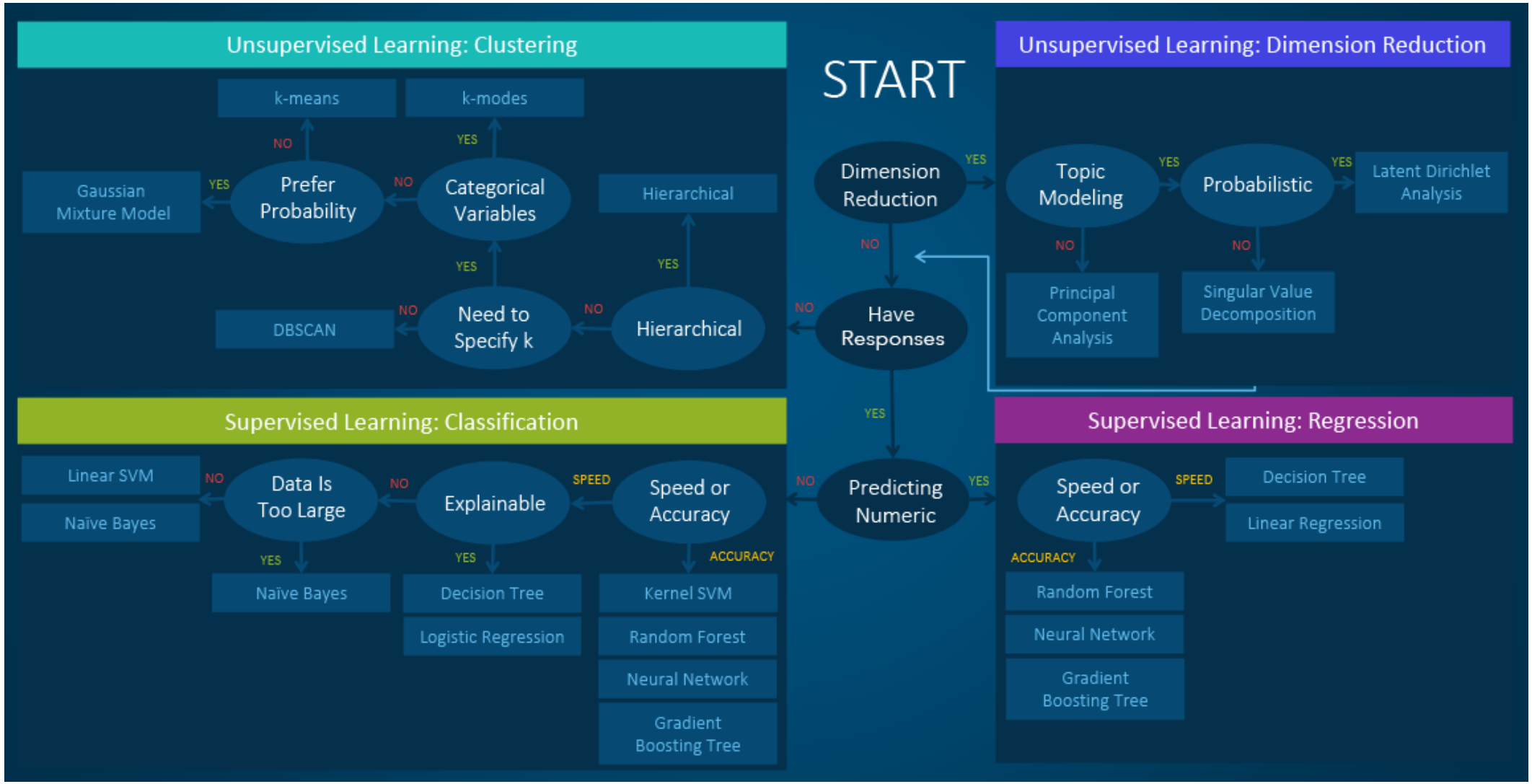
- Tabelas de Frequência: Frequência Absoluta, Relativa e Acumulada
- Visualização de Dados: Box-Plots, Histogramas, Gráficos de Linhas ou Colunas
- Medidas Resumo: Média, Mediana, Quartis, Desvio Padrão

- Correlação de Pearson
- Coeficiente de Determinação
- *Information Value* (IV)

Framework de Análise de Dados

Resumo das principais técnicas de Machine Learning

Preditiva.ai



Fonte: https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

Framework de Análise de Dados

Etapa 4: Desenvolvendo o estudo ou modelo

Preditiva.ai

Desenvolvimento

- ❑ Escolha da técnica estatística que responde o problema
- ❑ **Desenvolvimento do estudo analítico ou modelo**

O que deve ser feito ?

Finalmente chegamos ao desenvolvimento do estudo analítico ou modelo estatístico/machine learning. Após o entendimento do problema de negócio, coleta e preparação dos dados, iniciamos a análise em busca de insights e oportunidades de melhoria do processo com o uso das técnicas e ferramentas analíticas.

Como realizar ?

Vai depender do tipo de técnica escolhida. Em estudos descritivos são utilizadas técnicas de estatística descritiva como tabelas de frequência, medidas resumo ou gráficos. Já estudos de inferência precisamos validar se a amostra representa bem a população antes de realizar os cálculos e comparações. Para modelos estatísticos e de Machine Learning, uma série de cuidados devem ser tomados, entre elas, a quantidade e qualidade das variáveis explicativas.

Dicas da Preditiva:

- Vejamos na próxima página um apanhado de dicas para desenvolvimento de estudos ou modelos.

Framework de Análise de Dados

Dicas gerais de desenvolvimento de estudos ou modelos

Preditiva.ai

Atenção para as premissas e pontos fracos e fortes de cada técnica

Invista um tempo suficiente para "limpar" a base, pois caso contrário seu estudo/modelo pode não trazer valor.

"Garbage In, Garbage Out"

Sempre valide um modelo utilizando técnicas de **validação cruzada**

Tome cuidado com o **Data Leakage** (quando uma variável em um período "futuro" ou da partição de teste "vaza" para a partição de treino do modelo)

Converse com as áreas de negócio/clientes constantemente. Eles podem fornecer muito direcionamento para interpretar (e validar) os *insights* obtidos

Busque estudos anteriores na literatura. É pouco provável que você esteja fazendo um trabalho totalmente inédito. Busque papers, estudos anteriores para buscar inspiração

Comece com um modelo mais simples e vá sofisticando se houver necessidade

Tente combinar variáveis. Você pode ser surpreender com o resultado

Desenvolva seu estudo ou modelo já pensando em como ele pode ser implantado. Nada pior que um projeto interessante mas inviável

Interprete seu modelo. Não apenas informe sua performance

Extraia uma "foto" dos seus dados e não mude mais. Sempre use random-state nos pacotes do Python. Isso é importante para reperformar seus resultados na etapa de validação

Seja ***Data Centric*** **em vez de** ***Model Centric*****.** Os modelos não fazem milagre se você não tiver boas variáveis

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 5: Validação do trabalho

Framework de Análise de Dados

Etapa 5: Validação do Trabalho

Preditiva.ai

Validação

- ❑ **Verificação dos critérios de sucesso e Validação do Estudo/Modelo**

O que deve ser feito ?

Nesta etapa verificamos se estamos satisfeitos com os resultados obtidos. Além disso, revisamos todo o processo em busca de falhas para validar o estudo ou modelo antes de entrar em implantação.

Como realizar ?

Um bom estudo ou modelo é aquele que dá boas perspectivas de melhoria do processo ou negócio envolvido. Lembra que na primeira etapa da metodologia nós definimos os Critérios de Sucesso? Se você está convencido(a) o suficiente de que seu trabalho alcançou os critérios, você já tem uma **primeira versão do projeto**.

Com isso, falta validar se o seu estudo/modelo está correto do ponto de vista estatístico ou mesmo que todas as etapas operacionais foram cumpridas corretamente. Dependendo do tamanho da empresa ou do quanto regulado é o setor, existem áreas específicas focas em validação. Se não é o seu caso, um colega de trabalho ou seu gestor podem ser os validadores. Ou seja, realize um ***peer review*** de seu trabalho.

Dicas da Preditiva:

- **Desapegue**! Desenvolva um estudo bom o suficiente e desapegue! É muito comum ficarmos presos no sonho de perfeição. Assim como em várias frentes de nossa vida, lembre-se: Feito é melhor que perfeito!
- **Finalize a primeira versão da documentação** e entregue para o validador tentar reperformar seus resultados
- **Segure o ego e orgulho**. Erros acontecem a todos, principalmente em uma atividade que se você errar um vírgula o resultado já pode ser totalmente diferente. **Peça para o seu *peer reviewer* ser sincero(a)** e não se preocupar em criticar o seu trabalho. Essa etapa serve para que o estudo seja o mais preciso possível pois muita coisa (inclusive dinheiro) pode derivar daí.

Framework de Análise de Dados

Etapa 5: Validação do Trabalho

Preditiva.ai

Validação

- ❑ Verificação dos critérios de sucesso e Validação do Estudo/Modelo
- ❑ **Aprovação pelo cliente do trabalho e Atualização do Roadmap**

O que deve ser feito ?

Trabalho validado. E agora? Vamos chamar o cliente do trabalho para compartilhar esse super estudo/modelo e já contarem os dias para a implantação.

Como realizar ?

Convide para a mesma reunião seu gestor e seus clientes. Apresente seu trabalho com confiança (afinal, você validou os insights com o cliente durante o desenvolvimento e seu colega também validou com você as planilhas e códigos). Um apresentação de um projeto de dados deve conter, no mínimo, as seguintes etapas:

- Objetivo do projeto
- Premissas e alinhamentos iniciais
- Apresentação dos tipos de variáveis analisadas e suas respectivas análises exploratórias
- Propostas de solução para o(s) problema(s) de negócio (estudo, modelo, dashboard, etc)
- Próximos passos do projeto (ex: aprovação, implantação, acompanhamento etc)
- Roadmap

Dicas da Preditiva:

- Envie a apresentação aos presentes no final da reunião.
- Peça o “de acordo” formal do cliente para dar os próximos passos (implantação).
- O que não for possível realizar em tempo de projeto, deixe no Roadmap para as próximas versões.
- Estude técnicas de **Data Storytelling** e sempre pratique para melhorar sua **oratória**
- Para criar planos de ação, uma abordagem é utilizar a **metodologia 5W 2H**. Desta forma o plano fica mais claro. É importante também dar uma ideia de prioridade para o cliente, assim o foco fica nos planos que podem trazer mais resultado no curto prazo.

Framework de Análise de Dados

Exemplos de planos de ação sugeridos

Preditiva.ai

Projeto People Analytics: Minimização do Turnover de colaboradores

Prioridade	O que fazer?	Por que?	Quem?	Como fazer?	Qual o custo?
1	Revisão da política de Horas Extras	Colaboradores que fazem hora extra têm probabilidade de 31% de turnover.	Time de RH	Criar um grupo controle e teste de pessoas que podem fazer hora extra e não podem. Assim podemos mensurar a real diminuição de turnover ao longo do tempo com KPI's e Dashboards de acompanhamento.	Diminuição da produtividade de algumas áreas que precisam que os colaboradores trabalhem mais horas.
2	Revisão salarial	Colaboradores que ganham até 3000 reais têm probabilidade de 29% de turnover.	Time de Remuneração	Verificar se os salários da empresa estão atualizados de acordo com o mercado. Caso não estejam, estudar a possibilidade de adequação salarial.	Aumento da folha de pagamento.
3	Revisão da política de Viagens Corporativas	Colaboradores que viajam frequentemente têm probabilidade de 25% de turnover.	Time de RH	Criar um grupo controle e teste de pessoas que podem viajar e não podem. Assim podemos mensurar a real diminuição de turnover ao longo do tempo com KPI's e Dashboards de acompanhamento. Além disso, intensificar ferramentas de trabalho remoto na empresa, mudando processos e a cultura.	Diminuição da produtividade de algumas áreas que precisam que os colaboradores trabalhem localmente. No entanto, pode diminuir o curso com alternativas de trabalho remoto.
4	Revisão da Integração e Execução da Cultura	Colaboradores com até 2 anos de empresa têm probabilidade de 30% de turnover.	Time de Cultura e Onboarding	Verificar se o onboarding dos colaboradores está adequado. Isso envolve acessos, integração com o time, liderança e feedback. Também faze constantes mensurações de como a cultura organizacional está sendo seguida e executada pelos gestores.	Apenas horas de projeto dos times envolvidos.
5	Pesquisa de Satisfação mais detalhada	Colaboradores com baixa satisfação com o trabalho têm probabilidade de 25% de turnover.	Time de Cultura	Como não temos dados das causas da baixa satisfação, é importante conduzir uma pesquisa mais abrangente com perguntas qualitativas e abertas dos motivos para baixa satisfação.	Fornecedor externo com a ferramenta de pesquisa.

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 6: Deploy - Implantação

Framework de Análise de Dados

Etapa 6: Implantação

Preditiva.ai

Implantação

- ❑ **Plano de implantação e monitoramento do estudo ou modelo**

O que deve ser feito ?

Finalmente chegamos à última etapa do projeto. Após muito trabalho, chegou a hora de iniciar a colheita dos resultados. Nesta etapa precisamos criar um plano de implantação dos planos de ação validados pelo cliente e já estruturar uma forma de monitorar os resultados desse plano ou modelo implantado.

Como realizar ?

Naturalmente vai depender do tipo de negócio, tecnologia e processos da empresa. Mas de uma forma geral, se o seu plano de ação visa melhorar um indicador/KPI, é interessante criar um acompanhamento automático em um Dashboard com indicação clara do momento em que o plano de ação foi implantado. Já para modelos de Machine Learning, defina as métricas a serem utilizadas para o monitoramento do modelo, crie um Dashboard de acompanhamento do modelo e atualize a documentação. Algumas métricas de monitoramento comuns:

- KS/AuROC ou qualquer medida em treino.
- Distribuição dos scores/clusters em treino.
- Taxa do Target em treino.

Dicas da Preditiva:

- **Insira também os gatilho**s propostos para iniciar um processo de re-treino ou calibragem do modelo.
- Ao implantar a melhoria trazida pelo projeto, não faça tudo de uma vez com toda a sua base. **Crie grupos de controle e teste** e vá monitoramento as diferenças entre os grupos ao longo do tempo. Se de fato o grupo teste estiver melhor, aumente esse grupo até que todo o seu processo esteja passando pelas diretrizes na melhoria implantada. É muito mais seguro fazer desta forma.

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Etapa 6: Deploy – Mensurando o valor dos projetos

Valor de um Projeto de Dados

Mensuração de Resultados

Preditiva.ai

Uma parte fundamental dos **Projetos de Dados** é a **mensuração de resultados**. Afinal de contas, depois de percorrer todas as etapas, do **Entendimento do Negócio** ao **Deploy**, precisamos apurar quais foram os ganhos obtidos pela implantação do projeto, seja ele um estudo ou modelo.

O ***Backtest*** é muito utilizado para avaliar o **potencial** de resultado financeiro do projeto de dados, enquanto o **Teste / Controle** é o método utilizado para mensurar o **valor real** que o projeto está gerando para a empresa.

Backtest

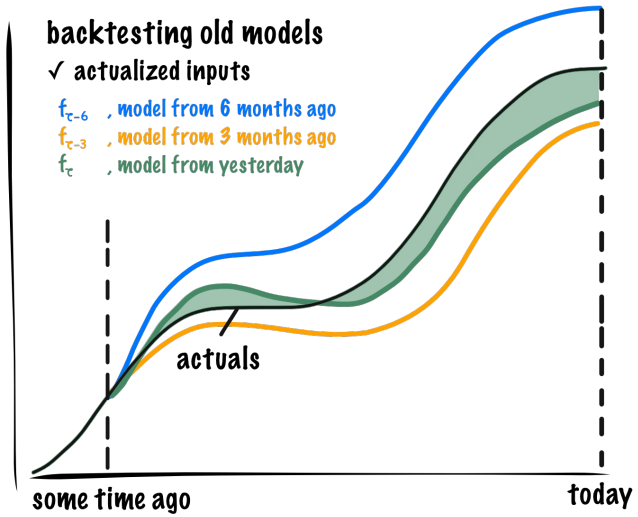


Imagem: https://medium.com/@FMZ_Quant/5-1-the-meaning-and-trap-of-backtesting-b4aff13f8285

Teste / Controle

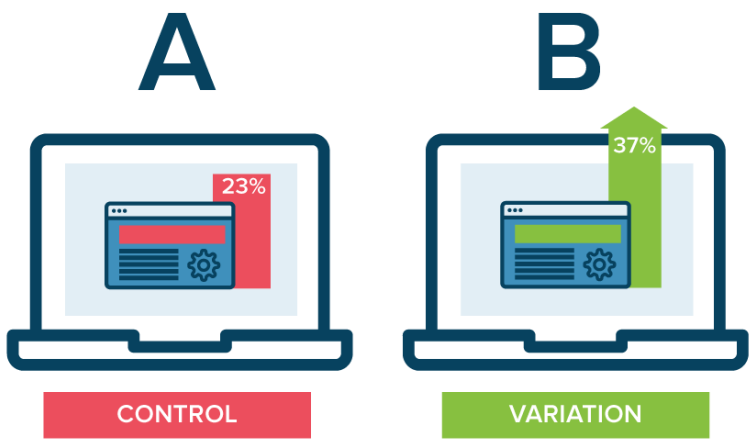


Imagem: https://www.optimizely.com/optimization-glossary/ab-testing/

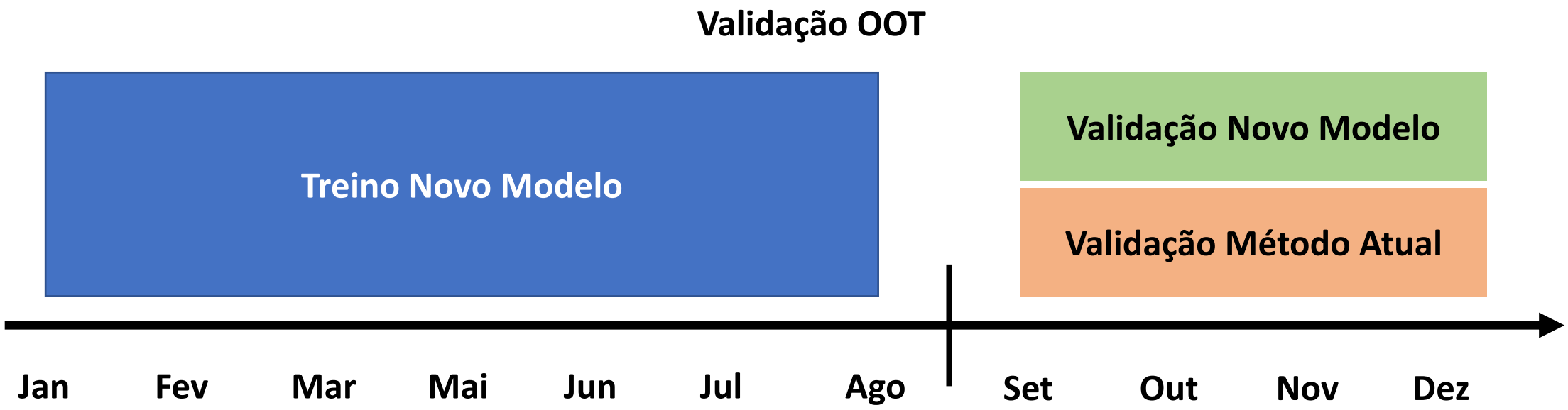
Backtesting

Valor de um Projeto de Dados

Mensuração de Resultados: *Backtest*

Preditiva.ai

O ***Backtest*** consiste em fazer uma comparação entre os resultados obtidos com o **Novo Modelo** vs. **Método Atual**, utilizando uma base de dados histórica, normalmente um dataset ***out-of-time***.



Valor de um Projeto de Dados

Mensuração de Resultados: *Backtest*

Preditiva.ai

Para apresentar a aplicação prática do ***Backtest***, considere o seguinte exemplo:

Uma empresa de telefonia celular, preocupada com a potencial migração de seus clientes para o concorrente, solicitou o desenvolvimento de um modelo para estimar a **probabilidade de cada cliente cancelar seu contrato**. Esse abandono é também conhecido como ***Churn***. Dessa forma, essa empresa poderia identificar os clientes com maior probabilidade de abandono e apresentar ofertas diferenciadas para eles.

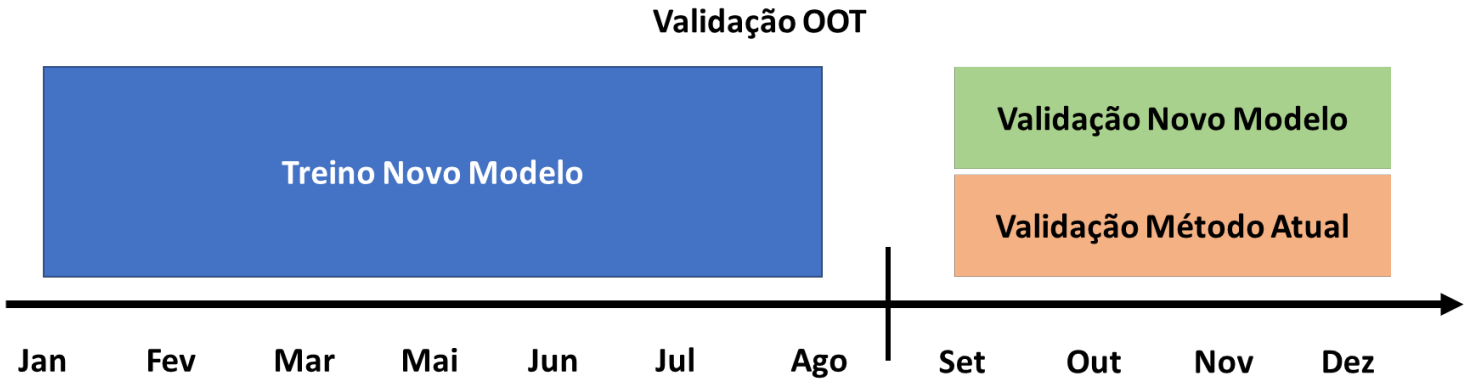
Photo by John Tuesday on Unsplash

Valor de um Projeto de Dados

Mensuração de Resultados: *Backtest*

Preditiva.ai

Após o **desenvolvimento** do modelo utilizando o dataset de **Treino**, utilizamos o modelo para fazer a **predição** do dataset de **Validação/Teste** e avaliamos o desempenho do modelo nesse conjunto de dados.



Métrica	Treino Novo Modelo	Validação Novo Modelo	Validação Método Atual
Acurácia	75%	72%	64%
Precision	87%	85%	78%
Recall	95%	94%	89%
...	...	...	...
AUROC	89%	87%	79%

O **Novo Modelo** está apresentando resultados melhores que o **Método Atual**, mas precisamos converter esse ganho de performance em **valores financeiros**!

Valor de um Projeto de Dados

Mensuração de Resultados

Preditiva.ai

Para medir o **impacto financeiro** gerado pelo **Novo Modelo** devemos avaliar as decisões tomadas a partir deste modelo e as consequências financeiras dessas decisões. Para isso, uma **matriz de custos**:

Real / Predito	*Churn*	Não *Churn*
Churn	**Real = Predito** Custos de retenção: -R$ 200,00 Receitas com cliente: R$ 500,00 **Resultado: R$ 300,00**	**Real** Custos de retenção: -R$ 200,00 Receitas com cliente: R$ 500,00 **Predito** Custos de retenção: R$ 0,00 Receitas com cliente: R$ 0,00 **Resultado: -R$ 300,00**
Não *Churn*	**Real** Custos de retenção: -R$ 0,00 Receitas com cliente: R$ 500,00 **Predito** Custos de retenção: -R$ 200,00 Receitas com cliente: R$ 500,00 **Resultado: -R$ 200,00**	**Real = Predito** Custos de retenção: R$ 0,00 Receitas com cliente: R$ 500,00 **Resultado: R$ 500,00**

Valor de um Projeto de Dados

Mensuração de Resultados: *Backtest*

Preditiva.ai

Devemos então aplicar a **Matriz de Custos** na **Matriz de Confusão** de cada método. Supondo que o dataset de **Validação** possua 500 clientes, obtivemos as seguintes matrizes de confusão:

Matriz de Custos $

Real / Predito	Churn	Não Churn
Churn	R$ 300,00	-R$ 300,00
Não Churn	-R$ 200,00	R$ 500,00

Novo Modelo *Churn*

Real / Predito	*Churn*	Não *Churn*
Churn	120	20
Não *Churn*	120	240

Método Atual

Real / Predito	*Churn*	Não *Churn*
Churn	120	140
Não *Churn*	40	200

Valor de um Projeto de Dados

Mensuração de Resultados: *Backtest*

Preditiva.ai

Considerando as premissas do **número de clientes da empresa** e a **diferença** obtida entre o **Novo Modelo** e o **Método Atual**, podemos concluir que ao utilizar o **Novo Modelo *Churn*** a empresa tem um **potencial** estimado de gerar um valor adicional de aproximadamente **R$16 milhões**.

Impacto Novo Modelo Churn

Real / Predito	Churn	Não Churn	Total
Churn	36.000,00	-6.000,00	**30.000,00**
Não Churn	-24.000,00	120.000,00	**96.000,00**
Total	**12.000,00**	**114.000,00**	**126.000,00**

Resultado Médio por Cliente	252,00
Ganho em relação ao Método Atual (R$)	80,00
Ganho em relação ao Método Atual (%)	47%
Número de clientes da empresa	200.000
Potencial de resultado pelo novo modelo	16.000.000,00

Impacto Método Atual

Real / Predito	Churn	Não Churn	Total
Churn	36.000,00	-42.000,00	**-6.000,00**
Não Churn	-8.000,00	100.000,00	**92.000,00**
Total	**28.000,00**	**58.000,00**	**86.000,00**

Resultado Médio por Cliente	172,00

Teste e Controle

Valor de um Projeto de Dados

Mensuração de Resultados: Teste / Controle

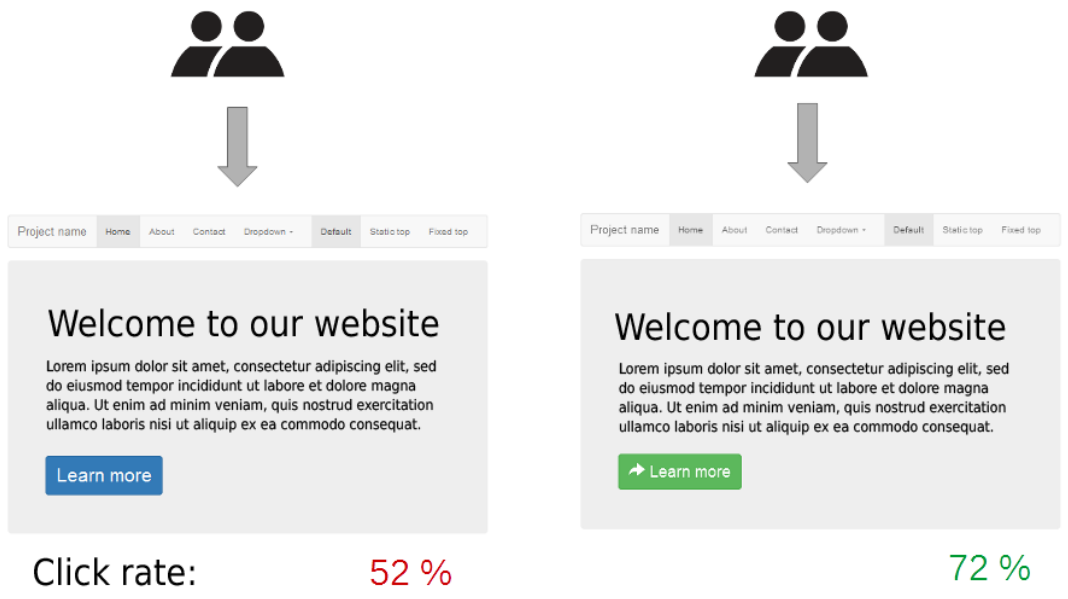
Preditiva.ai

O **Teste/Controle**, também conhecido como **Teste A/B**, é um método muito utilizado em pesquisas de experimentação com usuários.

O método consiste em apresentar **2 ou mais versões** diferentes para o usuário e **medir o resultado** obtido em cada versão.

Neste exemplo são apresentadas duas versões diferentes do site e é feita a medição na taxa de cliques no botão.

Com isso, são identificados **padrões** que geram **melhores resultados**, de acordo com os objetivos pré-definidos.



Fonte: https://upload.wikimedia.org/wikipedia/commons/2/2e/A-B_testing_example.png

Valor de um Projeto de Dados

Mensuração de Resultados: Teste / Controle

Preditiva.ai

O **Teste/Controle** em projetos de dados funciona de forma semelhante. Considere o mesmo exemplo da empresa de telefonia celular que apresentamos na aplicação da técnica ***Backtest***.

Após concluir o desenvolvimento do novo modelo, a empresa quer **mensurar** quanto de **valor** o novo modelo está **efetivamente** gerando. Para isso, decidiu aplicar a técnica de **Teste/Controle**.

Photo by John Tuesday on Unsplash

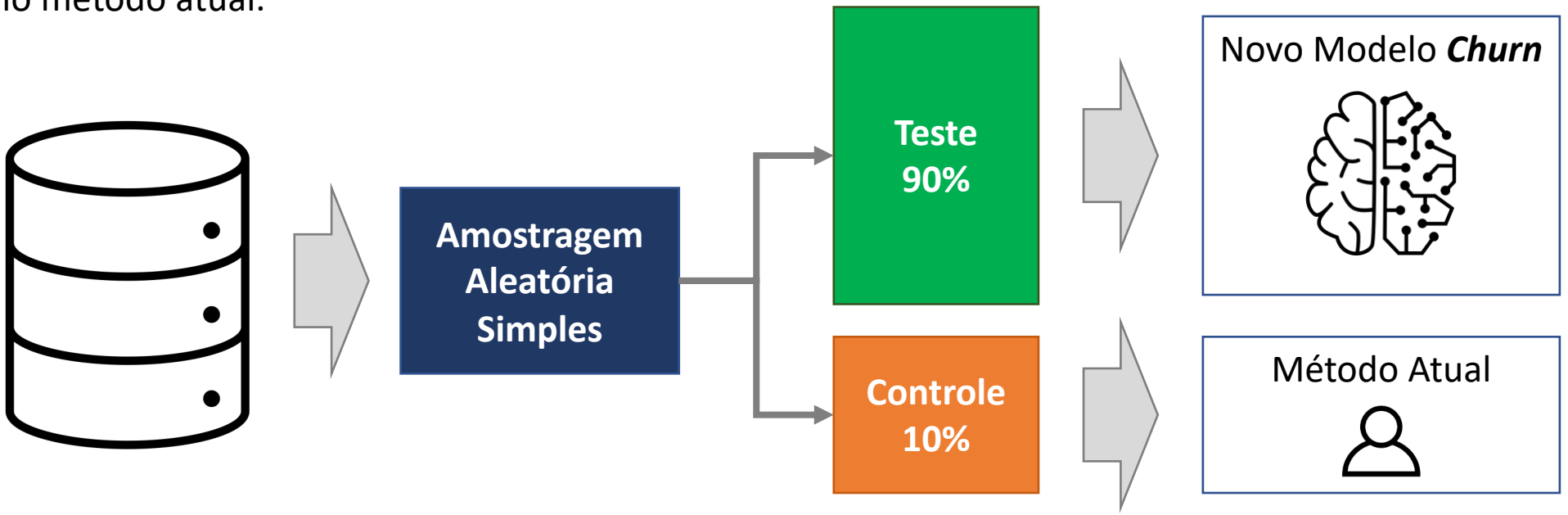
Valor de um Projeto de Dados

Mensuração de Resultados: Teste / Controle

Preditiva.ai

Para aplicar essa técnica a base de clientes será dividida aleatoriamente em 2 grupos: **Teste** e **Controle**. A separação dos clientes nos grupos **Teste** e **Controle** deve ser feita de forma que os clientes desses 2 grupos possuam **características semelhantes**. Qualquer diferença entre os grupos **compromete** o resultado da **comparação**.

O novo modelo será utilizado para estimar a probabilidade de ***Churn*** no grupo **Teste**, que neste exemplo contém 90% dos clientes. O grupo **Controle**, com 10% dos clientes restantes terão sua probabilidade de ***Churn*** estimada com base no método atual.



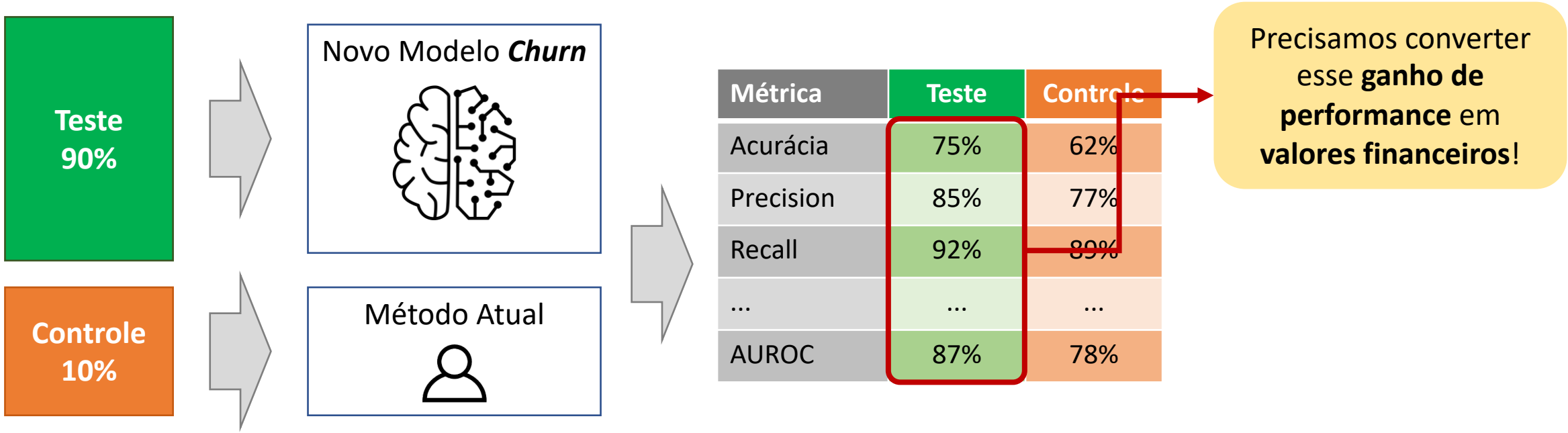
Valor de um Projeto de Dados

Mensuração de Resultados: Teste / Controle

Preditiva.ai

Durante o período de 1 mês foram calculadas as estimativas da probabilidade de ***Churn*** de 1.000 clientes, sendo **900** calculadas a partir do **novo modelo**, enquanto os **100** restantes tiveram a estimativa calculada a partir do **método atual**.

Comparamos então os grupos **Teste** e **Controle** utilizando os indicadores de performance adequados. Avaliando esses indicadores podemos ver que o **Novo Modelo de *Churn*** apresenta **melhores resultados**, confirmando a análise do ***Backtest***.



Valor de um Projeto de Dados

Mensuração de Resultados

Preditiva.ai

Para medir o **impacto financeiro** gerado pelo novo modelo devemos avaliar as decisões tomadas a partir deste modelo e as consequências financeiras dessas decisões. Para isso, uma **matriz de custos**:

Real / Predito	*Churn*	Não *Churn*
Churn	**Real = Predito** Custos de retenção: -R$ 200,00 Receitas com cliente: R$ 500,00 **Resultado: R$ 300,00**	**Real** Custos de retenção: -R$ 200,00 Receitas com cliente: R$ 500,00 **Predito** Custos de retenção: R$ 0,00 Receitas com cliente: R$ 0,00 **Resultado: -R$ 300,00**
Não *Churn*	**Real** Custos de retenção: -R$ 0,00 Receitas com cliente: R$ 500,00 **Predito** Custos de retenção: -R$ 200,00 Receitas com cliente: R$ 500,00 **Resultado: -R$ 200,00**	**Real = Predito** Custos de retenção: R$ 0,00 Receitas com cliente: R$ 500,00 **Resultado: R$ 500,00**

Valor de um Projeto de Dados

Mensuração de Resultados: Teste / Controle

Preditiva.ai

Considerando as premissas o **número de clientes que utilizaram o novo modelo** e a **diferença** obtida entre os grupos **Teste** e **Controle**, o **Novo Modelo Churn** gerou neste primeiro mês **R$ 73.800,00**.

Impacto Novo Modelo Churn

Real / Predito	Churn	Não Churn	Total
Churn	67.500,00	-40.500,00	**27.000,00**
Não Churn	-18.000,00	225.000,00	**207.000,00**
Total	**49.500,00**	**184.500,00**	**234.000,00**

Resultado Médio por Cliente	260,00
Ganho em relação ao Método Anterior (R$)	82,00
Ganho em relação ao Método Anterior (%)	46%
Número de clientes no novo modelo	900
Resultado gerado pelo novo modelo	73.800,00

Impacto Método Atual

Real / Predito	Churn	Não Churn	Total
Churn	6.600,00	-3.600,00	**3.000,00**
Não Churn	-5.200,00	20.000,00	**14.800,00**
Total	**1.400,00**	**16.400,00**	**17.800,00**

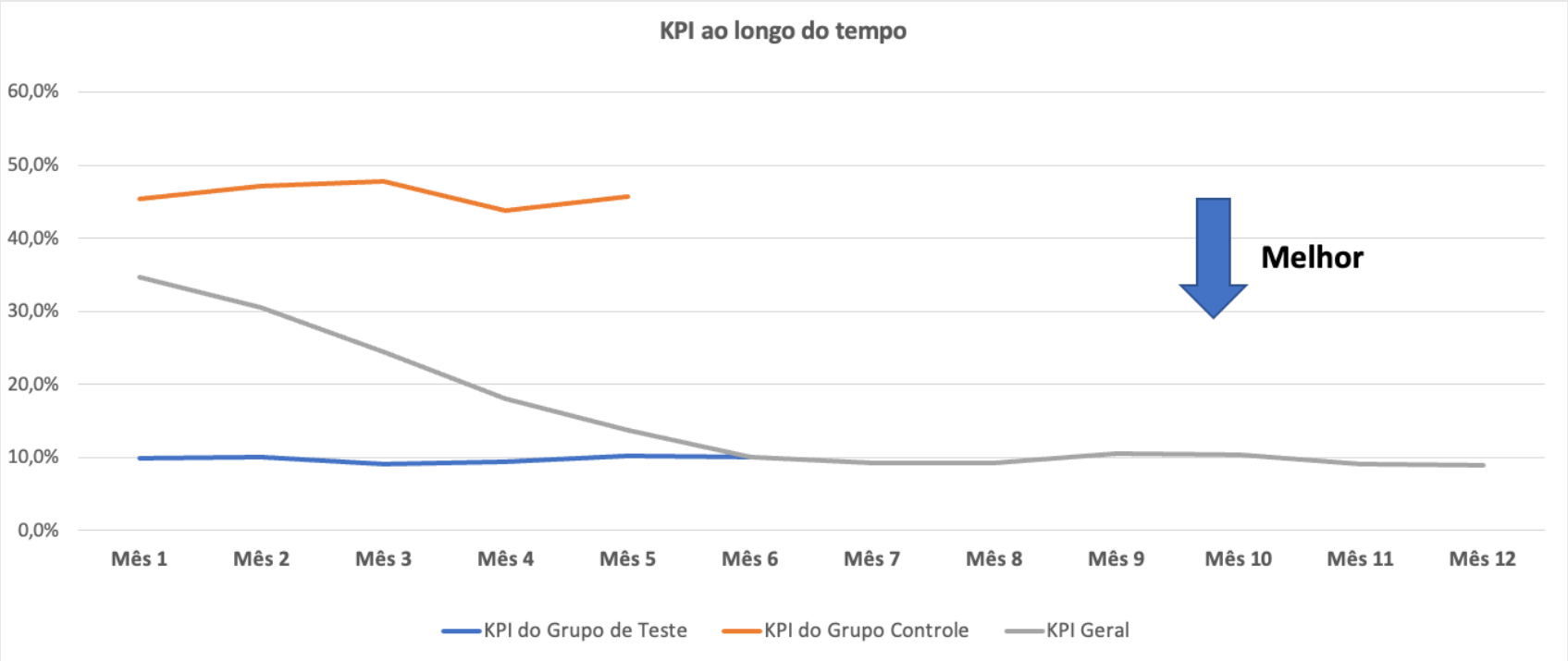
Resultado Médio por Cliente	178,00

Valor de um Projeto de Dados

Outro exemplo de acompanhamento por Teste e Controle

Preditiva.ai

Plano de Ação	Revisão da política de Horas Extras
Por que?	Colaboradores que fazem hora extra têm probabilidade de 31% de turnover.
O que foi feito?	Criação de grupo controle e teste de pessoas que podem fazer hora extra e não podem.
KPI em acompanhamento	Percentual de Turnover (% de colaboradores que deixaram a empresa)



	Mês 1	Mês 2	Mês 3	Mês 4	Mês 5	Mês 6	Mês 7	Mês 8	Mês 9	Mês 10	Mês 11	Mês 12
% do Grupo Teste (grupo com aplicação da mudança)	30%	45%	60%	75%	90%	100%	100%	100%	100%	100%	100%	100%
% do Grupo Controle	70%	55%	40%	25%	10%	0%	0%	0%	0%	0%	0%	0%
KPI do Grupo de Teste	9,8%	10,1%	9,0%	9,4%	10,2%	10,0%	9,2%	9,3%	10,6%	10,3%	9,1%	9,1%
KPI do Grupo Controle	45,4%	47,2%	47,8%	43,7%	45,7%							
KPI Geral	**34,7%**	**30,5%**	**24,5%**	**18,0%**	**13,8%**	**10,0%**	**9,2%**	**9,3%**	**10,6%**	**10,3%**	**9,1%**	**9,0%**

Preditiva.ai

Framework de Análise de Dados
Metodologia CRISP-DM

Considerações finais

Framework de Análise de Dados

Considerações finais

Preditiva.ai

Recapitulando os passos da metodologia. Utilize essas etapas como um "**check-list**" do seu projeto:

Entendimento do negócio

- ❑ Objetivo
- ❑ Premissas
- ❑ Riscos envolvidos
- ❑ Custos x Benefícios
- ❑ Critérios de sucesso
- ❑ Planejamento do projeto e início da documentação

Entendimento dos dados

- ❑ Descrição dos dados
- ❑ Coleta
- ❑ Análise Exploratória
- ❑ Verificação da qualidade dos dados (*Sanity Check*)

Preparação dos dados

- ❑ Seleção de variáveis
- ❑ Limpeza dos dados
- ❑ Formatação dos dados

Modelagem/ Análise

- ❑ Escolha da técnica estatística que responde o problema
- ❑ Desenvolvi-mento do estudo analítico ou modelo
- ❑ Finalização

Validação

- ❑ Verificação dos critérios de sucesso
- ❑ Validação do estudo/ modelo
- ❑ Aprovação pelo cliente do trabalho
- ❑ Atualização do Roadmap

Deploy (Implantação)

- ❑ Plano de implantação
- ❑ Monitoramento do modelo/KPI's
- ❑ Finalização da documentação

Framework de Análise de Dados

Metodologia Ágil para Data Science

Preditiva.ai

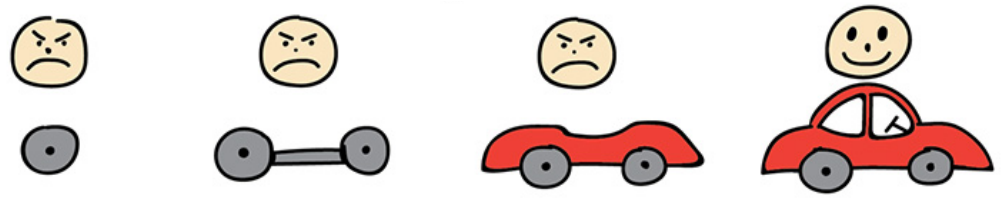
Por muito tempo, o **CRISP-DM** foi o principal framework de desenvolvimento. No entanto, com o advento das **metodologias ágeis** para desenvolvimento de produtos, muito se fala em como agregar os benefícios dessas metodologias para projetos de Ciência de Dados.

Alguns dos motivadores dessa mudança:

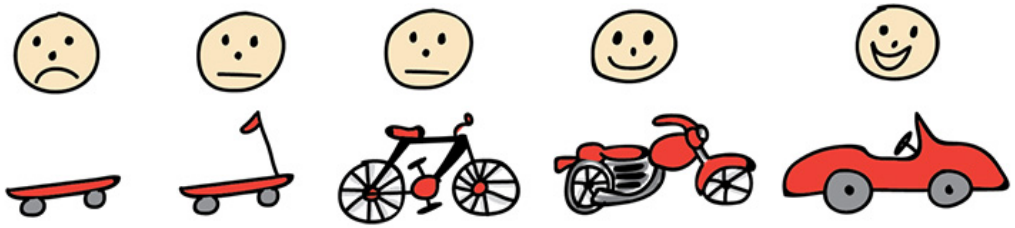
- Como o próprio nome diz, **Ciência** de Dados é uma área baseada em experimentação. Diferentemente de Engenharia de Software, onde os resultados esperados são determinísticos, em **Ciência de Dados o resultado final em geral é imprevisível**. Portanto, metodologias lineares como o CRISP-DM não são ideais.

- Por ser Waterfall, o cliente só terá insights e benefícios no final do projeto, que pode demorar muito. Por exemplo, em bancos tradicionais, em geral **um projeto de modelagem dura entre 5 a 10 meses**.

Abordagem Tradicional. Ex: CRISP-DM



Nova Abordagem: Metodologia Ágil para Data Science

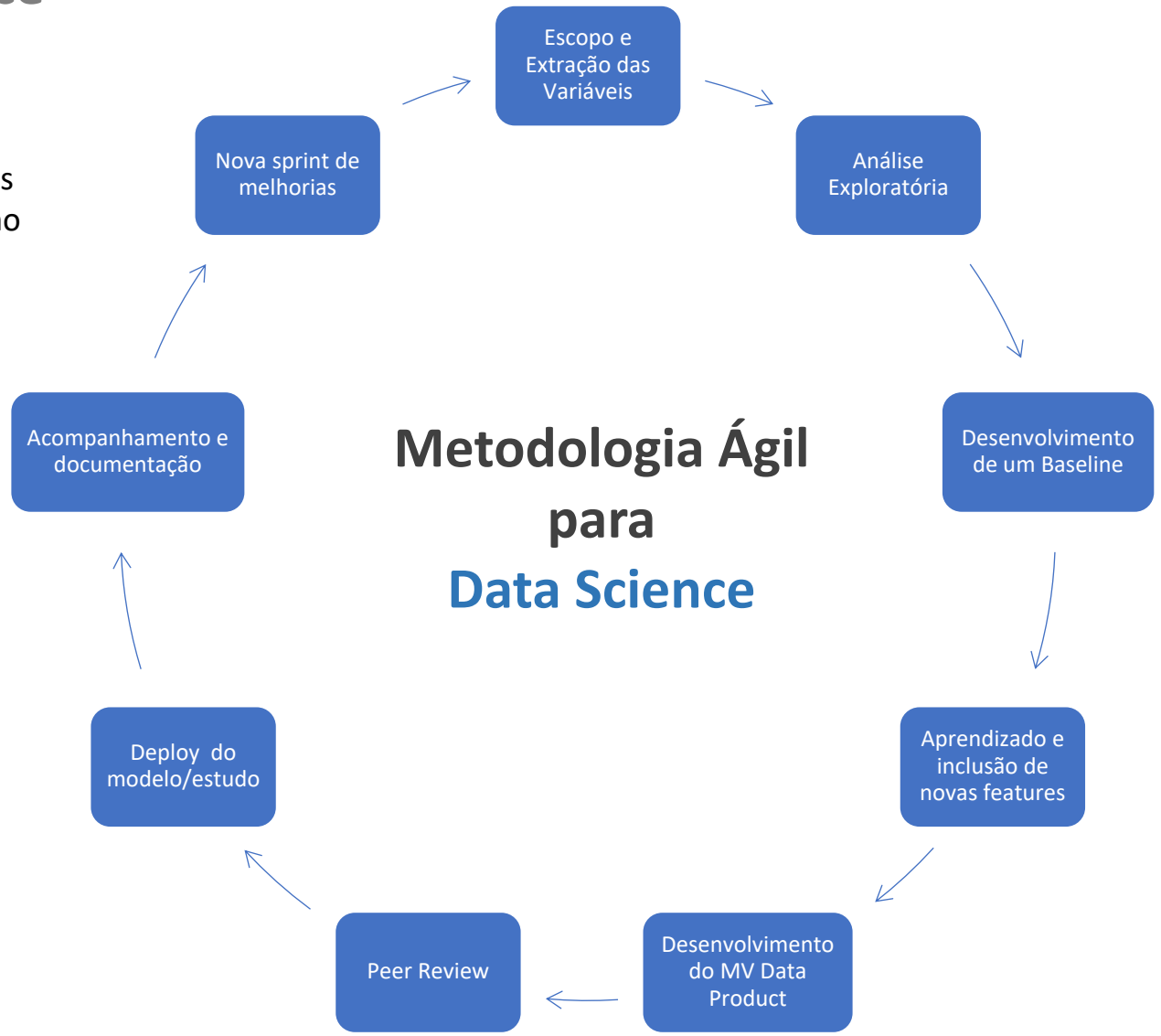


Framework de Análise de Dados

Metodologia Ágil para Data Science

Preditiva.ai

Ainda não temos uma metodologia definitiva que mescla os conceitos do CRISP-DM com os de Metodologias Ágeis como o **Scrum** ou **KanBan**. No entato, a maioria dos frameworks seguem basicamente as etapas do ciclo ao lado:



Framework de Análise de Dados

Sua capacidade de desenvolver bons projetos melhora com o tempo

Preditiva.ai

Match the Right Level of Data Science Seniority to the Tasks to Be Done

☆☆☆☆ Core ☆☆☆ Significant ☆☆ Decent ☆ Some

	Guide, Inspire and Storytelling	Formulate/ Prioritize Projects	Collect and Integrate Data	Prepare and Refine Data	Explore and Understand Data	Build ML Models	Operation-alize ML Models
Junior Data Scientists	☆	☆☆	☆☆	☆☆☆☆	☆☆☆☆	☆☆☆☆	☆
Midlevel Data Scientists	☆☆☆	☆☆☆	☆☆☆	☆☆	☆☆	☆☆☆	☆☆
Senior Data Scientists	☆☆☆☆	☆☆☆☆	☆	☆	☆	☆	☆☆

Source: Gartner

720573_C

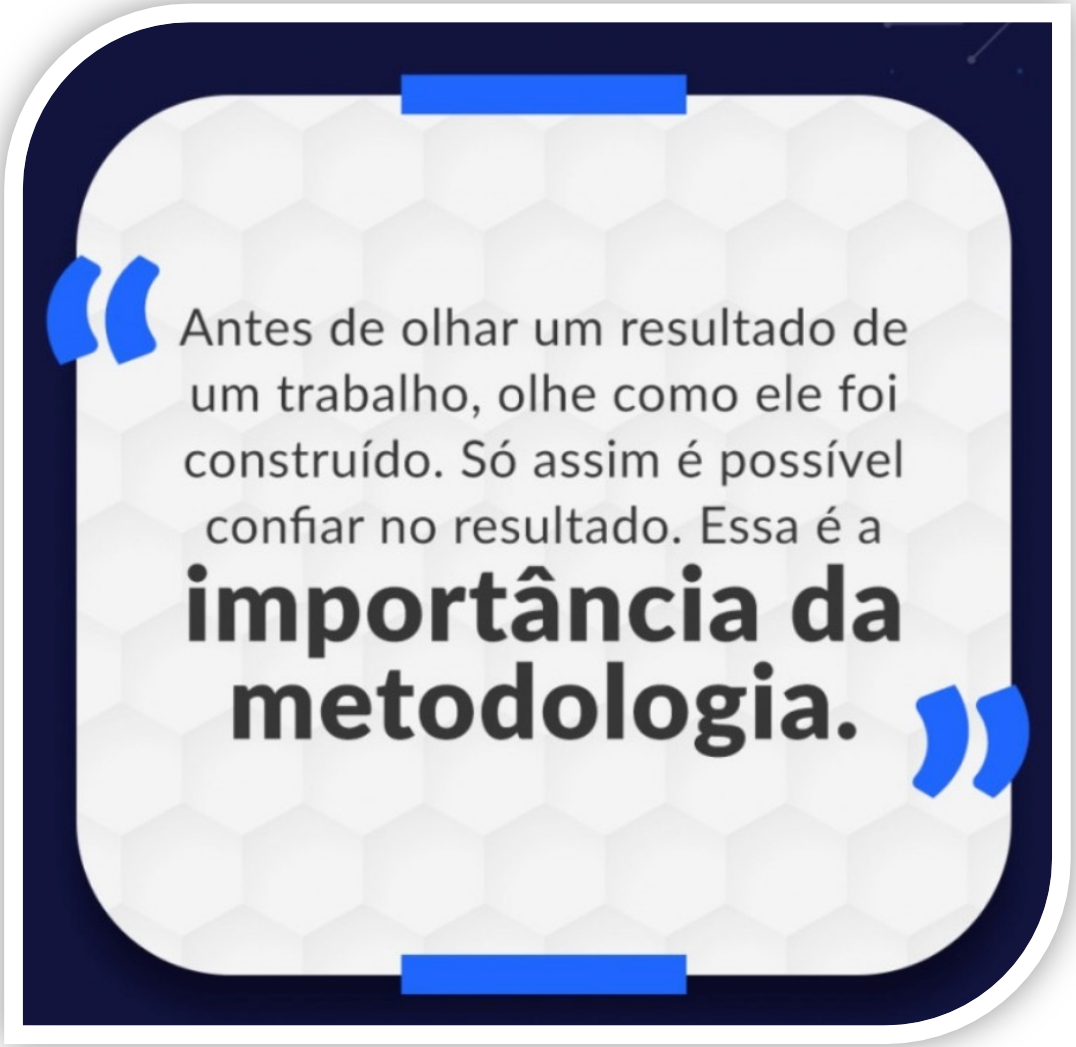
Gartner

Fonte: https://www.gartner.com/en/documents/3983458/staffing-data-science-teams-mapping-capabilities-to-key-

Framework de Análise de Dados

Considerações finais

Preditiva.ai



Preditiva.ai