



Projeto Integrador

Agosto/2020

➤ Agenda

- 1 **Introdução**
- 2 **Etapas**
- 3 **Avaliação**
- 4 **Exemplos**

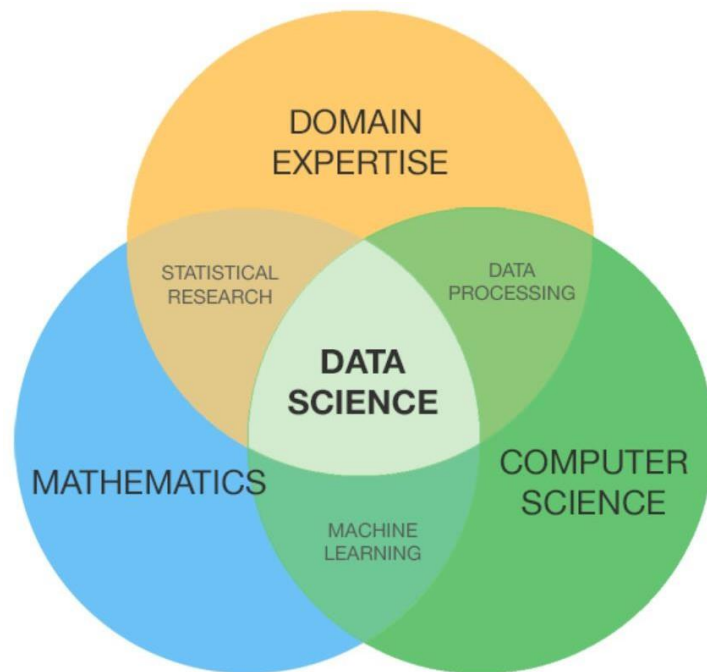


1. Introdução

➤ Organização

- O Projeto Integrador é o trabalho de conclusão do curso de Ciência de Dados e permite a aplicação prática de tudo o que foi aprendido durante o curso
- Cada grupo terá no máximo 6 alunos, para permitir que todos tenham participação relevante na constituição e desenvolvimento do projeto
- O grupo será formado buscando equilíbrio de competências, aproveitando ao máximo os conhecimentos e a bagagem de todos

➤ Perfil do Cientista de Dados



Source: Palmer, Shelly. *Data Science for the C-Suite*.
New York: Digital Living Press, 2015. Print.

➤ Critérios para formação dos grupos

Grupo	Idade	Exp.	Nivel.	Python	Estat	Business	media	tempo
A	25,0	3,0	73,3	2,6	3,0	3,2	2,9	4,6
B	32,2	5,7	63,8	2,8	3,0	3,5	3,1	7,5
C	27,8	2,7	62,5	2,8	3,2	3,7	3,2	4,8
D	33,2	5,0	59,4	3,0	2,7	3,2	2,9	6,3
E	28,2	4,0	72,9	3,2	3,2	3,0	3,1	5,6
F	29,6	6,2	61,7	2,6	2,8	3,2	2,9	5,6
G	30,8	5,5	62,5	3,0	3,5	2,8	3,1	5,8
H	29,8	6,0	75,7	2,7	3,0	3,2	2,9	6,0

> Grupos

Grupo 1

Aline Soares da Silva
Ivan Costa Passos
Lucas Ribeiro
Rafael Medeiros
Raiane Honorato
Stéphanie Vieira Gonçalves

Grupo 2

Ana Cláudia Aires Nogueira
Juliana Coneglian
Renato Souza
Rodrigo Cordeiro Portes
Rudolf Strohdiek

Grupo 3

Arana Paula Guimarães
Lucas Cabral Rosa
Patricia de Menezes Barbosa
Tiago Costa
Victor Brilha Cirillo
Victor Wood Machado

Grupo 4

Alexander Patrick O. Leite
Danilo Yamaguchi
Gustavo Deguti Kajiura
Letícia Berta
Vagner Roberto Junior

Grupo 5

Beatriz Amorim Matos
Brunno Da Silva
Eduardo Santana
Luis Henrique Anaya
Thiago da Silva

Grupo 6

André Audi
Beatriz Yokota
Luís Valverde Guimarães
Mozart Marin
Rafael Zerbini
Victor Wilm

Grupo 7

Guilherme Augusto Stefani
Helena Salgado Lacaz
João Carlos De Souza
Raynaia Maior
Rene da Conceição Lopes
Victor Nunes Botelho

Grupo 8

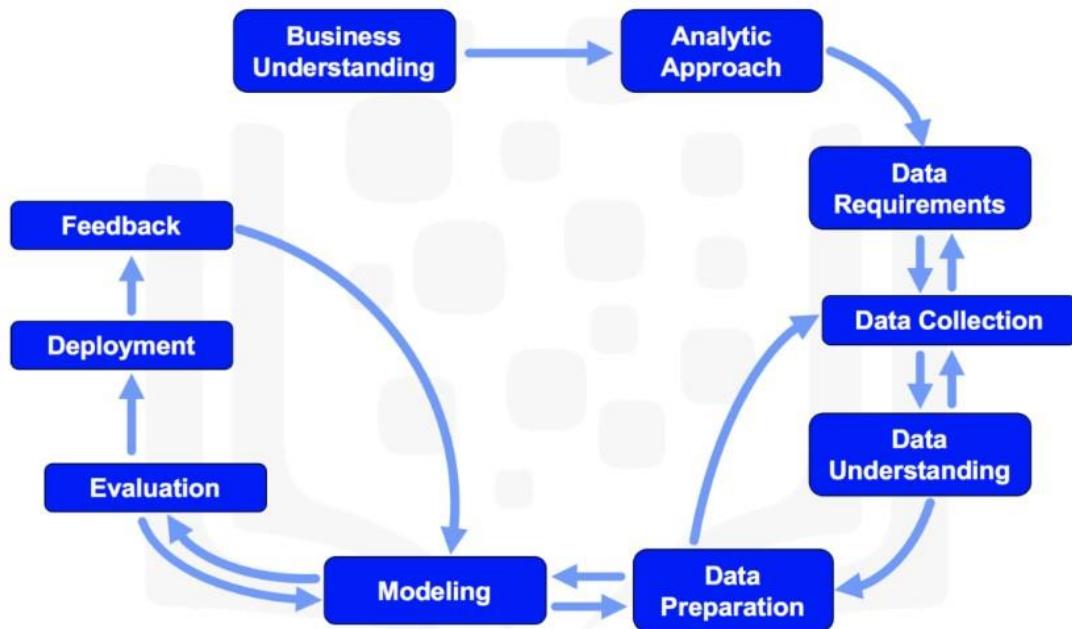
Adrian Ouriques Homrich
Douglas Souza
Henrique Padrao
Higor Da Silva
Maria Geórgia C. Senhorinho
Vinicius Puhl Godinho

➤ Objetivos

- Aplicação real de todas as etapas de projeto de ciência de dados, com atenção para as fases de preparação de dados e modelagem
- Desenvolvimento de soft-skills (organização, divisão de tarefas, trabalho em equipe), em especial no contexto de desenvolvimento de códigos em ciência de dados
- Trabalhar estratégias de como abordar o problema para obtenção de uma solução que possa criar valor para as partes interessadas e/ou impacto social
- Produção de materiais (notebooks, apresentações, etc) que suportem o processo e a solução a ser apresentada, considerando uma audiência técnica e não-técnica

2. Etapas

➤ Processos de um Projeto de DS



John Rollins
Data Scientist, IBM Analytics, IBM

➤ Cronograma

CkPt#	Etapa	Início	Entrega	Duração
1	Apresentação do Tema	14.08.20	31.08.20	17
2	Apresentação dos Dados	31.08.20	26.10.20	56
3	Problematização de Hipóteses e EDA	26.08.20	09.12.20	44
4	Modelagem	09.12.20	27.01.21	49
5	Pré-apresentação e Tuning	27.01.21	08.03.21	40
6	Apresentação Final e MVP	08.03.21	19.03.21	11

➤ 2.1 Apresentação do Tema do Projeto

- Ao final da aula, o grupo deverá apresentar uma lista de propostas, destacando a escolha final e critérios adotados para decisão. Procurem realizar o máximo antes da aula (brainstorms, avaliação, dinâmicas de decisão, etc)
- Abordagem das questões de contexto que envolvem o tema a ser trabalhado, de tal forma que: (1) identifique um público claro; (2) atenda a uma necessidade; e (3) gere valor. Pode-se, por exemplo, realizar entrevistas com partes interessadas
- Procure responder a algumas perguntas de contexto: (1) Qual a situação atual?; (2) Qual o histórico e como evoluiu?; (3) Quais os principais aspectos devem ser “atacados” para melhoria?
- Utilizem o período para definir as rotinas de trabalho comuns: dia e horário recorrente para reuniões, papéis e responsabilidades, repositórios, ferramentas, etc

➤ 2.2 Apresentação das Fontes de Dados

- Apresentar uma compreensão mais ampla do problema, com o entendimento sobre quais dados serão utilizados na construção da solução
- As particularidades do problema dentro do contexto também são relevantes pois abrem novas possibilidades para a obtenção de uma solução criativa e única para a situação em análise
- Garantir que as bases definidas sejam capazes de responder as hipóteses planejadas
- Pensando de maneira criativa: quais fontes de dados (ex: dados abertos, dados internacionais, dados de mídias sociais, scrapping, surveys, etc), poderiam ajudar?
- Lembrem que podem também consultar profissionais de mercado para apoiar seus projetos

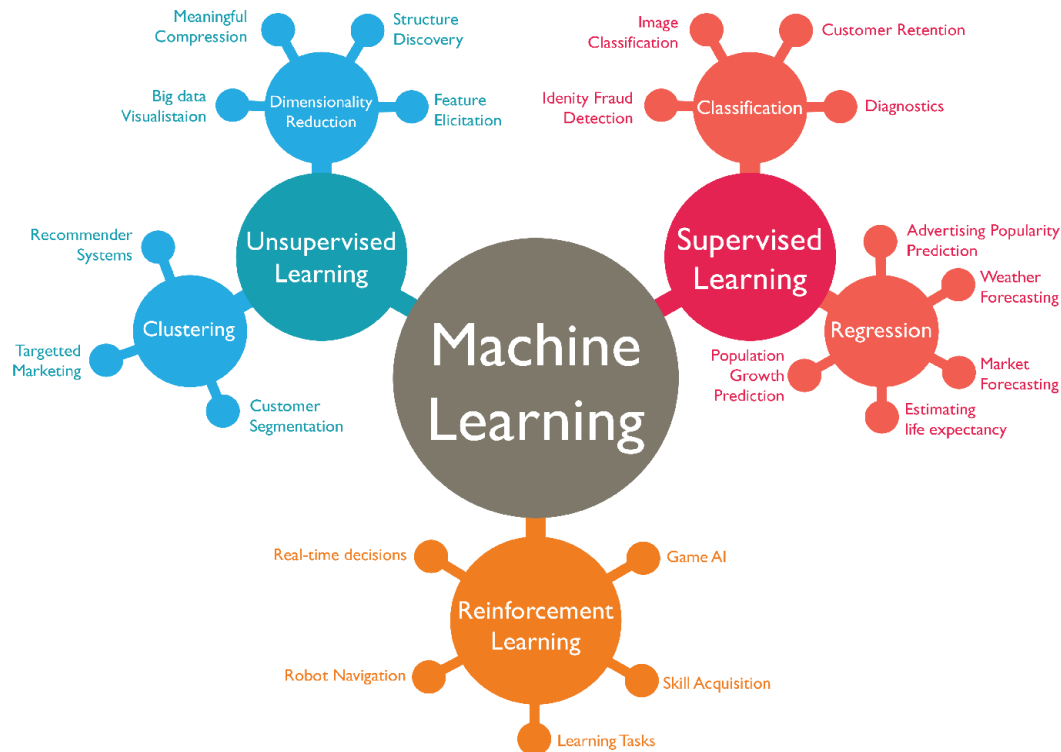
➤ 2.3 Problematização de Hipóteses e EDA

- Transformação do problema do projeto em hipóteses que devem ter passado por um processo de validação ou serem apresentadas em forma de solução com apelo de mercado ou impacto social
- Apresentação de perguntas centrais que vão nortear a solução (apresentação de 2 a 4 perguntas que, uma vez respondidas, trarão resultados tangíveis e mensuráveis)
- Apresentação da Exploratory Data Analysis com análises aprofundadas sobre o comportamento das variáveis que expliquem o problema, com foco nas hipóteses adotadas e validadas pelo mercado e/ou partes interessadas do projeto
- Apresentação de análises estatísticas com visualização de dados para identificar tendências que permitam encontrar padrões nos dados

➤ 2.4 Modelagem

- Discussão sobre as diferentes opções de modelos utilizados/testados de acordo com o objetivo do projeto e otimizações no processo (modelos supervisionados, não-supervisionados, etc)
 - As estratégias podem ser combinadas com técnicas como Stacking em conjunto com a modelagem de dados e Feature Engineering para melhorar a resposta
- Apresentação de respostas às perguntas centrais com resultados práticos, através da definição de uma ou mais métricas que auxiliarão no processo de validação
- Time series e sistemas de recomendação, serão ensinados após esse checkpoint. Em casos, favor procurar co-learnings. O grupo poderá adotar técnicas não abordadas no programa (ex.: redes neurais) caso se sinta confortável

➤ 2.4 Modelagem



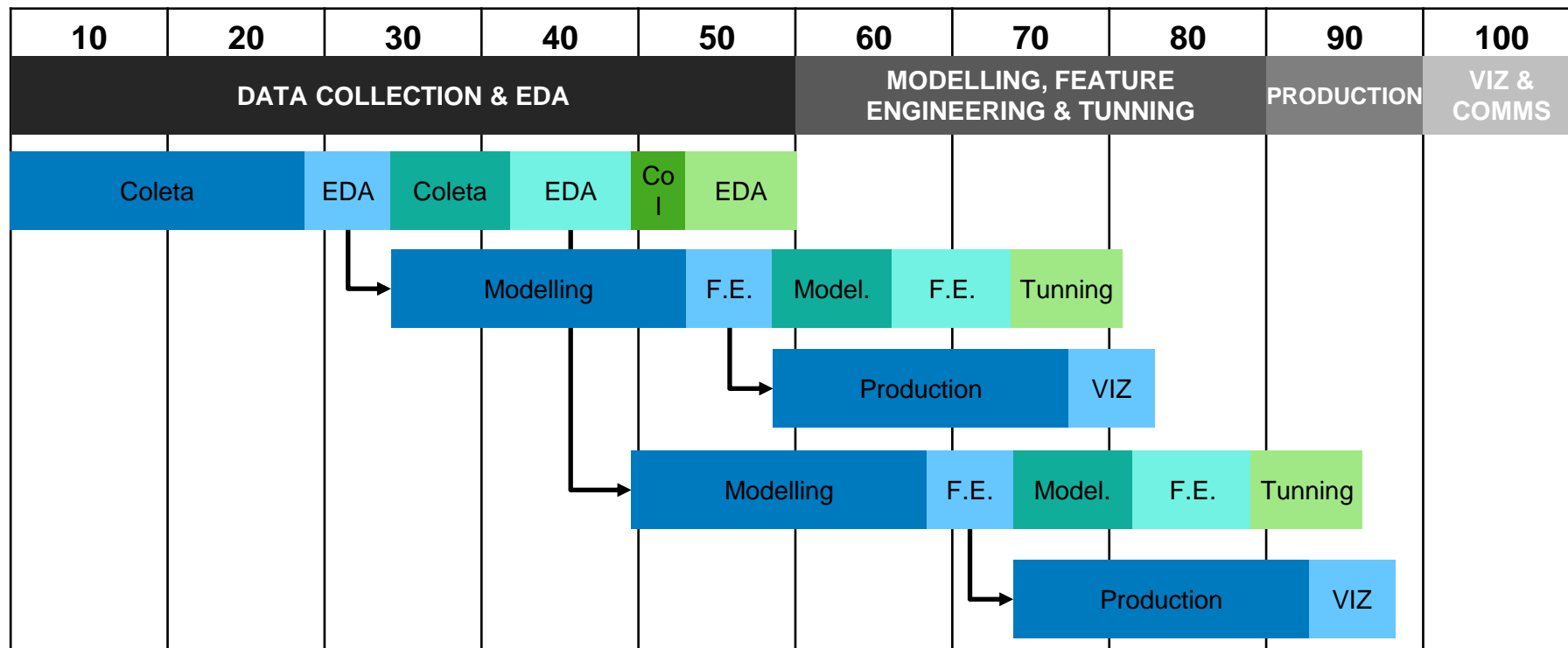
➤ 2.5 Pré-apresentação e Tunning

- É hora de fazer o primeiro teste com os colegas de sala e os professores. O objetivo é colher críticas e opiniões para permitir ajustes para a apresentação final
- A apresentação deve contemplar todos os critérios de avaliação, como uma forma de atendimento às expectativas do mercado em relação ao projeto de Data Science
- Técnicas específicas de visualização e apresentação de dados devem ser utilizadas: tabelas, gráficos diagramas, mapas, infográficos, painéis, etc.
- Adicionalmente, será feita uma discussão mais aprofundada sobre os hiper-parâmetros adotados (como melhorar as métricas de performance garantindo uma boa generalização do modelo) e outros ajustes adotados nas etapas da cadeia (feature engineering, data enrichment, etc)

➤ 2.6 Apresentação Final e MVP em Produção

- As apresentações terão duração de 15 minutos por grupo e todos os membros obrigatoriamente terão que se apresentar. A banca terá 15 minutos para comentar e questionar os membros
 - Serão realizadas em duas aulas, com a presença de todos, dias 22/03/21 e 24/03/21
- A entrega incluir: (1) o envio (geralmente em ppt) e realização de uma apresentação sobre o problema, a solução proposta e processo adotado; (2) os códigos utilizados (geralmente Jupyter)
 - A entrega de todos os materiais deve ser até às 23:59 de 19/03/21
 - Pode ser necessário navegar no código para dirimir eventuais dúvidas da banca
- A entrega também deve incluir a criação de uma aplicação real (por exemplo: API em endereço web) como forma de acesso para o público-alvo os resultados do projeto
- Estudem o perfil dos examinadores da banca e atentem para os critérios de avaliação

➤ Entregas vs processo





3. Critérios de Avaliação

➤ Critérios de avaliação

BANCA*



VALOR AGREGADO

Avalia a utilidade comercial ou social do projeto, problemas que resolve e relevância das descobertas



COMPLEXIDADE

Avalia os métodos utilizados em contraste ao que existe, bem como a amplitude com a qual esses métodos são utilizados



QUALIDADE DA APRESENTAÇÃO

Avalia a clareza na comunicação, a condução dos argumentos e o uso de recursos disponíveis (áudio, vídeo, interativos, etc)

PROFESSORES



ORGANIZAÇÃO

Avalia o cumprimento das datas e entregas estipuladas, assim como a qualidade dessas entregas



QUALIDADE TÉCNICA

Avalia boas práticas, tratamentos adotados, conclusões extraídas e a correta aplicação das técnicas estatísticas e de modelagem


PARES



CONTRIBUIÇÃO INDIVIDUAL

Avalia a contribuição de cada indivíduo para o projeto, com base na avaliação dos outros membros

* Professores e profissionais de mercado convidados



4. Cases de turmas anteriores

➤ Projeção do IDH

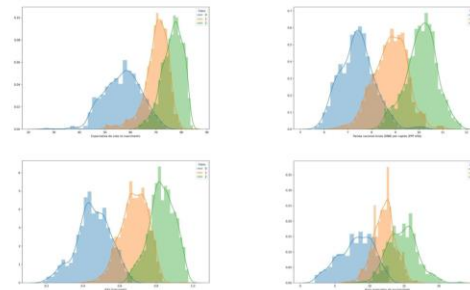
ANÁLISE DO IDH

ESTUDO DO ÍNDICE, PREDIÇÃO E PROPOSTA PARA NOVO CÁLCULO

Andre Vandor de Oliveira
Willy Taksawa
Synara Brito

Clusterização

- Análise
 - Feito isso pudemos analisar o comportamento de cada feature dada a clusterização e identificamos que o comportamento dos dados era condizentes como agrupamento proposto



➤ Previsão de furtos em tempo real

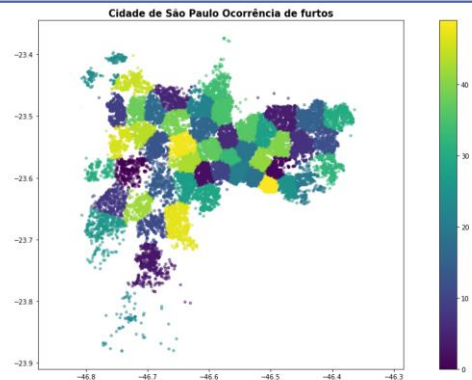


Modelo Probabilístico para Empresa de
Rastreamento de Veículos



João Procópio
Ricardo Torres
Sibelle Castro
Wagner Santana

SSP-SP – Distribuição de Furtos



➤ Previsão de rentabilidade de filmes

ACME
DATA

Grupo 2:
ANDREA FILGUERAS
CAMILA BEZERRA
FELIPE GONÇALVES
GABRIEL MONTEIRO
PAULO RAMOS ROMANSINI LOPES

Analises de Erros

	Inicial	Final
MAE	0.8105	0.5941
MAPE	0,06	0,04
R2	0.5921	0.7311

