



Insper

Ciência dos dados

Introdução à disciplina e ao Pandas

Aula de hoje

2

- Boas Vindas
- Time de professores
- Horários de atendimento e ninjas
- Aplicações
- O que é Ciência dos dados?
- Programa de ensino (conteúdo e critérios)
- Jupyter:
 - Atividade em sala: Pandas
 - Exercício: Pandas – APS1



Professores:

Maria **Kelly** Venezuela

T2A

Maria **Kelly** Venezuela

T2B

Maciel Calebe Vidal

T2C

Professor Auxiliar:

Marcio Fernando Stabile Junior

T2A

Marcio Fernando Stabile Junior

T2B

Antonio Deusany de Carvalho Junior

T2C

Ninjas:

Tales Ivalque Taveira de Freita

T2A

Isabella dos Santos de Amorim

T2B

Gabriel de Araujo Alves

T2C

Horário de Atendimento

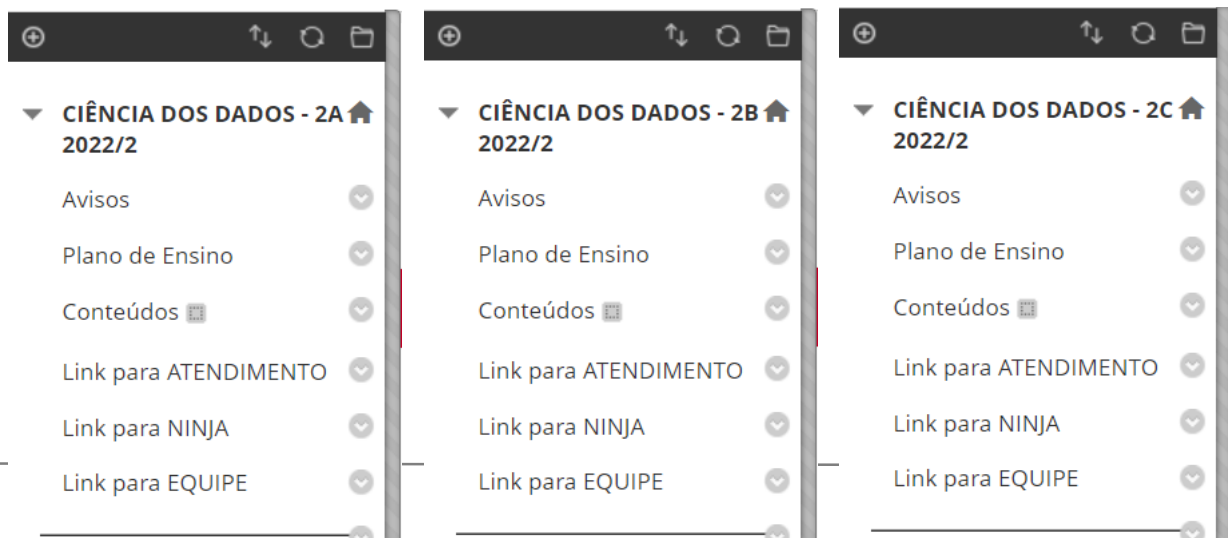
4

Turma 2A: Quintas das 10h00min às 11h30min

Turma 2B: Terças das 10h00min às 11h30min

Turma 2C: Quartas das 13h45min às 15h15min

Remoto via Teams



Teams - Atendimento e Ninjas

5

< Todas as equipes

22-2 Atendimentos e Ninjas - CD... ⋮

Caderno
Tarefas
Notas
Reflect
Insights

Canais

Geral

Atendimento - 2A
Atendimento - 2B
Atendimento - 2C
Ninja - 2A
Ninja - 2B
Ninja - 2C

Bem-vindo(a) a 22-2 Atendimentos e Ninjas - CDados
Escolha onde você deseja começar

Carregar Materiais de Aula

Configurar Caderno

Nova conversa

Aula de hoje

6

- Boas Vindas
- Time de professores
- Horários de atendimento e ninjas
- Aplicações
- O que é Ciência dos dados?
- Programa de ensino (conteúdo e critérios)
- Jupyter:
 - Atividade em sala: Pandas
 - Exercício: Pandas – APS1



Análise de dados da COVID-19 no Brasil

29/04/2021

Nowcasting de Mortes no Painel de Resultados

13/11/2020

Volta das atualizações

06/11/2020

Instabilidade nos dados do MS

13/10/2020

Cenário em 13 de Outubro

Painel Brasil

Dados atualizados em: 15/08/2021 14h17min

Total acumulado e novos casos e mortes nas 24h anteriores à atualização.
Variação percentual relativa ao dia anterior (▲|▼)

20.319.000

Casos acumulados
Confirmados

33.933

Novos casos
Confirmados

567.862

Mortes acumuladas
Confirmadas

966

Novas mortes
Confirmadas

Previsão para os próximos 7 e 14 dias

Total acumulado de casos e mortes previstos, usando modelo estatístico.
(+/-) é o erro padrão preditivo percentual.

20.551.103

+/- 1,81%

Casos acumulados
Previstos para 7 dias

20.738.920

+/- 2,75%

Casos acumulados
Previstos para 14 dias

575.003

+/- 0,75%

Mortes acumuladas
Previstas para 7 dias

581.117

+/- 1,39%

Mortes acumuladas
Previstas para 14 dias

0,88

Nº de Reprodução (R)

Número médio de pessoas contaminadas
por cada infectado. (atraso de 1 semana)

Aplicação

COVID-19 por Municípios

O Brasil é um país heterogêneo e desigual. O período quando o vírus chega e a velocidade com que se propaga em cada localidade depende de várias características: conectividade com outras localidades (nacionais ou internacionais) em diferentes estágios da pandemia, densidade demográfica, condições sanitárias, políticas adotadas localmente, entre outros fatores socioeconômicos.

Ao olhar os dados agregados para o Brasil, é plausível que a dinâmica das curvas de casos e mortes, assim como as projeções subjacentes, estejam dominadas por municípios nos quais a Covid-19 chegou mais cedo.

Visando capturar esse fenômeno de maneira mais fidedigna, dividimos os municípios em dois grupos: os 20% com maior Índice de Desenvolvimento Humano Municipal (IDHM), onde habitam 56% da população, e 80% com os demais municípios. A hipótese é que os municípios com maior IDHM já estariam em um estágio mais avançado da pandemia, por conta do fluxo de pessoas entre estes municípios e centros internacionais. A divisão é arbitrária, e ainda estamos testando outros recortes do Brasil.

O IDHM é um indicador sintético que compreende indicadores de três dimensões do desenvolvimento humano: saúde/longevidade, educação e renda. Utilizando dados de 2010 a nível municipal, o limiar que divide os dois grupos é de 0.727.

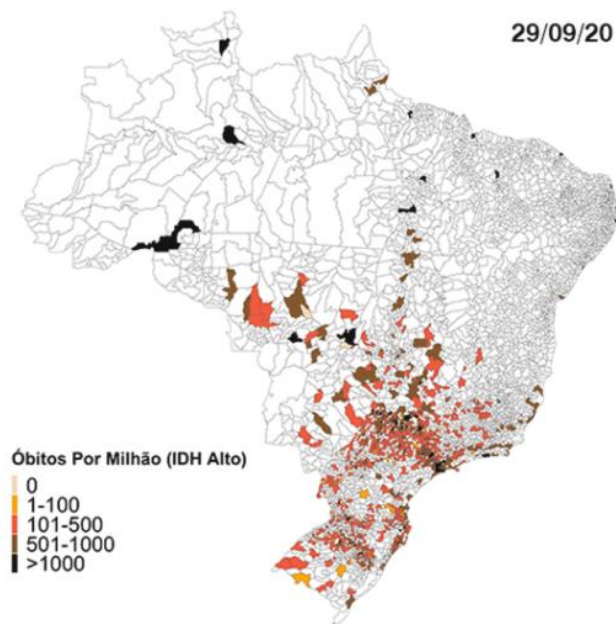


Aplicação

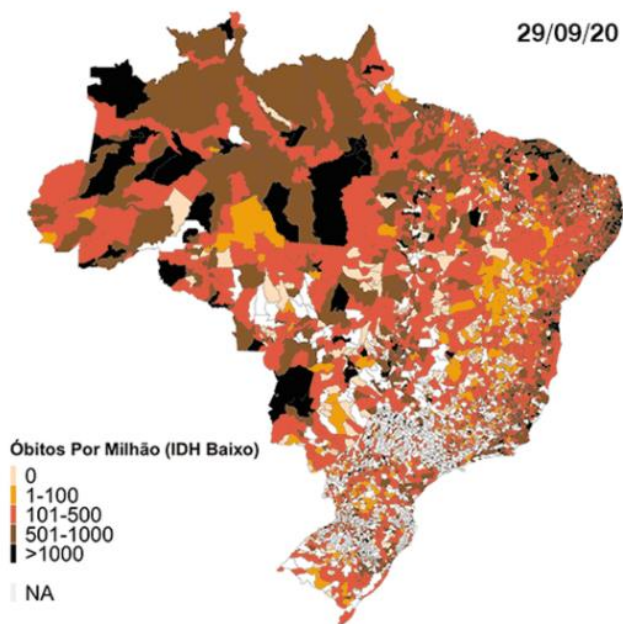
10

Óbitos por municípios no Brasil

Os mapas a seguir apresentam o total atualizado de mortes de COVID-19 por municípios. O primeiro mapa considera municípios com IDH maior que 0.727 (20% do total de municípios), ao passo que o segundo considera os demais municípios.



Óbitos COVID-19 em municípios com IDH alto



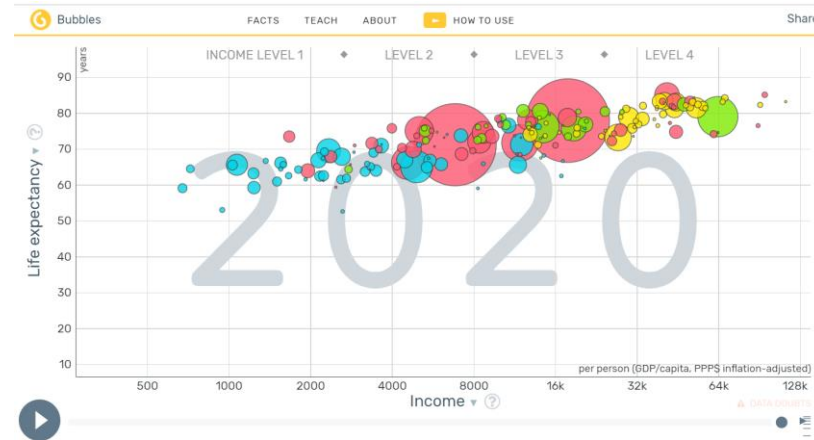
Óbitos COVID-19 em municípios com IDH baixo

Aplicação

11

Gapminder:

- Possui diversas estatísticas socioeconômicas de quase todos países do mundo que podem ser facilmente obtidas pelo link:
<https://www.gapminder.org/data/>
- Veja também:
<https://www.gapminder.org/tools/>



Aplicação

Reconhecimento facial pela polícia de SP:

- A Polícia Civil de SP adquiriu uma plataforma de reconhecimento facial
- Ferramenta vai auxiliar investigações e evitar fraudes na emissão de documentos.
- Duas das funções é detectar rostos de uma lista prévia em imagens e coletar rostos exibidos em vídeos.



Reconhecimento facial em ação
Imagem: Getty Images/123stockphoto

Texto extraído de <https://www.uol.com.br/tilt/noticias/redacao/2019/11/15/big-brother-urbano-como-vai-funcionar-o-reconhecimento-facial-em-sao-paulo.htm>

Pizza de Dados

20
APR

Episódio 020: Estatística e Machine Learning

April 20, 2019

Chamamos uma Estatística e Engenheira de Aprendizado de Máquina pra contar pra gente como ela passou de Ciências Sociais para Cientista de Dados no Ifood! No episódio de hoje a querida Júlia Tessler conta as aventuras de uma estatística trabalhando como cientista, faz uma ode de amor à Regressão Linear e fala sobre o dia a dia trabalhando com dados de Pizza!

Agradecimento especial aos nossos Parceiros

Esse episódio não seria possível sem o apoio especial dos nossos parças do Data Bootcamp, o maior bootcamp de Data Science do Brasil! Aprenda a organizar, extrair e interpretar os dados da sua empresa com as tecnologias mais avançadas usadas no mercado. Confira as datas dos próximos cursos em <https://databootcamp.com.br/calendar>.



171,218

Downloads

28

Episodes

+ Follow

et Share



O primeiro podcast Brasileiro sobre ciência de dados

Search...



Exemplos de projetos de alunos na disciplina

- O que vou achar de um filme da Netflix que ainda não assisti?
- O acesso à cultura de um aluno ajuda a prever sua nota do ENEM?
- A raça/cor de um indivíduo suspeito é levada em consideração na abordagem de um policial americano?
- O valor de Mercado de jogadores de futebol se explica só pelo que fazem em campo?

O que é Ciência dos dados?

Objetivos / Problemas:

- Associação
- Causa e efeito
- Previsões
- Identificar tendências

Uso:

Situações que não sejam determinísticas, fazemos uso de modelos probabilísticos considerando conjunto de dados (amostra).



Aula de hoje

16

- Boas Vindas
- Time de professores
- Horários de atendimento e ninjas
- Aplicações
- O que é Ciência dos dados?
- Programa de ensino (conteúdo e critérios)
- Jupyter:
 - Atividade em sala: Pandas
 - Exercício: Pandas – APS1

Disciplina: Conteúdo

O que teremos e faremos, neste semestre?

Objetivos de aprendizado

Ao final do semestre, o aluno deverá ser capaz de:

- Elaborar **análises exploratórias de dados** (univariadas e multivariadas), utilizando **ferramentas estatísticas e computacionais adequadas**;
- Especificar as **distribuições de probabilidades** adequadas para as variáveis quantitativas discretas e contínuas;
- Conduzir **testes inferenciais** adequados que possam dar base à tomada de decisão; e
- **Analisar relações entre as variáveis**, utilizando ferramentas estatísticas inferenciais adequadas.

Disciplina: Critério

O que teremos e faremos, neste semestre?

APS – Atividade Prática Supervisionada

20

DATA INÍCIO	DEADLINE às 23h59	APSs
16/08 no dia da Aula 1	25/08 no dia da Aula 4	APS1: Manipulação de df com pandas
30/08 no dia da Aula 5	01/09 no dia da Aula 6	APS2: Explorando Duas Variáveis Quantitativas Deadline APS1
01/09 no dia da Aula 6	15/09 no dia da Aula 10	APS3: Teoria da probabilidade Deadline APS2
08/09 no dia da Aula 8	08/09 no dia da Aula 8	APS4: Teorema de Bayes com Texto Deadline APS4 (final da aula)
15/09 no dia da Aula 10	27/09 no dia da Aula 13	APS5: Variáveis aleatórias discretas Deadline APS3
20/09 no dia da Aula 11	27/09 no dia da Aula 13	APS6: Modelos probabilísticos discretos
13/10 no dia da Aula 17	25/10 no dia da Aula 20	APS7: Modelos probabilísticos contínuos
20/10 no dia da Aula 19	27/10 no dia da Aula 21	APS8: Propriedades de Esperança e Variância com soma de v.a.'s
10/11 no dia da Aula 25	17/11 no dia da Aula 26	APS9: Teste de Hipóteses para Média Populacional
22/11 no dia da Aula 27	22/11 no dia da Aula 27	APS10: Regressão Linear Deadline APS10 (final da aula)

APS – Atividade Prática Supervisionada

21

Durante o semestre, teremos:

- Os enunciados dos exercícios são disponibilizados no github e os testes para validação das respostas são liberados no Blackboard.
- Os testes poderão ser refeitos indeterminadas vezes até que expire prazo de entrega.

IMPORTANTE:

- ✓ É preciso ter **pelo menos 80% de acerto em cada APS** para a mesma ser considerada como **entrega satisfatória**.
- ✓ Ainda, é preciso ter metade ou mais de APSs com **entrega satisfatória**.
- ✓ Caso não haja esses dois quesitos acima validados, **o aluno será automaticamente reprovado na disciplina, independente das demais notas.**
- ✓ **Não serão aceitos possíveis APSs com atraso.**

Avaliações

DATA	AVALIAÇÕES
Início 06/09 às 18h00 até dia 11/09 às 21h	QUIZ 1 (Aulas 01 a 05) via Blackboard
29/09 ou 04/10 (ver Calendário Insper)	AVALIAÇÃO INTERMEDIÁRIA (AI)
Início 01/11 às 18h00 até dia 06/11 às 21h	QUIZ 2 (Aulas 16 a 20) via Blackboard
01/12 ou 06/12 (ver Calendário Insper)	AVALIAÇÃO FINAL (AF)

IMPORTANTE:

- ✓ As datas das avaliações, na semana de provas, seguem o calendário do Insper.
- ✓ A Avaliação Substitutiva irá englobar **todo o conteúdo** e deverá substituir apenas **uma das DUAS avaliações (AI ou AF)**.
- ✓ Caso falte em mais do que uma avaliação, a nota da Avaliação Substitutiva será usada apenas na avaliação de maior peso.
- ✓ Não há prova substitutiva para os QUIZZES, a menos que haja justificativa médica para todo o período que os mesmos ficarão disponíveis no Blackboard.

DATA	PROJETO
30/08 no dia da Aula 5	Início Projeto 1
13/09 no dia da Aula 9	Aula Estúdio Projeto 1
22/09 no dia da Aula 12	FIM Projeto 1 às 23h59
18/10 no dia da Aula 18	Início Projeto 2
29/11 no dia da Aula 28	Aula Estúdio Projeto 2
29/11 no dia da Aula 28	FIM Projeto 2 às 23h59

IMPORTANTE:

- ✓ Todos os projetos devem ser entregues e nenhum deles pode ser considerado com conceito I.

Notas NA e NP

NOTA DAS AVALIAÇÕES - NA

NOME DA AVALIAÇÃO	SIGLA	PESO EM %
QUIZ 1	Q1	10
AVALIAÇÃO INTERMEDIÁRIA	AI	30
QUIZ 2	Q2	15
AVALIAÇÃO FINAL	AF	45

NOTA DOS PROJETOS - NP

NOME DA AVALIAÇÃO	SIGLA	PESO EM %
MÉDIA APSs	APS	10
PROJETO 1	P1	45
PROJETO 2	P2	45

IMPORTANTE:

- ✓ Na média das APSs, serão consideradas todas as APSs disponibilizadas.
- ✓ Para os projetos, será utilizada a tabela oficial do Blackboard para converter de conceito para nota numérica.

Nota Final da disciplina

A nota final da disciplina será calculada da seguinte forma:

- **média(NA, NP)**, se **NA** e **NP** forem maiores ou iguais a 5 simultaneamente;
- **min(NA, NP)**, caso contrário

IMPORTANTE:

- ✓ Para os projetos, será utilizada a tabela oficial do Blackboard para converter de conceito para nota numérica.
- ✓ É preciso ter pelo menos 50% das APSs entregues para o critério da nota final acima ser aplicada; caso contrário, aluno será reprovado.

- MONTGOMERY, D. Estatística Aplicada e Probabilidade para Engenheiros (6a edição). LTC, 2016.
- MAGALHÃES, M. N.; DE LIMA, A. C. P. Noções de Probabilidade e Estatística (7a edição). Edusp, 2013.
- GRUS, J. Data Science do Zero: Primeiras Regras com Python. Alta Books, 2016.

- Blackboard (material e avisos)
- Github (jupyter notebook e base de dados)

<https://github.com/Insper/CD22-2>

Atenção: clonar repositório até semana que vem.

- Lista de exercícios (além dos exercícios das aulas)

Conteúdo no Blackboard

28

CIÊNCIA DOS DADOS - 2A 2022/2

Conteúdos

CIÊNCIA DOS DADOS - 2A 2022/2

Avisos

Plano de Ensino

Conteúdos

Link para ATENDIMENTO

Link para NINJA

Link para EQUIPE

Notas

Enviar e-mail

Central de ajuda

Gerenciamento do curso

Painel de controle

Coleção de Conteúdo

Ferramentas do curso

Avaliação

Centro de notas

Usuários e grupos

Personalização

Pacotes e utilitários


Ajuda


Conteúdos


Criar conteúdo


Avaliações

Ferramentas

 GitHub de Ciência dos dados

 AULAS

 APSS

 LISTAS DE EXERCÍCIOS

er

Bibliotecas Python para *Data Science* 29

- Bibliotecas Principais e Estatísticas:
Numpy; SciPy; Pandas; StatsModels
- Visualização:
Matplotlib; Seaborn; Plotly; Bokeh; Pydot
- Machine Learning
Scikit-learn; XGBoost; Eli5
- Deep Learning
TensorFlow; PyTorch; Keras
- Consulte sobre essas e outras bibliotecas em:
<http://datascienceacademy.com.br/blog/top-20-bibliotecas-python-para-data-science/>

Notebook Atividade – em aula

Manipulando base de dados reais:

- Download pelo Blackboard (a partir da semana que vem, será apenas pelo GitHub)
- Fazer juntos e discutir em sala

Notebook Exercício – APS1

Manipulando base de dados reais:

- Download pelo Blackboard (a partir da semana que vem, será apenas pelo GitHub)
- Fazer individual e discutir na mesa
- Submeter INDIVIDUAL via Blackboard (APS1)

Próxima aula...

Leitura prévia necessária:

- Montgomery & Runger, seções 6.1 e 6.3
- Magalhães e Lima (7ª. Edição): Cap. 1 – destaque para tipos de variáveis, tabelas e gráficos para variáveis qualitativas.