


FIAP

Statistics for Data Science & Machine Learning

Aula 1

Prof. Jones Egydio

profjones.egydio@fiap.com.br



Objetivos

- Apresentação da disciplina;
- Apresentação pessoal;
- Bibliografia da disciplina;
- Organização do curso;
- Avaliações;
- Exemplos com Python;
- Conclusão;
- Perguntas.

Quem sou eu

Formação Acadêmica

- **USP – Universidade de São Paulo (2023 – em andamento).**

Pós-graduação Stricto Sensu – Doutorado em Engenharia Elétrica.

- Área de pesquisa: Engenharia de Sistemas e Inteligência Artificial.

- **UFABC – Universidade Federal do ABC (2014).**

Pós-graduação Stricto Sensu – Mestrado em Engenharia Elétrica.

- Dissertação: Análise de Interferências Eletromagnéticas em Veículos Automotores utilizando Métodos Numéricos.

- **PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul (2019).**

Pós-graduação Lato Sensu – Filosofia e Autoconhecimento.

- **USCS – Universidade de São Caetano do Sul (2011).**

Pós-graduação Lato Sensu – Docência no Ensino Superior.

- Pesquisa: Uso de Recursos Tecnológicos como Ferramenta de Aprendizagem em Disciplinas de Matemática no Curso de Engenharia.

- **IMT – Instituto Mauá de Tecnologia (2002 – 2007).**

- Graduação em Engenharia Eletrônica.

- **Centro Universitário Fundação Santo André (1998 – 2001).**

- Graduação em Bacharelado em Matemática.

Experiência profissional

Itaú – MLOps Engineer (fev./2024).

FIAP e Mackenzie – Professor (fev./2024).

IMT – Instituto Mauá de Tecnologia (fev./2008 – dez/2023).

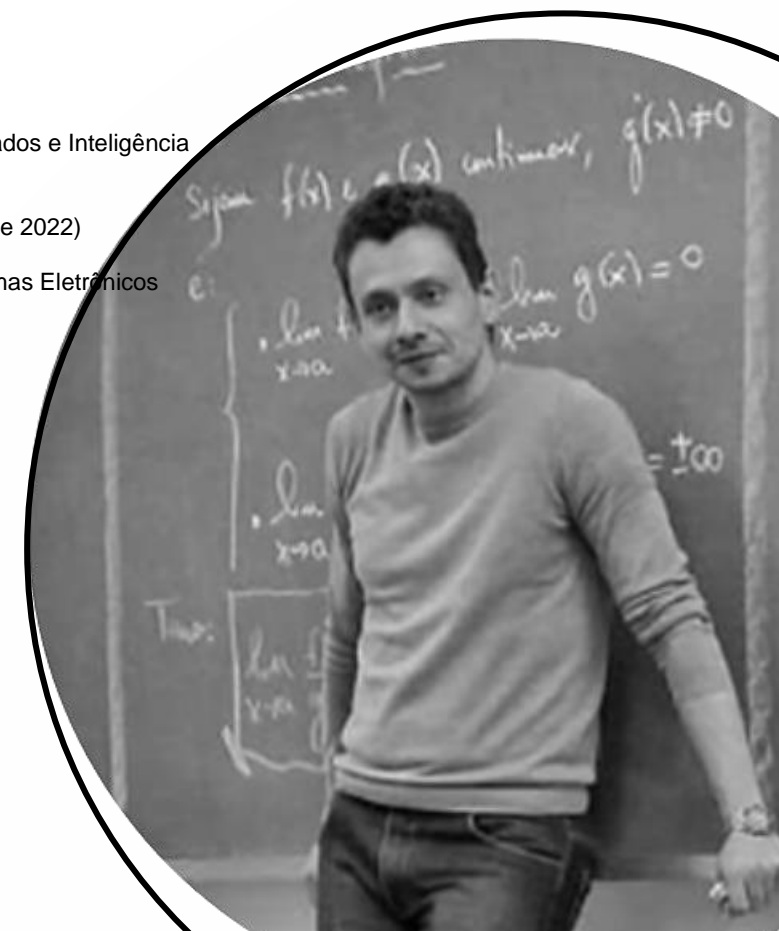
- Professor assistente – Departamento Fundamental
- Coordenador do Programa de Pós-Graduação em Ciência de Dados e Inteligência Artificial (desde 2022)
- Coordenador do Programa de Minor em Ciência de Dados (desde 2022)
- Professor e pesquisador do Laboratório NSEE (Núcleo de Sistemas Eletrônicos Embarcados) (desde 2022)
- Orientação de alunos em Trabalhos de Conclusão de Curso, participação em projetos envolvendo a FAPESP, Faculdade de Saúde Pública da USP e Fundação Oncocentro do Estado de São Paulo;
- Departamento Fundamental (desde 2008)
 - Preparação e ministração de aulas para os cursos de Engenharia da Computação, Sistemas da Informação e Ciência da Computação do Instituto Mauá de Tecnologia.
 - Preparação e ministração de aulas para o curso de Cálculo Diferencial e Integral, Estatística, Matemática Computacional (Cálculo Numérico) e Vetores, Curvas e Superfícies (Geometria Analítica) e Álgebra Linear do Instituto Mauá de Tecnologia;

GM do Brasil (jan./2008 – set./2019).

- Engenheiro de desenvolvimento de produto
- Engenheiro de simulação elétrica (desde nov./2010);

[LinkedIn](#)

[Lattes](#)



Por que matemática?

- Fundação Teórica
- Modelagem e Simulação
- Análise de Dados
- Algoritmos e Otimização
- Inovação e Pesquisa
- Inteligência Computacional
- Comunicação e Padrões Universais

21/02/2024

Profissões ligadas à matemática equivalem a 4,6% do PIB





Objetivos do curso

“Capacitar o aluno no entendimento de conceitos básicos de estatística e análise de dados.”

“Preparar os alunos para entender e desempenhar conceitos futuros relacionados à Análise Exploratória de Dados e Machine Learning.”

Bibliografia básica do curso

- 1. SINICH, M.B. Estatística para Administração e Economia. 10ªed. São Paulo: Pearson Education do Brasil, 2009.*
- 2. SWEENEY, D.J.; WILLIAMS, T.A.; ANDERSON, D.R. Estatística aplicada à administração e economia, 2015.
- 3. FÁVERO, L.P. Análise de dados; modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 200

Bibliografia do curso – complementar

- 1. CASTANHEIRA, N.P. Cálculo aplicado à gestão e aos negócios. 1ºed.Curitiba: Editora Intersaberes, 2016.*
- 2. GROLEMUND,G; WICKHAM,h. R for Data Science. 1ªED. O´Reilly. Disponível em; <https://r4ds.had.co.nt/>,2015.
- 3. McKinney,W. Phyton for data analysis. 1ªed. O´Reilly. Disponível em: <http://wesmckinney.com/pages/book.html>,2013
- 4. JAMES,G.,et al. An Introduction to Statistical Learning with applications in R (free book). 1ª e. Springer. Disponível em <http://bit.ly/Gardeen>, 2014.*
- 5. LARSON,R.;Farber,B. Estatísticas Aplicada.6ªed. São Paulo; Pearson Education do Brasil, 2015.*



Conteúdo Programático Anual

- Conceitos estatísticos: discutir a coleta, organização e representação de dados numéricos utilizando linguagem Python.
- Distribuições probabilísticas: conceitos introdutórios sobre as principais distribuições de probabilidade alusivas aos modelos de Machine Learning.
- Medidas descritivas: apresentar algumas técnicas introdutórias que darão forte embasamento para etapas futuras de preparação e análise exploratória de dados.



Metodologia

- Aulas com objetivos específicos.
- Conteúdo proposto associado a sua aplicação no dia a dia.
- Integração entre as disciplinas.
- Preparação para o mercado de trabalho e aos projetos propostos pela FIAP.
- Exercícios em grupos, como "mini projetos"

Avaliação

As notas semestrais na FIAP são compostas:

- 40% Project Checkpoint Challenge&Feedback (**2 Challenges** + **3 Checkpoints**)
- 60% Global Solution (solução de tarefas de Cases reais)

$$\text{MS1} = (\text{PCC\&F} \times 0.4 + \text{GS} \times 0.6)$$

A média anual é ponderada, ou seja, os semestres possuem pesos diferentes:

$$\text{MA} = (\text{MS1} \times 0.4 + \text{MS2} \times 0.6)$$

Avaliação

CRITÉRIOS DE APROVAÇÃO

Média Anual	Situação
0 a 3.9	Reprovado
4.0 a 5.9	Exame
6.0 a 10	Aprovado

CASO O ALUNO FIQUE DE EXAME:

Nota para aprovação = $(12 - \text{Média Anual})$



Calendário

JUNHO

**27/05
a 07/06**

Período de avaliação semestral
(Global Solutions).

10 a 14

Período de avaliação semestral de disciplinas
de dependência.

10 a 14

Período de solicitação de avaliações
substitutivas regulares e de dependência.

17 a 21

Período de avaliações substitutivas
regulares e de dependência.

24 a 28

Vistas de provas

28

Final das fases / divulgação dos resultados
das avaliações semestrais

Em resumo

Global Solution
Final do
semestre

1



2

Sprint
entregas

3

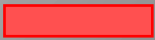
Checkpoint
Entregas durante
o semestre



Avaliação

Calendário (sujeito a alteração) 1º semestre:

- CP1: TBD
- CP2: TBD
- CP3: TBD



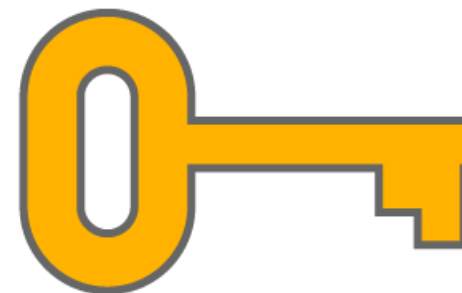
Os pioneiros na Ciência de Dados

- JOHN ECKERT E JOHN MAUCHLY, OS CRIADORES DO PRIMEIRO COMPUTADOR DO MUNDO;
- WILLIAM CLEVELAND E O TREINAMENTO DOS CIENTISTAS DE DADOS;
- ADA LOVELACE, A PRIMEIRA PROGRAMADORA;
- O TEAR MECÂNICO DE JOSEPH-MARIE JACQUARD
- ALAN TURING, O PAI DA CIÊNCIA DA COMPUTAÇÃO

Fonte: <https://blog.voomp.com.br/graduacao/ciencia-de-dados/6-figuras-importantes-para-a-historia-da-ciencia-de-dados>

Por que dados são tão importantes?

- **DADOS:** Símbolos que apresentam propriedades de objetos e eventos . São obtidos por coletas de pesquisa e armazenam em algum repositório
- **INFORMAÇÃO:** São dados processados de forma a torná-los úteis para uma determinada ação, tarefa ou tomada de decisão. Busca a conexão dos dados com a realidade.
- **CONHECIMENTO:** é a interação com a informação e a realidade (a realização da tarefa ou tomada de decisão). É realizar as instruções com propriedade e responder as questões com propriedade.
- **SABEDORIA;** Combinação de informação, conhecimento e entendimento de forma trazer valor a algo, fazer crescer ou fazer se desenvolver. Obtida por meio da reflexão, em exercício para o pensamento.



DESAFIOS do Big Data





Desafios

- DEMANDA CONTINUA CRESCENTE NOS PRÓXIMOS ANOS
- Faltam profissionais para iniciarem as demandas ou identificar aplicações.
- TRAZER VALOR DE SOLUÇÕES:
- Foco na solução, benefícios e não nas tecnologias.
- PREPARAÇÃO MÃO DE OBRA
- Nas empresas: qual é o mais difícil? Matemática/Estatística, Programação ou Negócios.
- Nas universidades e escolas: adaptação de ementas, professores qualificados

Habilidades

MACHINE LEARNING

MUST HAVE SKILLS

BASIC
LINEAR
ALGEBRA

CODING
KNOWLEDGE

STATISTICS
AND
PROBABILITY

ALGORITHMS
USAGE

UNIX
TOOLS

Mercado de trabalho

Average Salary for Skill: Apache Hadoop

Salary	Hourly Rate	Bonus
Apache Hadoop Median Salary by Job		
More Charts ▼		
Job	Average	
Data Scientist 242 profiles	\$97,729	
Data Engineer 182 profiles	\$99,999	
Data Scientist, IT 139 profiles	\$92,417	
Data Scientist / Engineer 76 profiles	\$91,652	

https://www.payscale.com/research/US/Skill=Apache_Hadoop/Salary

Average Salary for Skill: Apache Spark

Salary	Hourly Rate	Bonus
Apache Spark Median Salary by Job		
More Charts ▼		
Job	Average	
Data Scientist 304 profiles	\$101,798	
Senior Data Scientist 58 profiles	\$128,780	
Data Engineer 126 profiles	\$101,042	

https://www.payscale.com/research/US/Skill=Apache_Spark/Salary

Average Salary for Skill: R

Salary	Hourly Rate	Bonus
R Median Salary by Job		
More Charts ▼		
Job	Average	
Data Scientist 1142 profiles	\$88,323	
Data Scientist, IT 237 profiles	\$86,624	
Senior Data Scientist 183 profiles	\$122,046	

<https://www.payscale.com/research/US/Skill=R/Salary>

Ferramentas utilizadas

RDBMS/NOSQL	CLOUD	SCRIPTING	DEVOPS	MONITORING	VISUALIZATION
Mongodb	Oracle cloud	R/python (dataframes)	Jenkins	Prometheus	Grafana
Oracle nosql	Azure cloud	Ansible	Puppet	Zabbix	Tableau
Cassandra	Aws	Dsl language	Chef	Nagios	Kibana
Elasticsearch	Google platform	Cloudformation/terraform	Agile tools		Jupyter notebooks
Apache kafka	Openstack	Shell	Terraform		Zeppelin notebooks
Postgres	Database as service	Powershell	Any ci/cd tools		
Bigdata	Migration tools	Rest			

Assuntos que serão abordados

FUNDAMENTOS	TÉCNICAS DE DATA DE SCIENCE	OBJETIVOS
Funções, integrais, probabilidade	Distribuições de probabilidade	Definir a regra de probabilidade de dados
Funções, média, variância, desvio-padrão, quartis, integrais	Teste de hipóteses	Testar hipóteses
Funções, média, variância, desvio-padrão, quartis	Análise exploratória de dados	Entender o desenho geral dos dados
Funções, derivadas	Regressão linear	Explicar uma variável dependente contínua por meio de variáveis independentes
Funções, derivadas	Regressão logística	Explicar uma variável dependente binária por meio de variáveis independentes
Funções, derivadas	Árvores de decisão	Explicar uma variável dependente por meio de variáveis independentes criando regras
Média, variância, covariância e correlação	Técnicas de redução de dimensão	Criação de índices
Distância euclidiana	Técnicas de agrupamentos (cluster)	Descobrir indivíduos semelhantes num banco dados

Perguntas?

Obrigado!