Data Science

TechGuide - Alura, FIAP e PM3

Data Science

Nivel 1

- [ ] **Data Science - Fundamentals**:
  - Data Science is the act of using algorithms and systems to extract, organize, and analyze data from various sources in order to detect patterns and make business decisions. The application areas are endless, such as in business, biology, medicine, engineering, etc.
  - Knowing the concept of Data Mining
  - Learning the main functions, such as 'describe', 'info', etc
  - Understanding the role of visualizations such as histograms and boxplots
  - Knowing what nominal and ordinal categorical variables are
  - Exploring data in Python with Pandas, Matplotlib, Seaborn and Numpy libraries
- [ ] **Feature Engineering**:
  - Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling, in order to improve the performance of Machine Learning algorithms.
  - Generating new variables from the available data
  - Transforming raw data into characteristics

- Highlighting problems with predictive models
- Improving model accuracy for new data
- Selecting and creating features using Pandas and Scikit-learn libraries

- [ ] **Data Extraction and Processing**:

- Data extraction is the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely unstructured.
- Obtaining the data to be analyzed
- Treating the obtained data, transforming it, changing its structure and values in order to make the database more coherent and ensure that the data to be worked on is in the best conditions to be analyzed
- Using the Pandas and Scikit-learn libraries to treat the data

- [ ] **Python for Data Science**:

- Python is a high-level interpreted programming language that supports multiple programming paradigms, such as imperative, object-oriented, and functional. It is a language with dynamic typing and automatic memory management.
- Learning Python programming logic
- Learning the language fundamentals such as variables, functions, lists, conditionals and imports
- Creating data analyses
- Using Matplotlib to generate graphs
- Using and manipulating lists to group data
- Knowing the NumPy library
- Knowing the Pandas library

- [ ] **Jupyter & Colab notebooks**:

- Jupyter Notebook and Google Colaboratory are Notebooks that allow the creation of text blocks and code blocks

- Notebooks make it easy to write Data Science projects, because you can see the result of the execution right after the code snippet
- Google Colaboratory allows you to write and execute Python code directly in the browser, with little to no configuration required
- These tools make it easy to share projects among the team

- [ ] **R for Data Science**:

- R is a programming language commonly used in computational statistics and data analysis.
- Learning to analyze, clean and visualize data
- Creating graphs
- Joining databases
- Drawinging statistical summaries
- Learning predictive models in R

- [ ] **Statistics and Math - Fundamentals**:

- Equations, Functions, and Limits
- Logarithms
- Matrices, determinants, vectors and vector space
- Derivatives and integrals
- Difference between mean, median and mode
- Frequency distribution
- Variance and standard deviation
- Binomial, Poisson and normal distributions
- Confidence level and confidence interval
- Sampling techniques
- Linear regression
- Time series

- [ ] **Data Visualization**:

- Data visualization is a contemporary expression of visual communication that consists of the visual representation of data.
- Mapping abstract data into visual representations
- Visually representing data that is present in our real world
- Using Python, Matplotlib and Seaborn to generate data visualizations

Nivel 2

- [ ] **Machine Learning - Fundamentals**:

- Machine Learning is a subfield of Engineering and Computer Science that evolved from the study of pattern recognition and the theory of computational learning in artificial intelligence.
- Supervised Learning
- Using classification algorithms
- Using regression algorithms
- Using Scikit-learn to create machine learning models

- [ ] **Machine Learning - Unsupervised Learning**:

- Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.
- Clustering is the set of data mining techniques that aim to automatically group data according to their degree of similarity.
- Knowing exploratory analysis
- Using K-means, DBSCAN and Mean shift methods to cluster unclassified data
- Evaluating the quality of a clustering
- Parameterizing clustering methods using the maximum silhouette coefficient
- Understanding the mathematics behind the Silhouette, Davies Bouldin and Calinski Harabasz validation metrics

- Knowing dimensionality reduction techniques

- [ ] **Machine Learning - Model Evaluation**:
  - Model Evaluation is the process that uses metrics to help us analyze whether a trained model will perform well in predicting when exposed to new data sets.
  - Knowing different strategies for model evaluation and optimization
  - Using a pipeline for training and validation
  - Knowing the main metrics for evaluating machine learning models
- [ ] **SQL - Fundamentals**:
  - Structured Query Language (SQL) is a standardized programming language that is used to manage relational databases and perform various operations on the data in them.
  - Knowing the most common SQL commands
  - Using SELECT to query a table
  - Using INSERT to insert data into a table
  - Using UPDATE to update a table
  - Using DELETE to remove data from a table
  - Using JOIN to connect data from multiple tables
  - Knowing the clauses (FROM, ORDER BY, etc.)
- [ ] **Web Scraping**:
  - Web scraping or web data scraping is the extraction of data from websites.
  - Using BeautifulSoup and Python to collect data
  - Searching and browsing the HTML
  - Accessing the content and attributes of HTML tags
  - Building datasets with scraping results
- [ ] **Statistical Tests**:

- Statistical tests are used to examine relationships between variables and test hypotheses.
- Creating confidence intervals for samples
- Comparing groups of samples
- Performing statistical tests
- Designing experiments to collect data
- Proposing mathematical models to understand a given problem
- Building color maps to aid data interpretation

Nivel 3

- [ ] **Deep Learning**:

- Deep Learning is a branch of Machine Learning based on a set of algorithms that attempt to model high-level abstractions of data using a deep graph with multiple processing layers, composed of various linear and nonlinear transformations.
- Building and training models with Keras
- Building and training models with Tensorflow
- Selecting the layers of a model
- Classifying images
- Understanding the concepts of weights and biases
- Neural networks for regression
- Understanding the concept of recurrent networks

- [ ] **Reinforcement Learning**:

- Reinforcement Learning is an area of Machine Learning that is concerned with how intelligent agents should take action in an environment in order to maximize the notion of cumulative reward.
- Understanding the concepts of agent and reward
- Understanding the difference between positive and negative reinforcement

- Knowing the Markov Decision Process model
- Understanding the concept of Feedback
- Using the Q-learning algorithm

☐ **Computer Vision**:

- Computer Vision is an interdisciplinary scientific field that deals with how computers can gain high-level knowledge from digital images or videos. From an engineering perspective, it seeks to understand and automate tasks that the human visual system can do.
- Extracting regions of interest from an image
- Normalizing and pre-processing image data
- Building classifiers for face recognition
- Extracting regions of the human face based on facial landmarks
- Analyzing different conditions of each component of the human face
- Using OpenCV

☐ **Natural Language Processing**:

- Natural Language Processing (NLP) is a subfield of artificial intelligence and linguistics that studies the problems of automatic generation and understanding of natural human languages.
- Sentiment Analysis
- Creating visualizations to facilitate the analysis of textual data
- Knowing the NLTK and Scikit-Learn libraries
- Normalizing texts
- Using TF-IDF and Ngrams to improve classification
- Using SKlearn
- Using Regex in PLN
- Knowing Word2Vec
- Combining word vectors to represent texts and classify them

Habilidade Auxiliar: Cloud, Big Data and Systems

- [ ] **Big Data and Data Engineering**:
    - Big data is the area of knowledge that studies how to process, analyze, and derive information from data sets that are too large to be analyzed by traditional systems.
    - Data engineering involves the development and organization of processes for collecting, storing, and transforming large-scale data.
    - Creating a data pipeline
    - Knowing Apache Airflow
    - Knowing Apache Spark
    - Knowing the Data Lake concept
    - Interacting with servers in the cloud
    - Creating Spark applications
- [ ] **Cloud - Fundamentals**:
    - Cloud, or cloud computing, is the distribution of computing services over the Internet using a pay-as-you-go pricing model. A cloud is composed of various computing resources, ranging from the computers themselves (or instances, in cloud terminology) to networks, storage, databases, and everything around them. In other words, everything that is normally needed to set up the equivalent of a server room, or even a complete data center, will be ready to use, configured, and run.
    - Knowing the difference between IaaS, PaaS and SaaS
    - Knowing the largest cloud providers
    - Specializing in a specific provider of your choice
- [ ] **Git & GitHub - Fundamentals**:
    - Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

- GitHub is a hosting service for software development and version control using Git.
- Creating a repository
- Cloning a repository
- Committing, pushing and pulling to and from the repository
- Reversing a commit
- Creating branches and pull requests
- Handling merge and conflicts

☐ **Linux - Fundamentals**:

- Linux is a term popularly used to refer to operating systems that use the Linux Kernel. Distributions include the Linux Kernel as well as system software and libraries.
- Knowing the Linux directory system
- Compacting and uncompressing files
- Editing files in the console with VI
- Managing the processes running on the machine
- Knowing the environment variables and PATH
- Managing packages
- Performing remote communication with SSH and SCP

Habilidade Auxiliar: Business

☐ **Business Process Management**:

- Business Process Management (BPM) is a discipline that uses various methods to discover, model, analyze, measure, improve, optimize, and automate business processes.

☐ **Business Intelligence (BI) - Fundamentals**:

- Business Intelligence is a set of theories, methodologies, processes, and technologies that enable the transformation of "raw" data into highly

relevant information for a company's decision making.

- Knowing the ETL process
- Performing modeling and structuring of tables in a Data Warehouse
- Creating visualizations that make sense
- Knowing PowerBI

- [ ] **Storytelling with data**:

- Storytelling is a way of telling stories that engage and grab the attention of the person listening. Within data analysis, it is very important to convey information to the receiver in a way that he or she understands not only the data, but also the whole context.

- [ ] **Excel**:

- Microsoft Excel is a spreadsheet editor produced by Microsoft with calculation and table building tools.
- Performing the basic mathematical operations with its operators (addition, subtraction, multiplication and division)
- Knowing the main formulas, such as 'AVERAGE', 'ROUND', 'MAX', 'MIN', etc.
- Performing column searches with the 'VLOOKUP' function
- Creating graphs and charts

- [ ] **Communication skills**:

- A good level of communication facilitates the achievement of objectives, problem solving, and increases productivity, because each professional will know exactly what is expected of them, and will transmit their ideas clearly.
