Workshop
Data Science Project

B. Sc. Cristian Lazo Quispe
Computer Vision Scientist at Rimac

clazoq@uni.pe

https://github.com/CristianLazoQuispe

https://www.linkedin.com/in/cristian-lazo-quispe/

aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

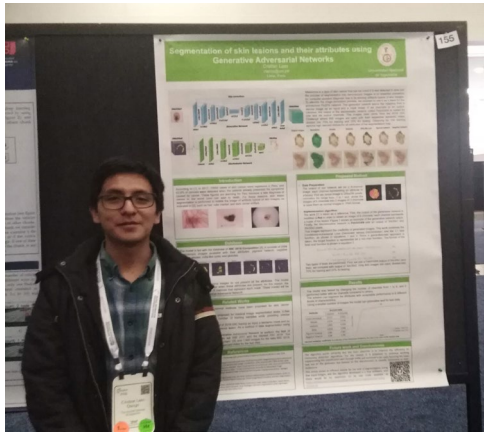
About me

Competitive Programming

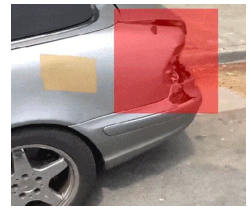
Medical devices

Datathon and Hackathon

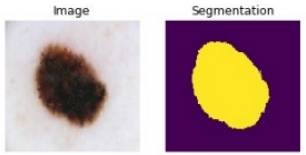
Conferences

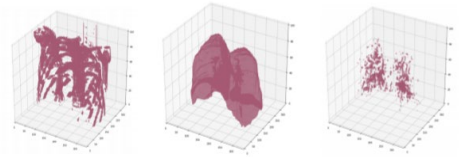


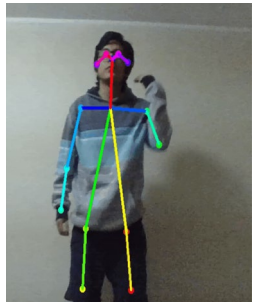
Work experience

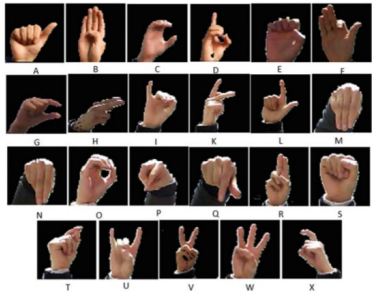


Research experience









https://scholar.google.com.pe/citations?user=t1AI_M4AAAAJ

aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

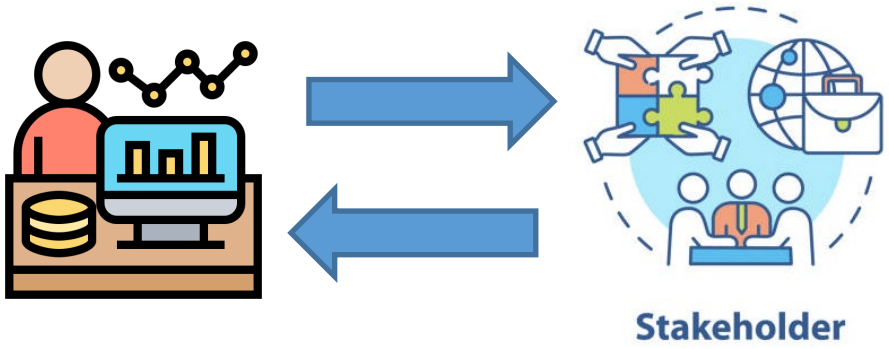
1. Overview

1. Overview
2. Who is a Data Scientist?
3. Data Science LifeCycle
4. Tools
5. Recommendations
6. Practice

aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

2. Who is a Data Scientist?

A data scientist is a professional responsible for collecting, analyzing, and interpreting extremely large amounts of data.



aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

3. Data Science Lifecycle

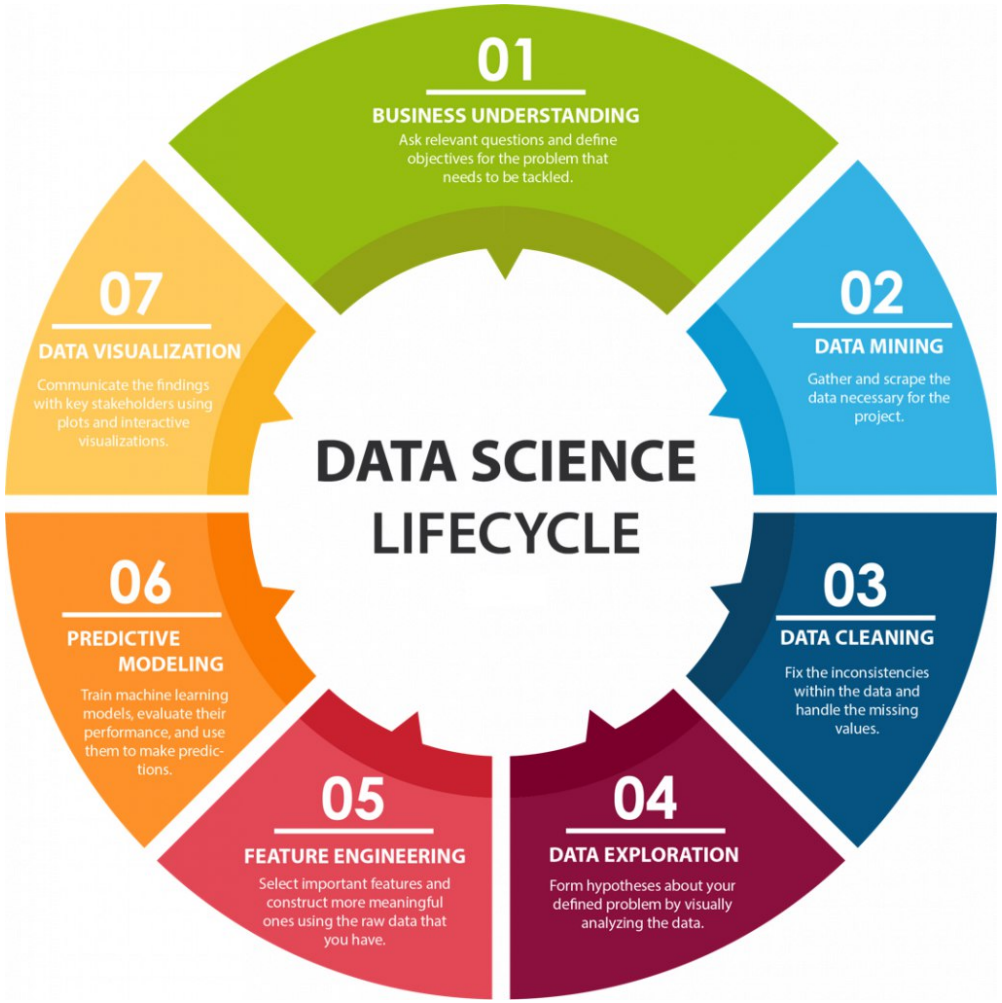
2.1. BUSINESS UNDERSTANDING

Define business problem:

•What is the project Goal?
•What is the problem do you want to solve?
•What data is available that might help solve this problem?
•What is the impact of your solution?
•How our solution affect the Key Performance Indicator (KPI)?

Define model:

•How much or how many? (regression)
•Which category? (classification)
•Which group? (clustering)
•Is this weird? (anomaly detection)
•Which option should be taken? (recommendation)

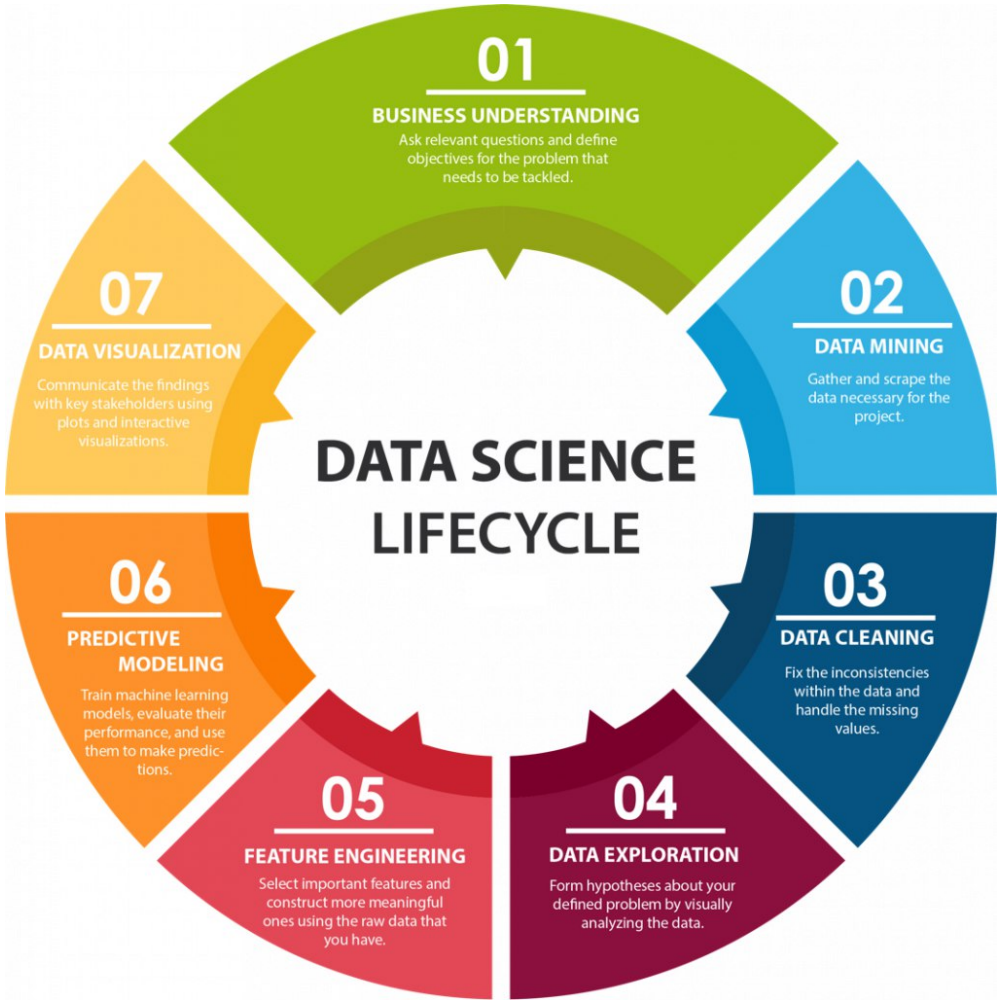


aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

3. Data Science Lifecycle

2.2. DATA MINING

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets.



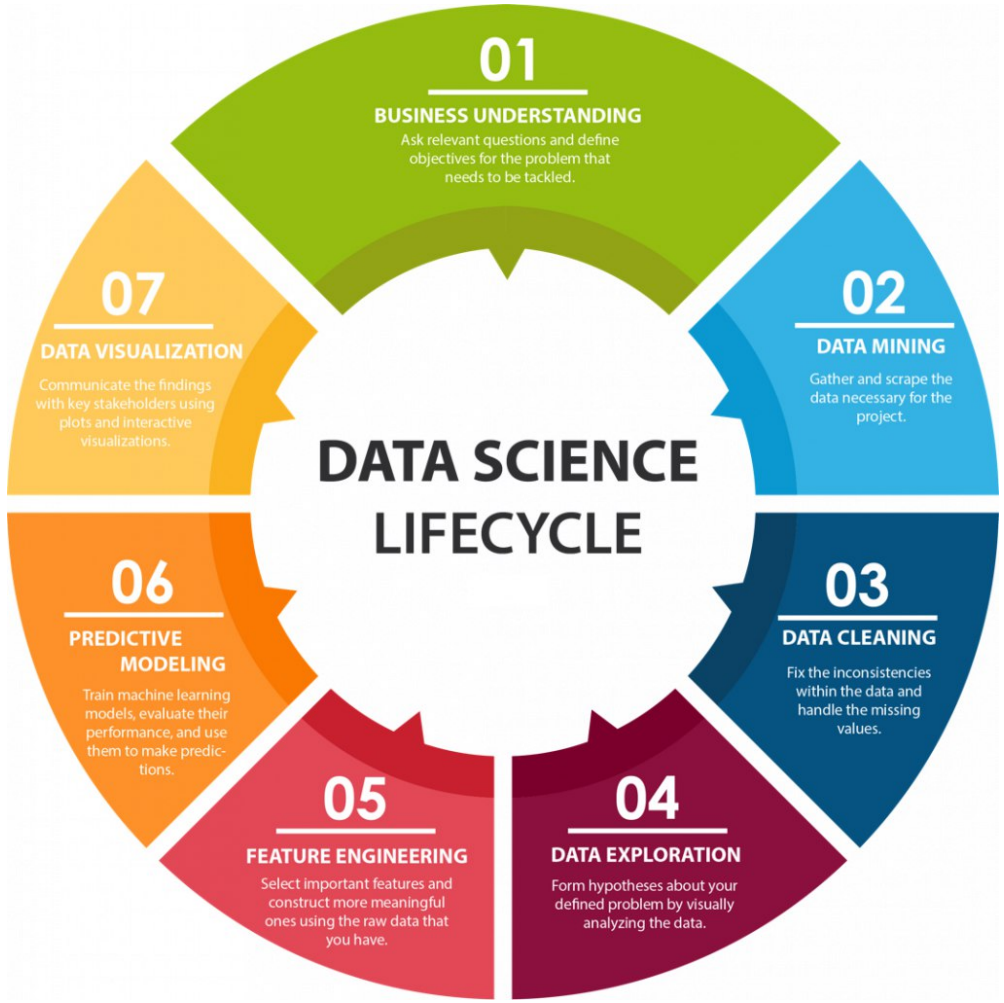
aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

3. Data Science Lifecycle

2.3. DATA CLEANING

Cleaning your data involves taking a closer look at the problems in the data that you've chosen to include for analysis.

Data Problem
Missing data
Data errors
Coding inconsistencies
Missing or bad metadata



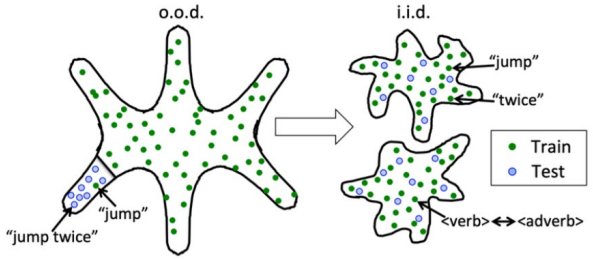
aporta BRECA IMPACTO SOCIAL | BREIT | MIT IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

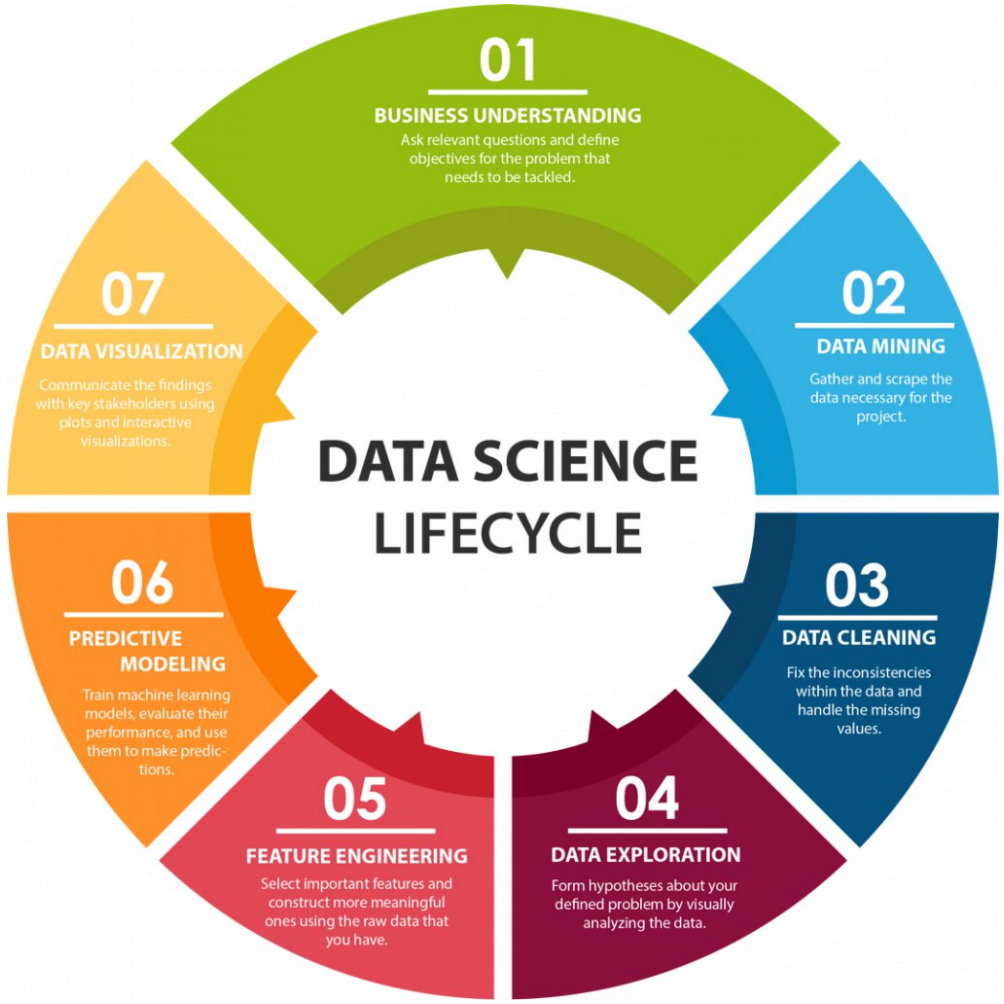
3. Data Science Lifecycle

2.4. DATA EXPLORATION

Create explorations to help you make more effective business decisions by exploring significant company data.

- Variable identification: define each variable and its role in the dataset
- Univariate analysis: for continuous variables, build box plots or histograms for each variable independently; for categorical variables, build bar charts to show the frequencies
- Bi-variable analysis - determine the interaction between variables by building visualization tools
- ~Continuous and Continuous: scatter plots
- ~Categorical and Categorical: stacked column chart
- ~Categorical and Continuous: boxplots combined with swarmplots
- Detect and treat missing values
- Detect and treat outliers





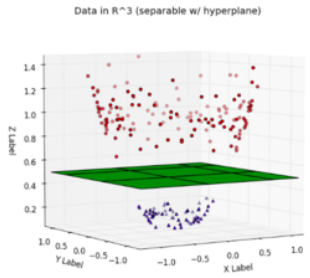
aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

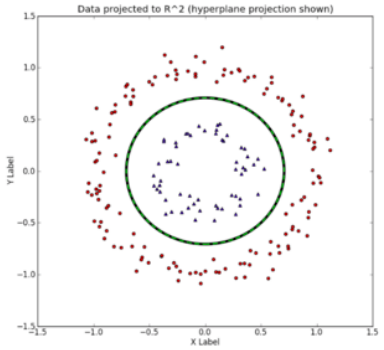
3. Data Science Lifecycle

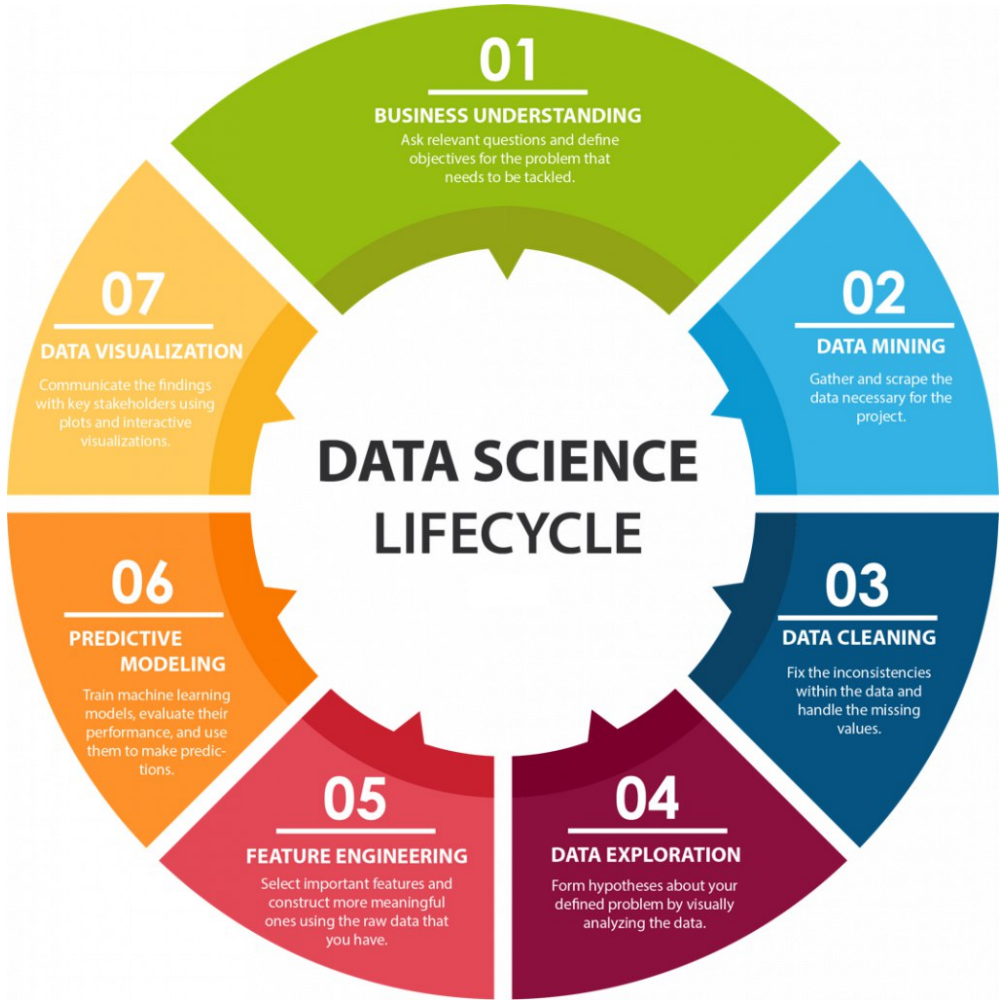
2.5. FEATURE ENGINEERING

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set

Date	Value	$Value_{t-1}$	$Value_{t-2}$
1/1/2017	200	NA	NA
1/2/2017	220	200	NA
1/3/2017	215	220	200
1/4/2017	230	215	220
1/5/2017	235	230	215
1/6/2017	225	235	230
1/7/2017	220	225	235
1/8/2017	225	220	225
1/9/2017	240	225	220
1/10/2017	245	240	225





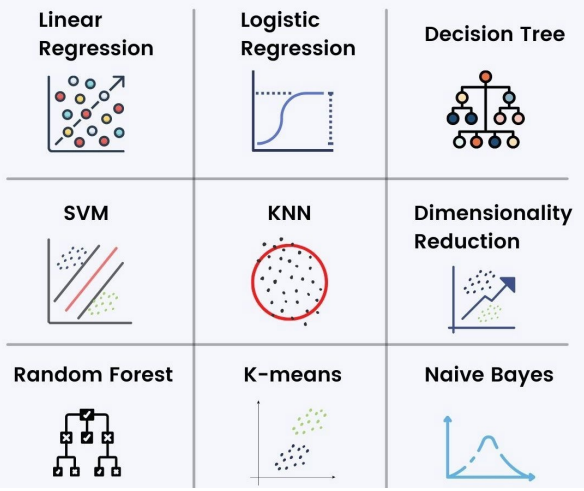


aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

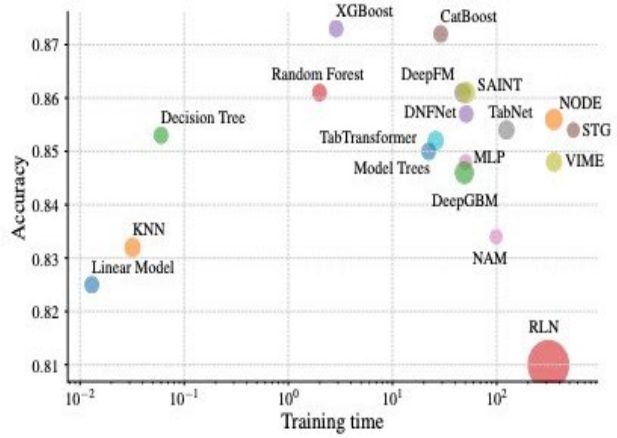
3. Data Science Lifecycle

2.6. PREDICTIVE MODELING

Predictive modeling is the process of taking known results and developing a model that can predict values for new occurrences. It uses historical data to predict future events.

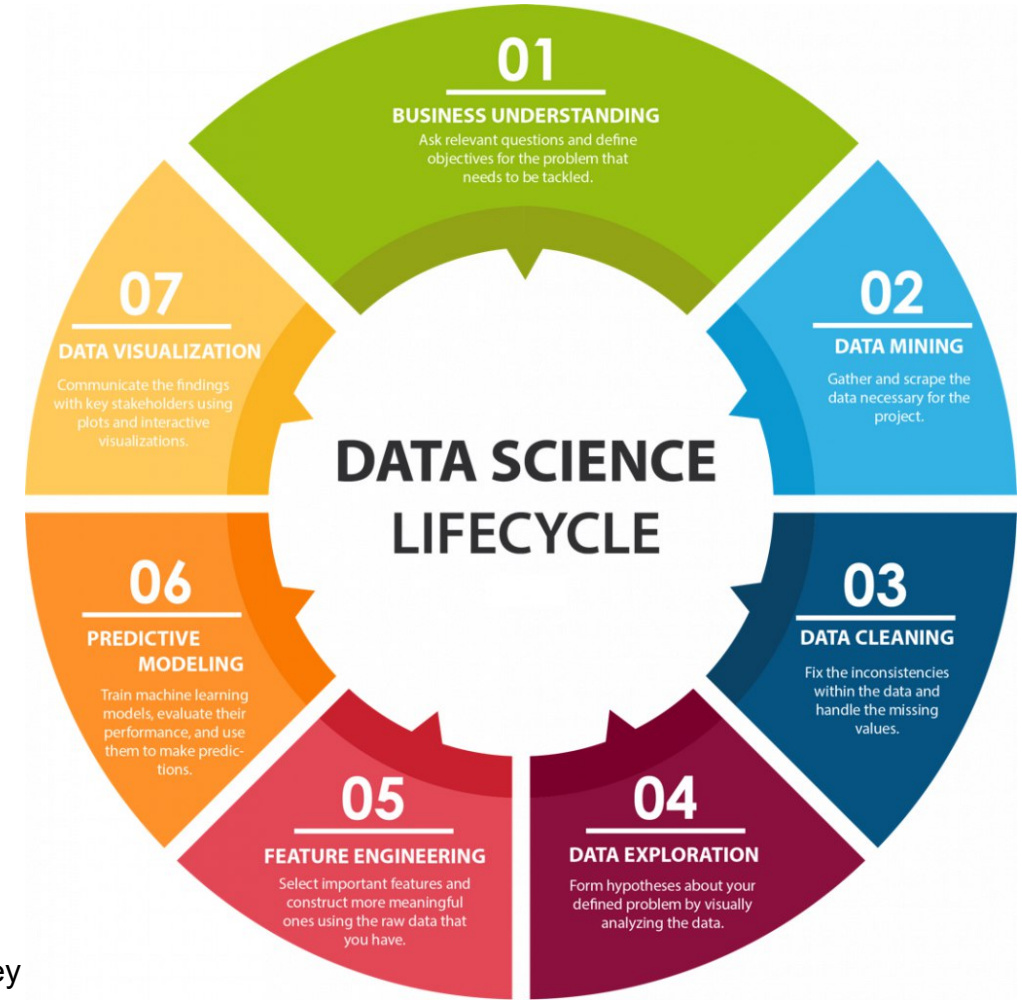


Review the state of the art



Deep Neural Networks and Tabular Data: A Survey

https://arxiv.org/pdf/2110.01889.pdf



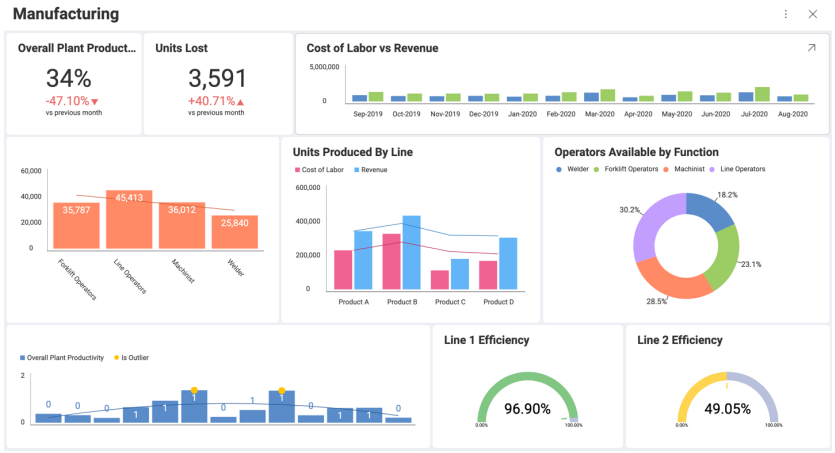
aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

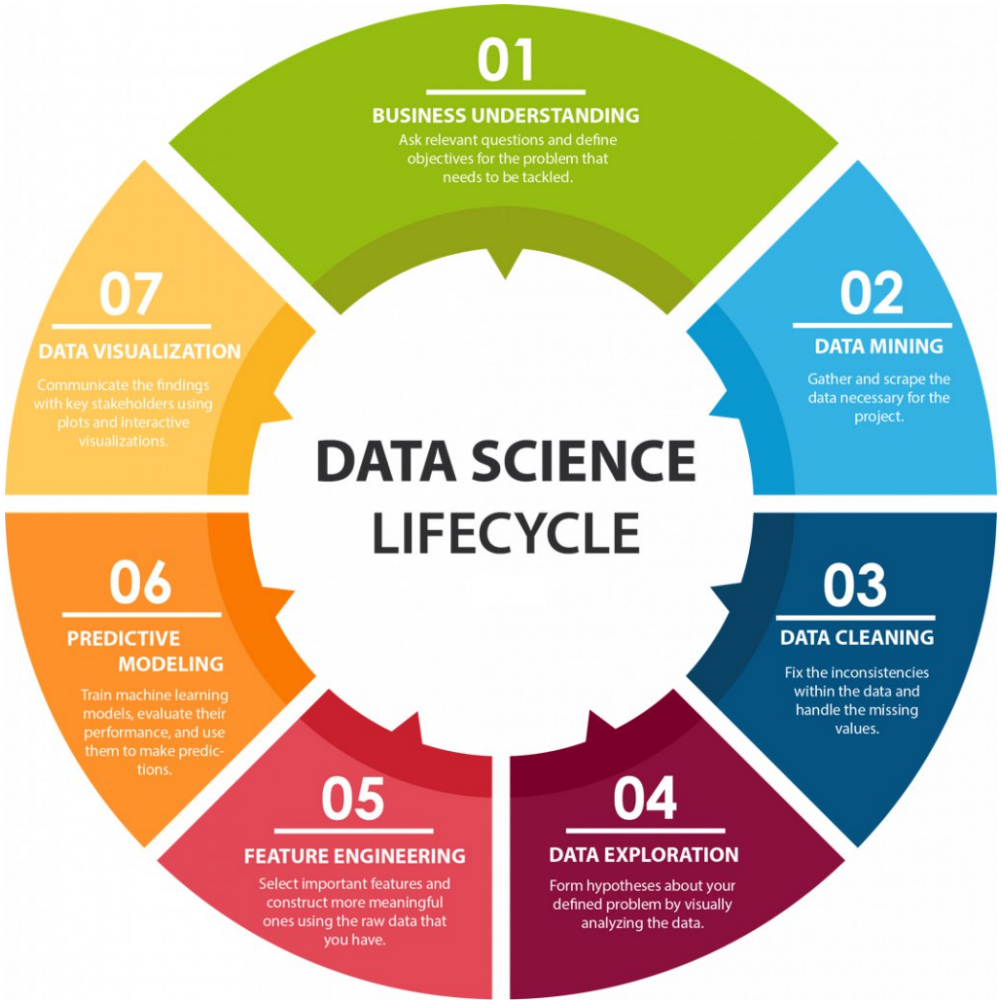
3. Data Science Lifecycle

2.7. DATA VISUALIZATION

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations.

These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.





aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

4.Tools

Version control tools

GitHub
Bitbucket
GitLab

Data Visualization tools


Power BI


Tableau


Google Data Studio

Machine Learning tools

aporta
BRECA IMPACTO SOCIAL

BREIT

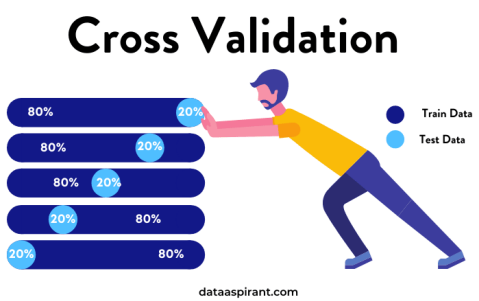
MIT

IDSS
MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY

edX MicroMasters®

5. Recomendations

Robust Validation



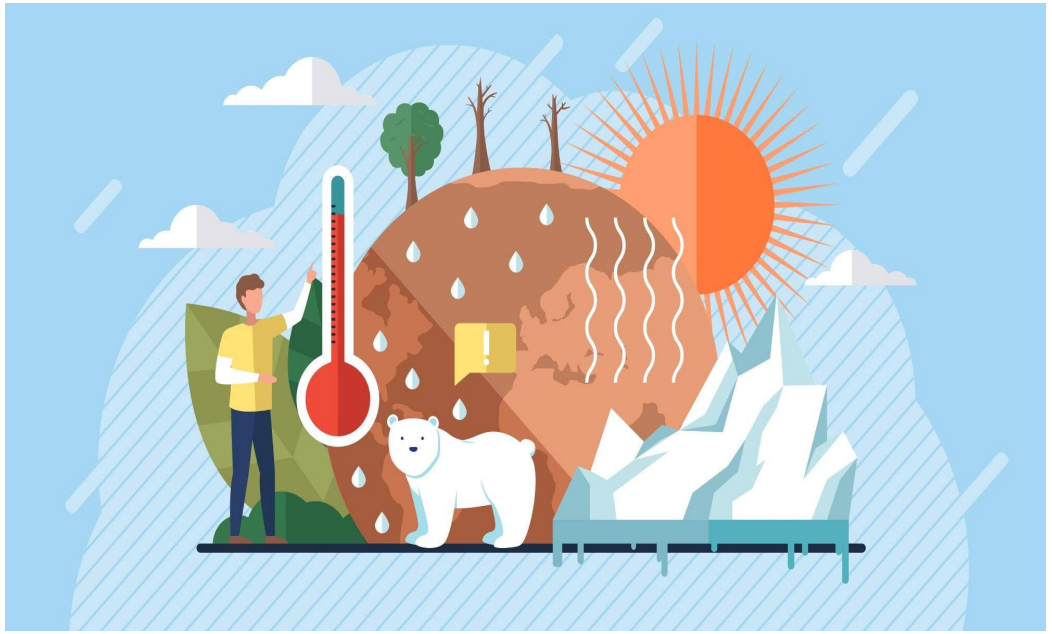
Agile Methodology

Effective Communication

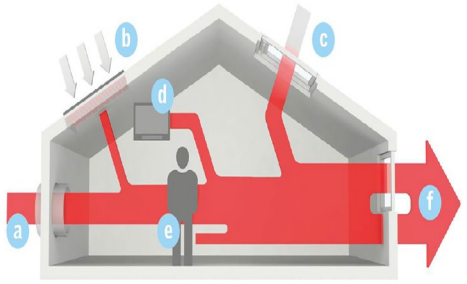
aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

6. Practice

https://www.kaggle.com/c/widsdatathon2022



Analyzing differences in building energy efficiency



https://github.com/CristianLazoQuispe/WorkshopDataScienceProject.git

aporta BRECA IMPACTO SOCIAL | BREIT | MIT | IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY | edX MicroMasters®

aporta
BRECA IMPACTO SOCIAL

BREIT

IDSS
MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY

edX MicroMasters®