AFRICA DATA SCHOOL

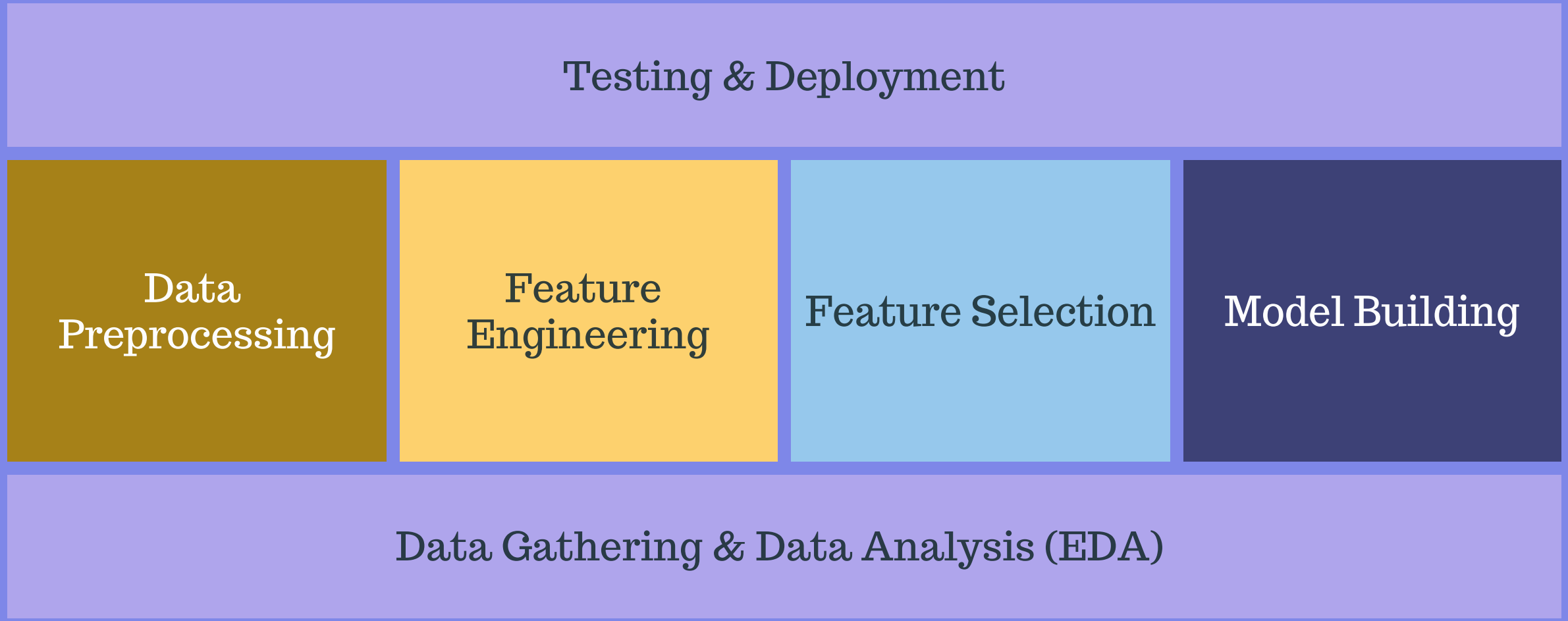
Machine Learning Pipeline

ANALYTICS INSIGHT

Intersection of data science and ML engineering, where the process of **training** and **testing** a model has been codified. The pipeline can then be run as **frequently** as needed.

AFRICA DATA SCHOOL

Top-Down ML Pipeline



AFRICA DATA SCHOOL

Data Gathering

Process of **gathering** and **measuring** information,

on **variables** of interest.

Collecting data for ML If you don't have any

Open Source Datasets

- Companies like Google are ready to give away data for Machine Learning.
- Real value is in internally collected data.

Collect data the right way

- Data collection with paper ledgers, (.xlsx, .csv files)
- Harder time with data preparation, but ML-friendly dataset.

Big Data

- Not about petabytes but the ability to process them right away.
- Larger dataset are harder to yield insights. Start small and reduce complexity of the data.

AFRICA DATA SCHOOL

Data Analysis (EDA)

An approach to analyzing data sets to **summarize** their main characteristics, often with **visual** methods.

AFRICA DATA SCHOOL

1 Gain intuition about the data.

2 Conduct sanity checks to ensure you're applying the right dataset.

3 Find out missing data and check any outliers.

4 Summarize the data.

Steps in EDA

Variable Identification	Univariate Analysis	Bivariate Analysis	Missing Value Treatment	Detecting Outliers
Discovering variable types	Individual characteristics of each feature.	Relationship between any two variables.	Find out if there is any specific reason why these values are missing and how we treat them	Understand presence of outliers.

AFRICA DATA SCHOOL

Types of Variables (Univariate Analysis)

Continous Variable

- Numeric variables thats have infinite number of values between any two values.

Graphical Technique

- Histogram
- KDE
- Box Plots, Q-Q Plot (outliers)

Categorical Variable

- Nominal variable; one that has two or more categories.

Graphical Technique

- Bar Plot
- Pie Chart
- Frequency Table

AFRICA DATA SCHOOL

Relationships (Bivariate Analysis)

	Graph Techniques
Continuous-Continuous	Scatterplot, Heatmap, Jointplot, Pairplot
Categorical-Continuous	Factorplot, SwarmMap, ViolinPlot, StripPlot
Categorical-Categorical	Crosstab, Stacked Bar, Bar Chart

AFRICA DATA SCHOOL

Data Preprocessing

The technique of preparing (**cleaning** and **organizing**) the raw data to make it suitable for a **building** and **training** Machine Learning models.

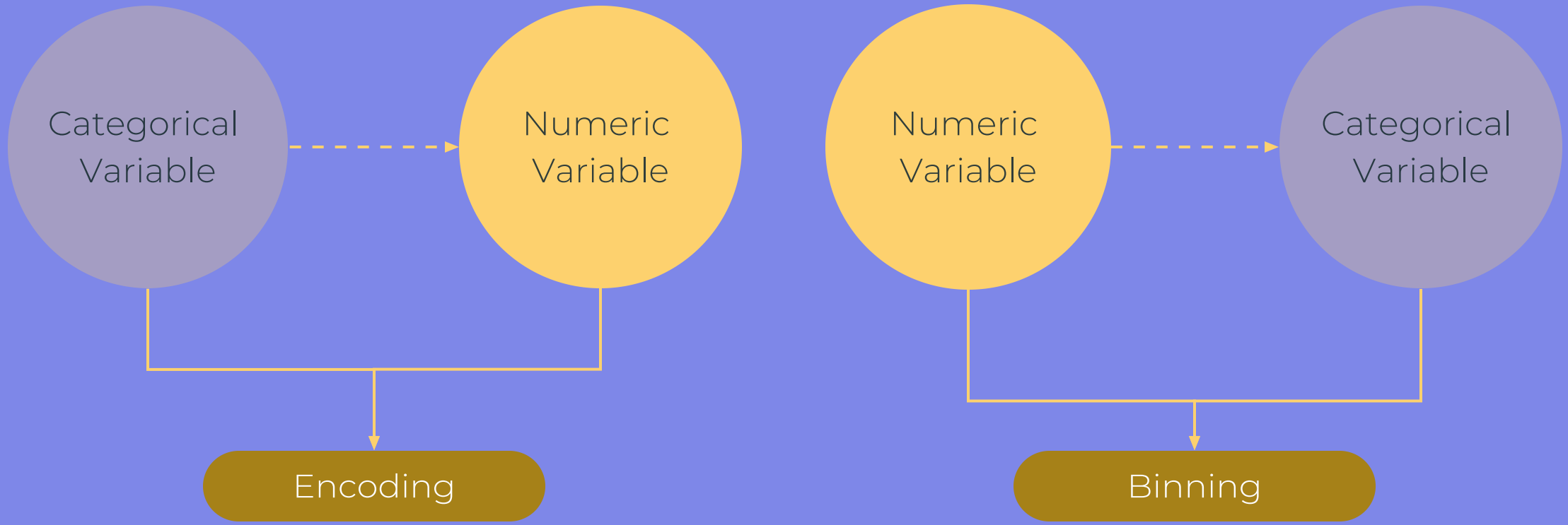
AFRICA DATA SCHOOL

Data preprocessing

Missing Value Imputation	Replacing missing data with subtituted values
One Hot Encoding	Representation of categorical variables as numbers
Ordinal encoding	Mapping each unique label to an integer value.
Cardinal Encoding	Encodes variables with many levels (high cardinality features.
Normalization	Adjust values to a common scale.

Transformation

AFRICA DATA SCHOOL



Feature Engineering

The process of using **domain knowledge** to extract features from raw data via data mining techniques.

Feature Engineering Techniques

Feature Interaction	Prediction based on two features
Polynomial Features	Features created by raising existing features to an exponent.
Trigonometry Features	Using trigonometric functions to extract features.
Group Features	Decidde aggregation functions of the features.
Bin Numeric Features	Turning continuous variables into categorical variable
Combine Rare Levels	Takes categorical variable and combines all levels with frequencies < threshold

AFRICA DATA SCHOOL

Feature Selection

Process of **reducing** the number of input variables when developing a predictive model.

To reduce the **computational cost** of modeling and, in some cases, to **improve the performance** of the model

AFRICA DATA SCHOOL

Feature Selection Techniques

Filter Method

Measure the relevance of features by their correlation with dependent variable.

Wrapper Method

Measure the usefulness of a subset of feature by actually training a model on it.

Embedded Method

Combine the qualities' of filter and wrapper methods.

AFRICA DATA SCHOOL

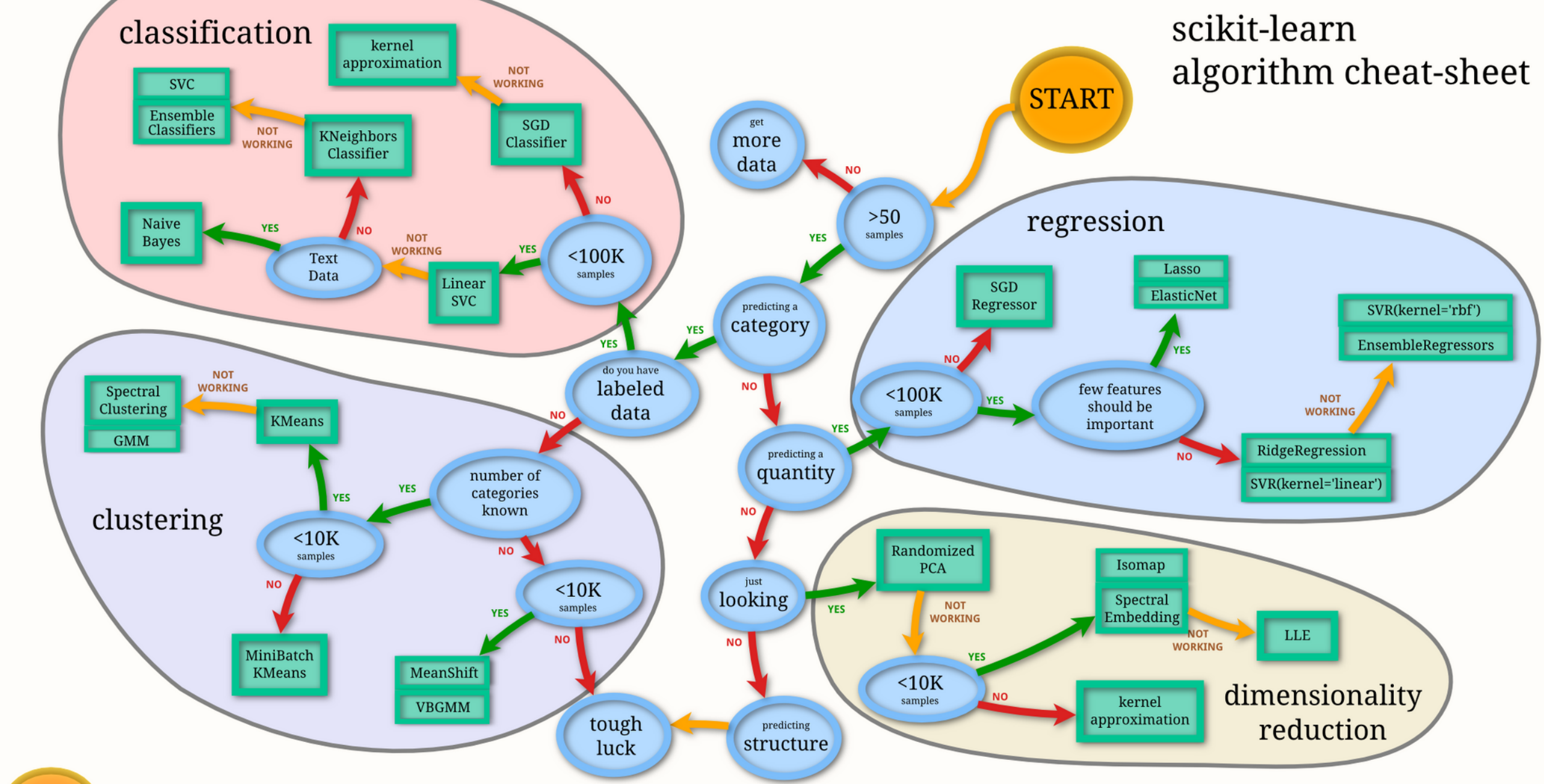
Modeling

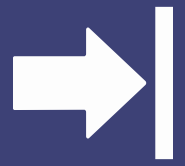
How to **train**, **fine-tune**, and **validate** a machine learning model

AFRICA DATA SCHOOL

Choosing the right estimator/ model

scikit-learn algorithm cheat-sheet





CONCLUSION

I have briefly discussed the importance of EDA in the Data Science pipeline and steps that are involved in Machine Learning Pipeline.

Additional Resources

Class Notebooks

- https://github.com/Africa-Data-School/ADS_Course_Material

Analytics Vidhya

- https://www.analyticsvidhya.com/blog/

Sklearn

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html