AN INITIATIVE OF:

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR 80th anniversary
Touching lives through innovation

nicis
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
DIRISA

Supervised Learning: Theory

DIRISA Datathon | Instructor: Kgaugelo Mmakola

Why Supervised Learning in a Datathon?

Many real-world problems involve predicting an outcome (price, label, class)

Supervised learning is the right tool when you have labeled data

Most datathon problems involve either:

Regression: Predicting continuous values (e.g., housing prices)

Classification: Predicting categories or classes (e.g., spam vs not spam)

What is Supervised Learning?

Definition: Supervised Learning is a type of machine learning where an algorithm learns from labeled data.

Each training example includes an input (features) and the correct output (label).

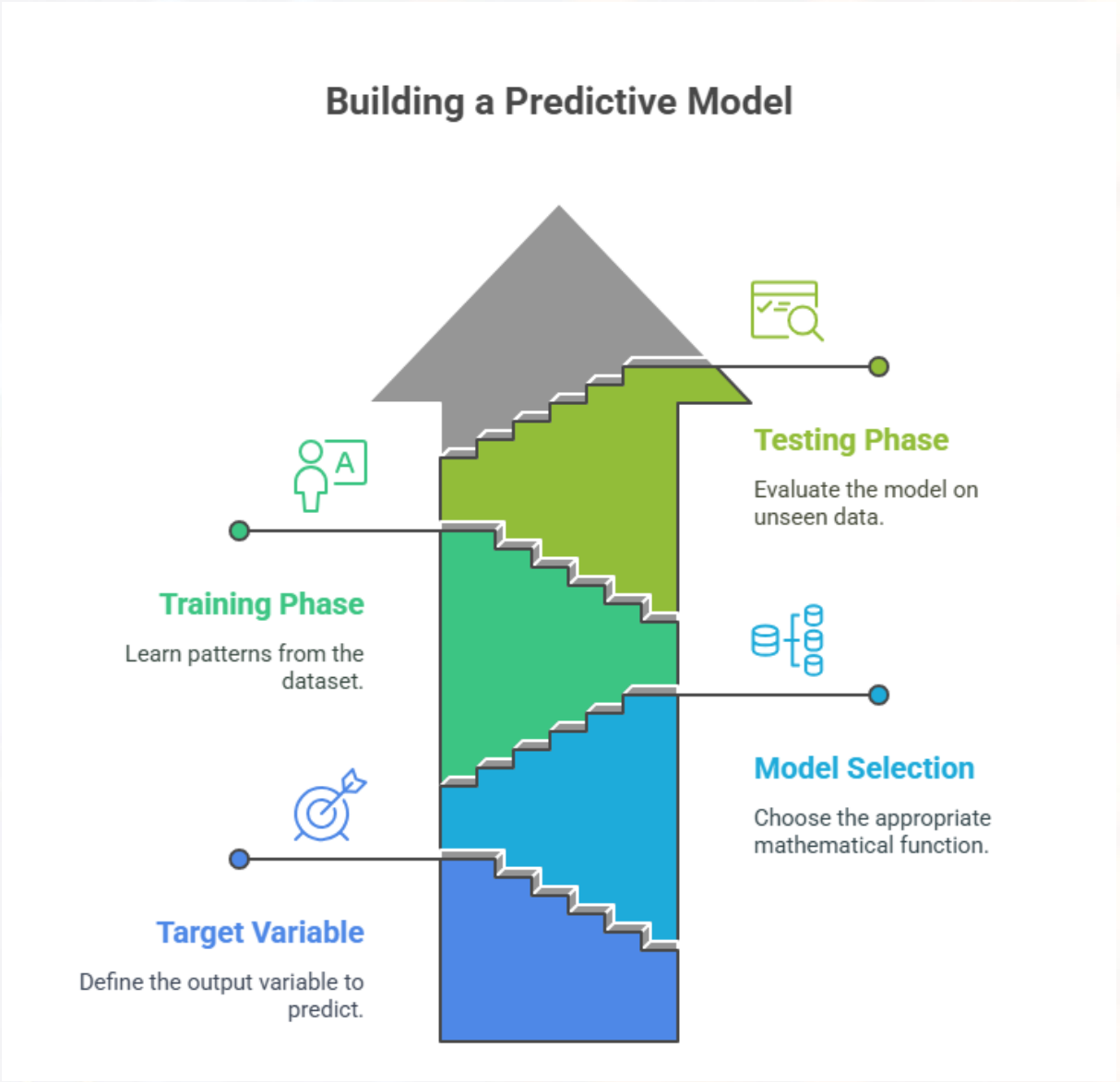
Key Ideas:

- The algorithm tries to learn a function that maps inputs (X) to outputs (y).
- Once trained, the model can predict outputs for new, unseen inputs.

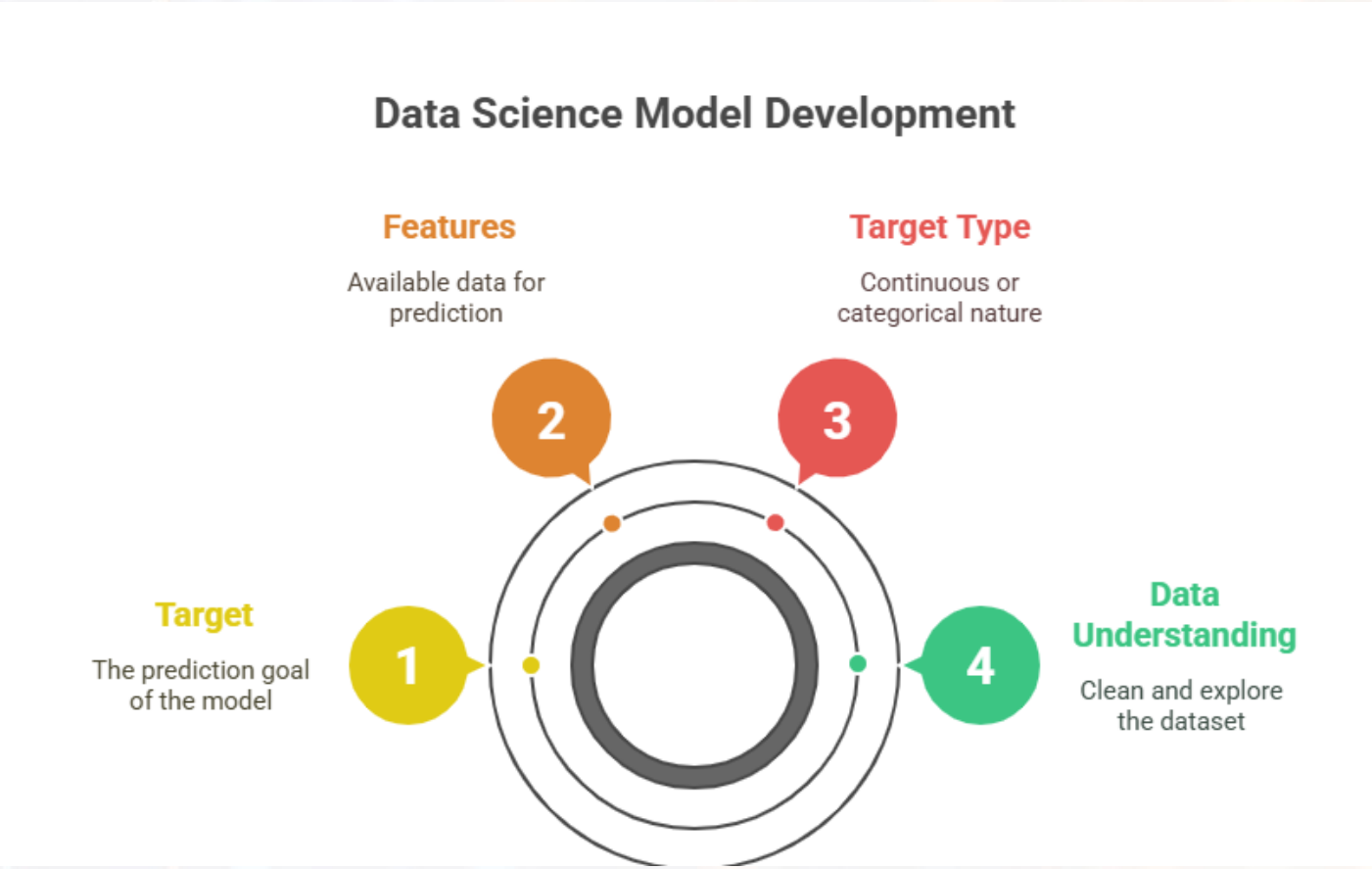
Real-world Examples:

- Predicting house prices (Regression)
- Identifying spam emails (Classification)

Key Concepts in Supervised Learning



How to Frame a Supervised Problem



Regression vs Classification

Choose the appropriate machine learning task for your prediction needs.

VS

Regression

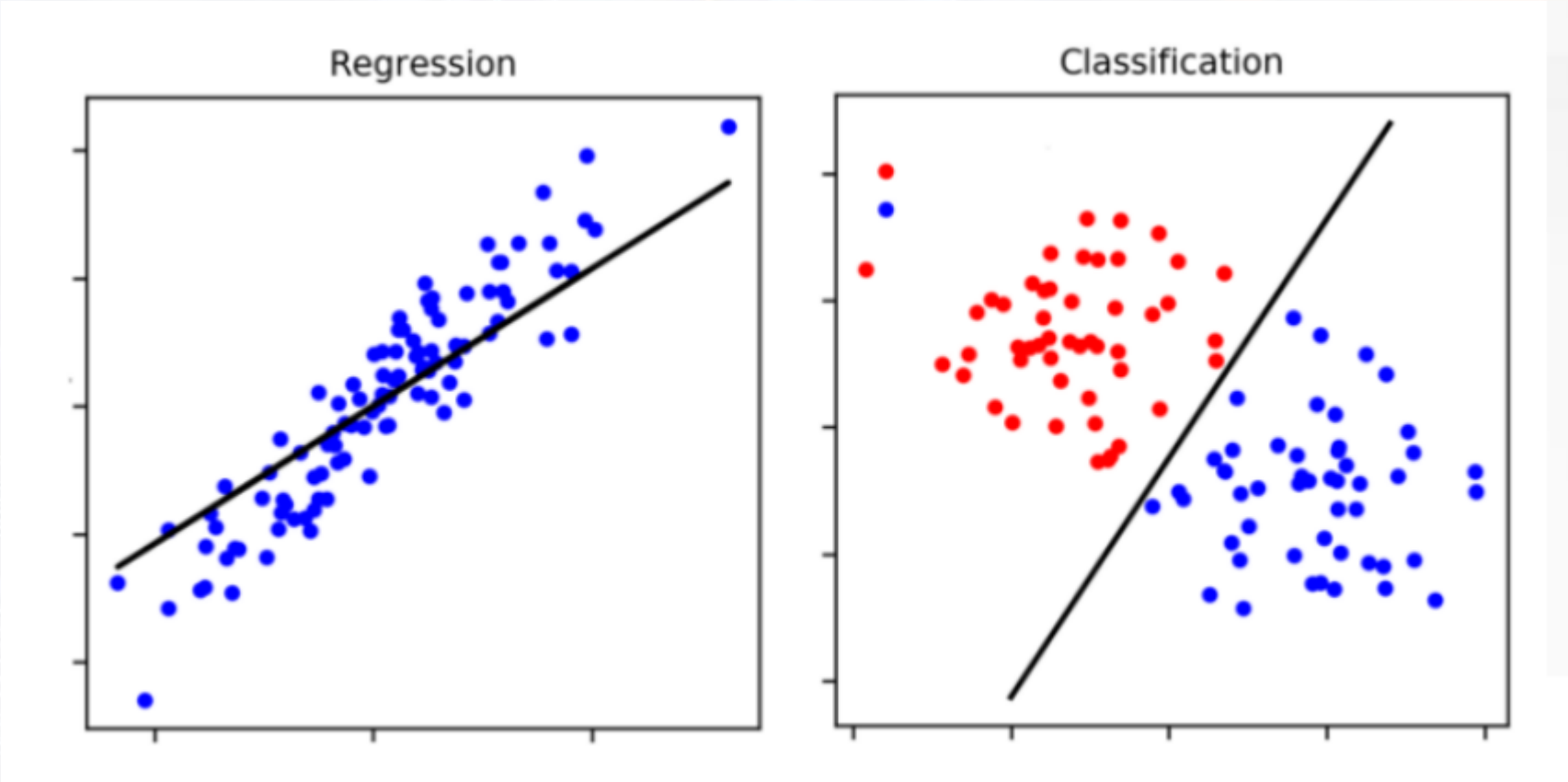
Predict numeric values

Classification

Predict categories

Regression vs Classification

category/class label vs continuous number



Regression


nicis
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM
DIRISA

Regression

Regression is a type of supervised learning used to predict continuous numerical values.

Regression Examples:

- Predicting number of people who will click a Google ad based on the ad content and data about the user's prior online behavior,
- Predicting the number of traffic accidents based on road conditions and speed limit,
- Predicting weather parameters (such as wind speed) based on historical weather behaviour.

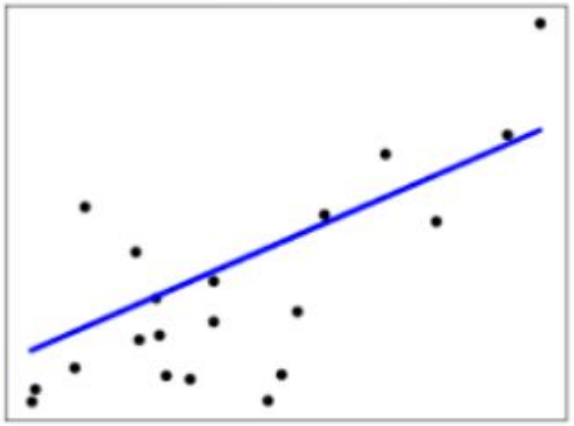
Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y)

y = f(X), X = input features, y = target variable

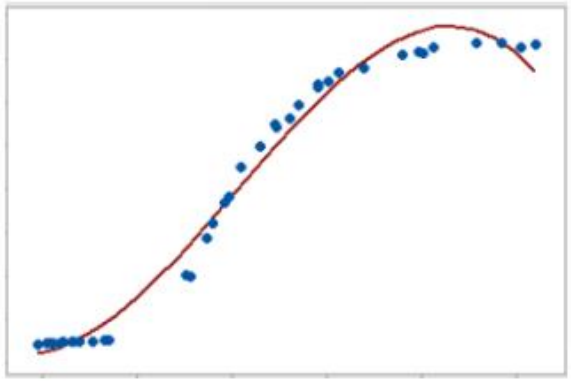
Regression Algorithms

Optimal regression model depend on dataset

- Linear



- Non-Linear



Regression Algorithms - Linear

- Simple linear regression
- Ordinary Least Squares
- Stochastic Gradient Descent etc.

Regression Algorithms – Non-Linear

- Decision Trees
- Random Forest Regression (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)
- Support vector regression (kernel = linear, polynomial, rbf) (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html)
- Kernel ridge regression
- Multi layer perceptron (Deep learning)
- Artificial Neural network (LSTM/ Recurrent Neural Network) (Deep learning)

Regression Models Explained

Linear Regression

Assumes a linear relationship and fits a straight line.

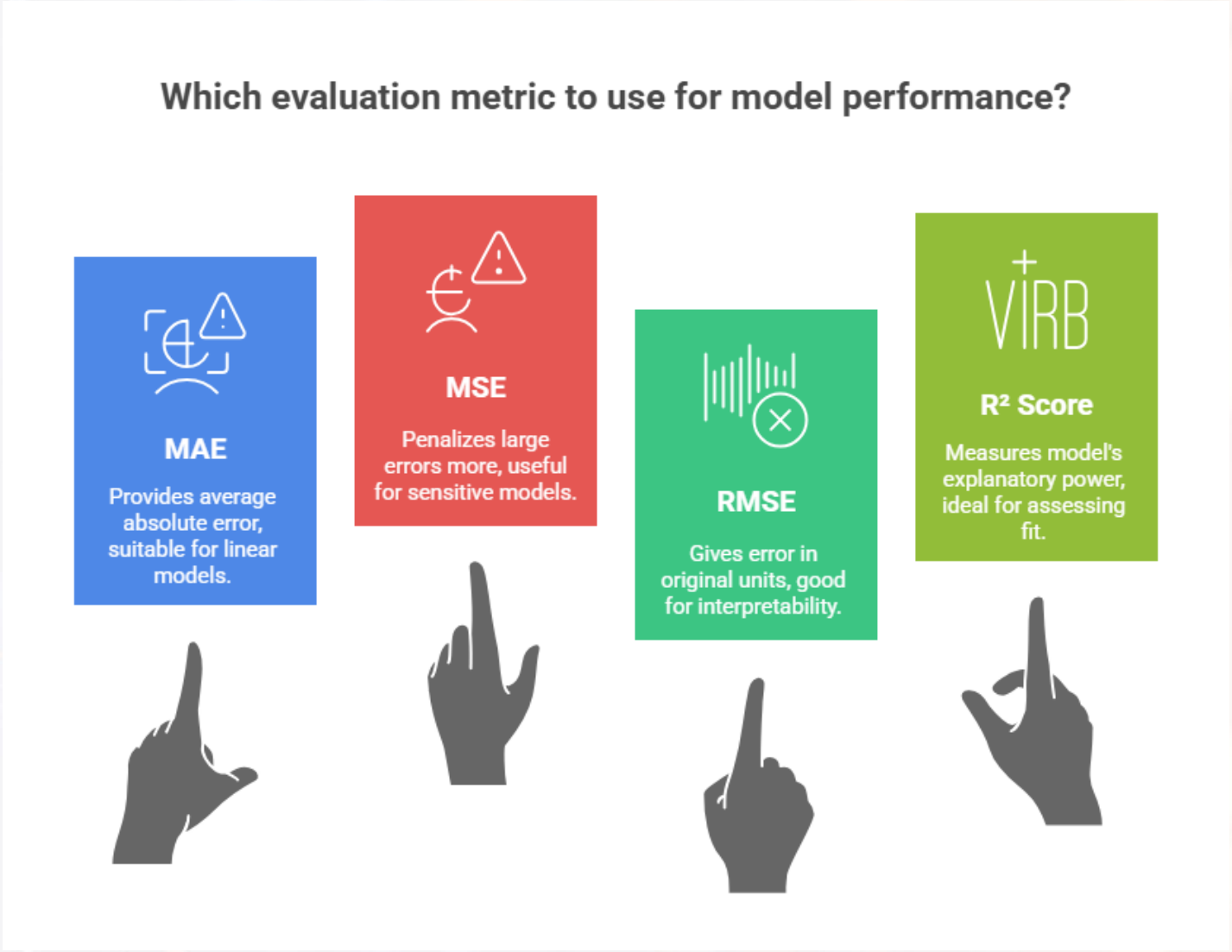
Decision Tree Regressor

Splits data into regions and averages outputs.

Random Forest Regressor

Builds multiple decision trees and averages their outputs for accuracy.

Regression Metrics



Classification

Classification

Classification is a type of supervised learning used to predict categories or labels.

Classification is used when the target, or value to predict, is a discrete class label.

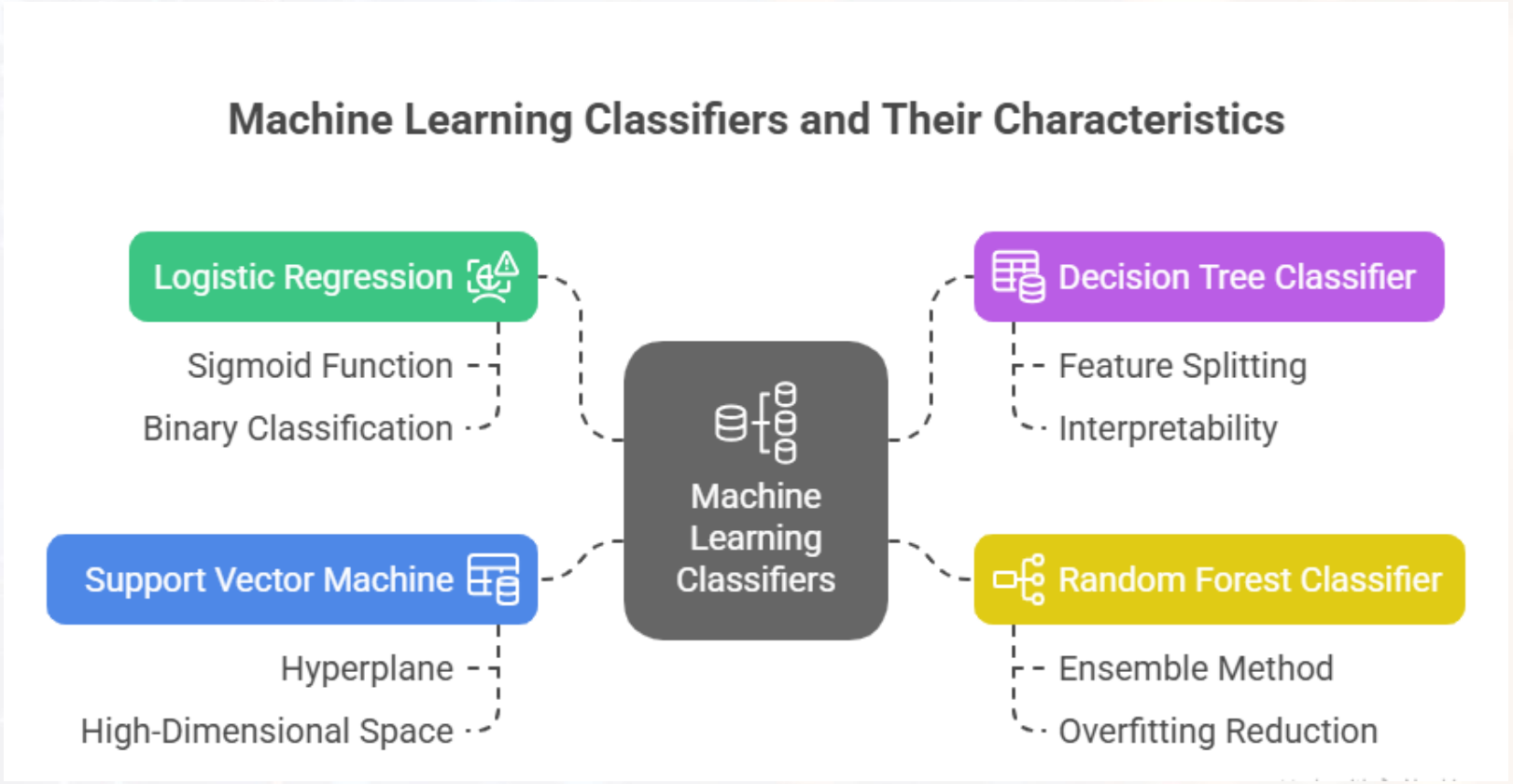
Classification examples:

- Spam filtering
- Identifying an object in an image
- Customer behaviour prediction

Classification Algorithms

- Some examples of algorithms are:
- Logistic Regression (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
- Decision trees (https://scikit-learn.org/stable/modules/tree.html)
- Random Forest (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)
- XGBoost (https://pypi.org/project/xgboost/)
- Support Vector Machines (https://scikit-learn.org/stable/modules/svm.html)
- Multi layer perceptron
- Artificial neural networks (e.g., Convolutional neural networks)

Classification Models Explained



Classification Metrics

Model Evaluation Metrics

Metric	Description	Formula
Accuracy	Correct prediction rate	(TP+TN)/(TP+TN+FP+FN)
Confusion Matrix	Performance visualization	TP, FP, TN, FN
Precision	Correct positive predictions	TP / (TP + FP)
Recall (Sensitivity)	Detected actual positives	TP / (TP + FN)
F1 Score	Balances precision and recall	2 * (Precision * Recall) / (Precision + Recall)

Before you start

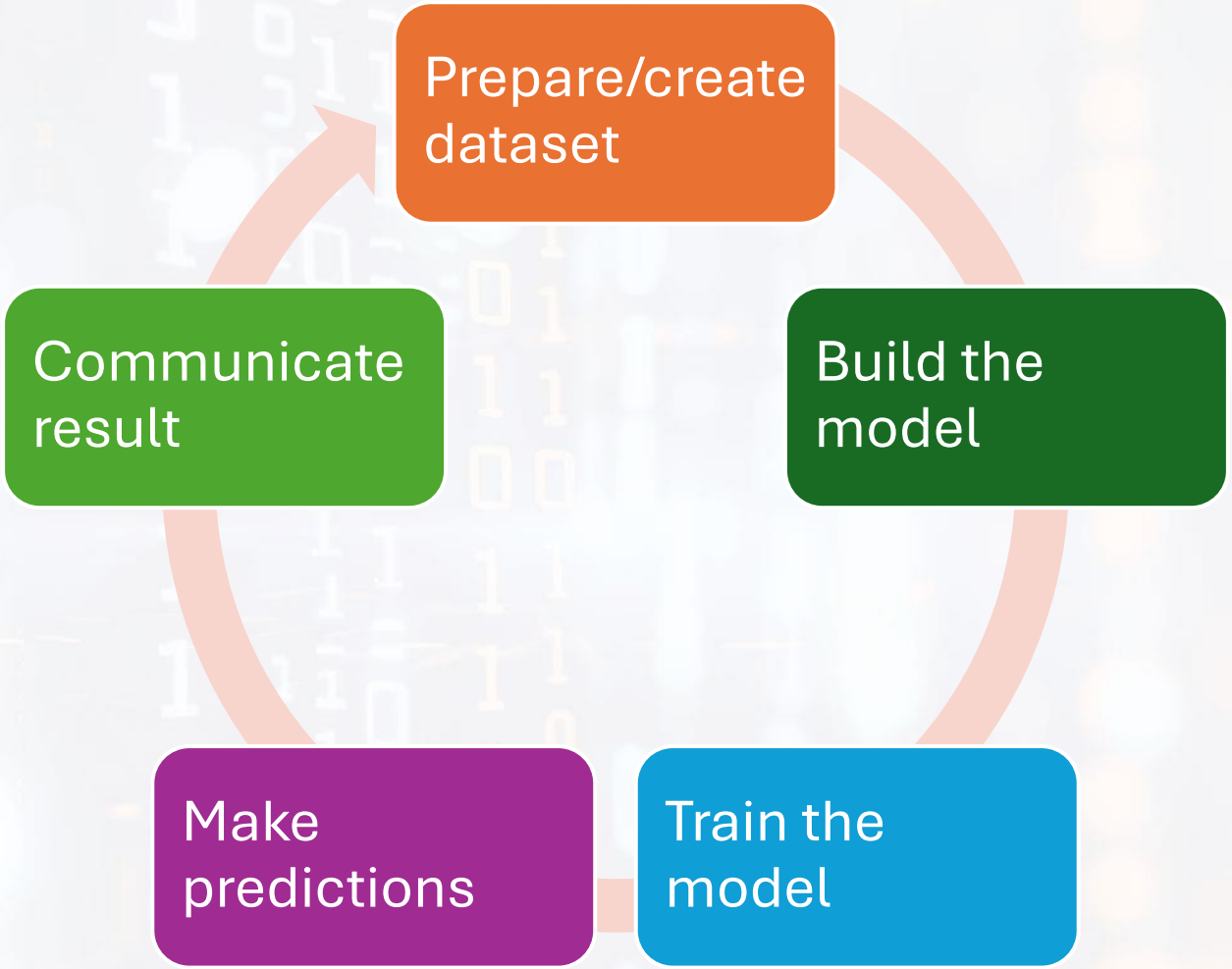
Preparing Data For Regression or Classification

- Rescale Inputs (normalization/standardization)
- Randomisation
- Remove Collinearity (if exist)
- Test/Train split

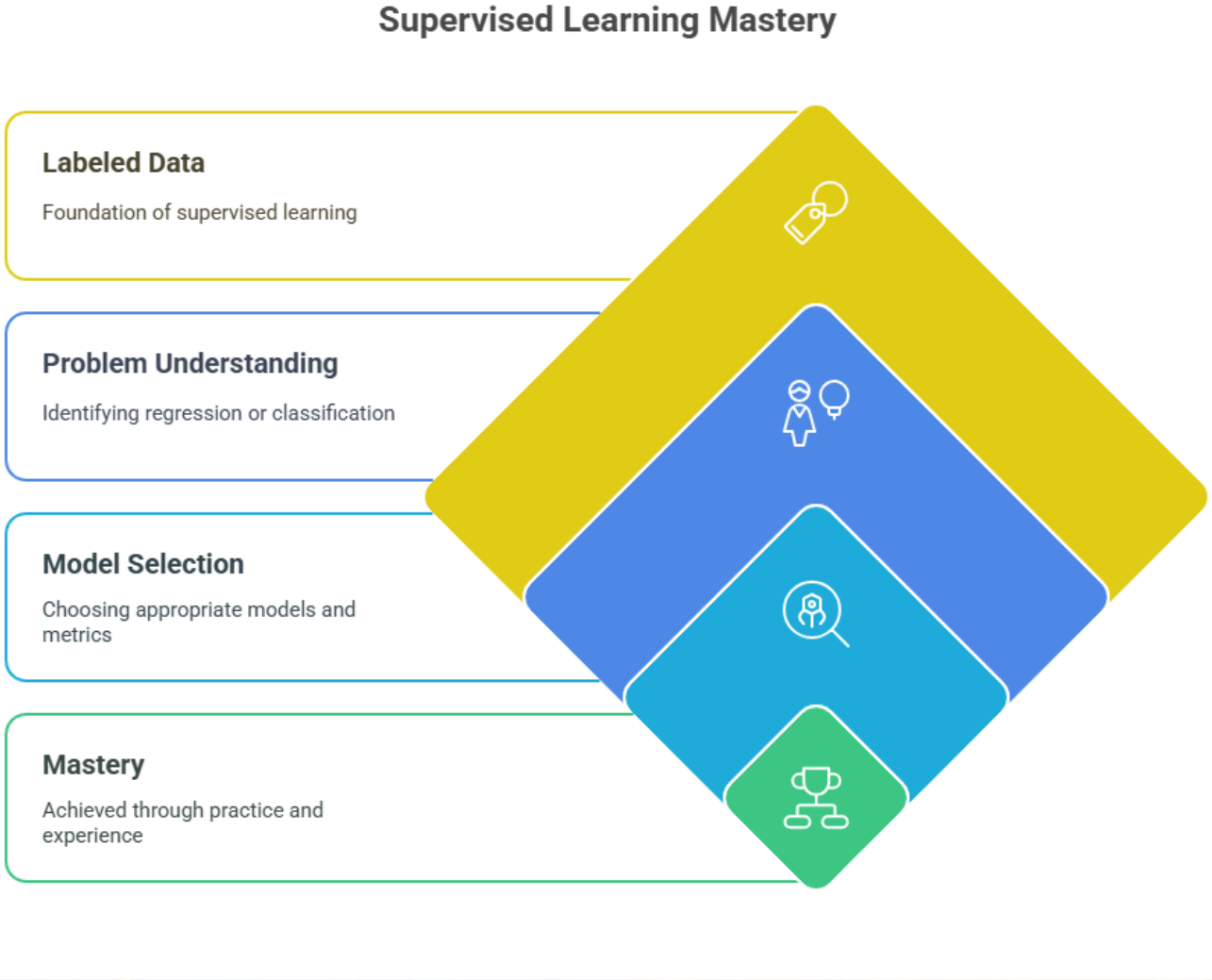
Additional operations if needed (applicable to Linear regression)

- Linear Assumption
- Remove Noise

Workflow


Prepare/create dataset
Build the model
Train the model
Make predictions
Communicate result

Summary & Takeaways




Thank you!!
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM
DIRISA