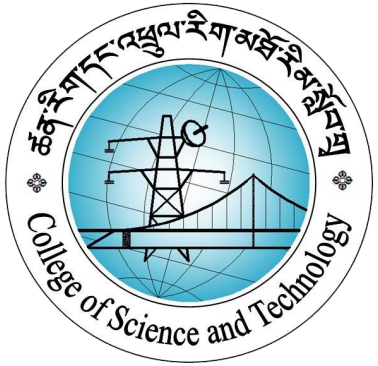


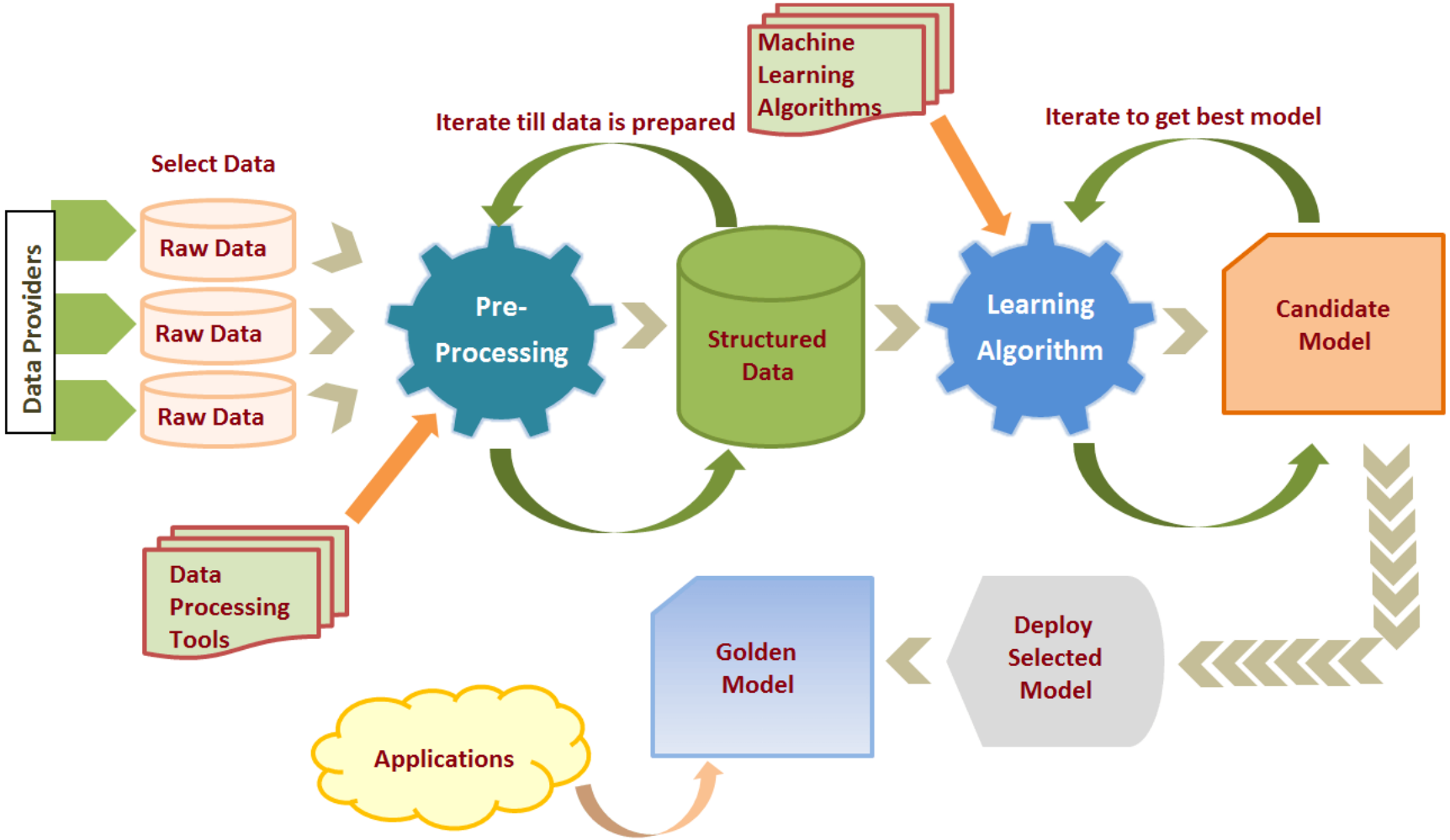
Royal University of Bhutan

Unit III: Regression

CTE309 Machine Learning

AS2024: BE Information Technology

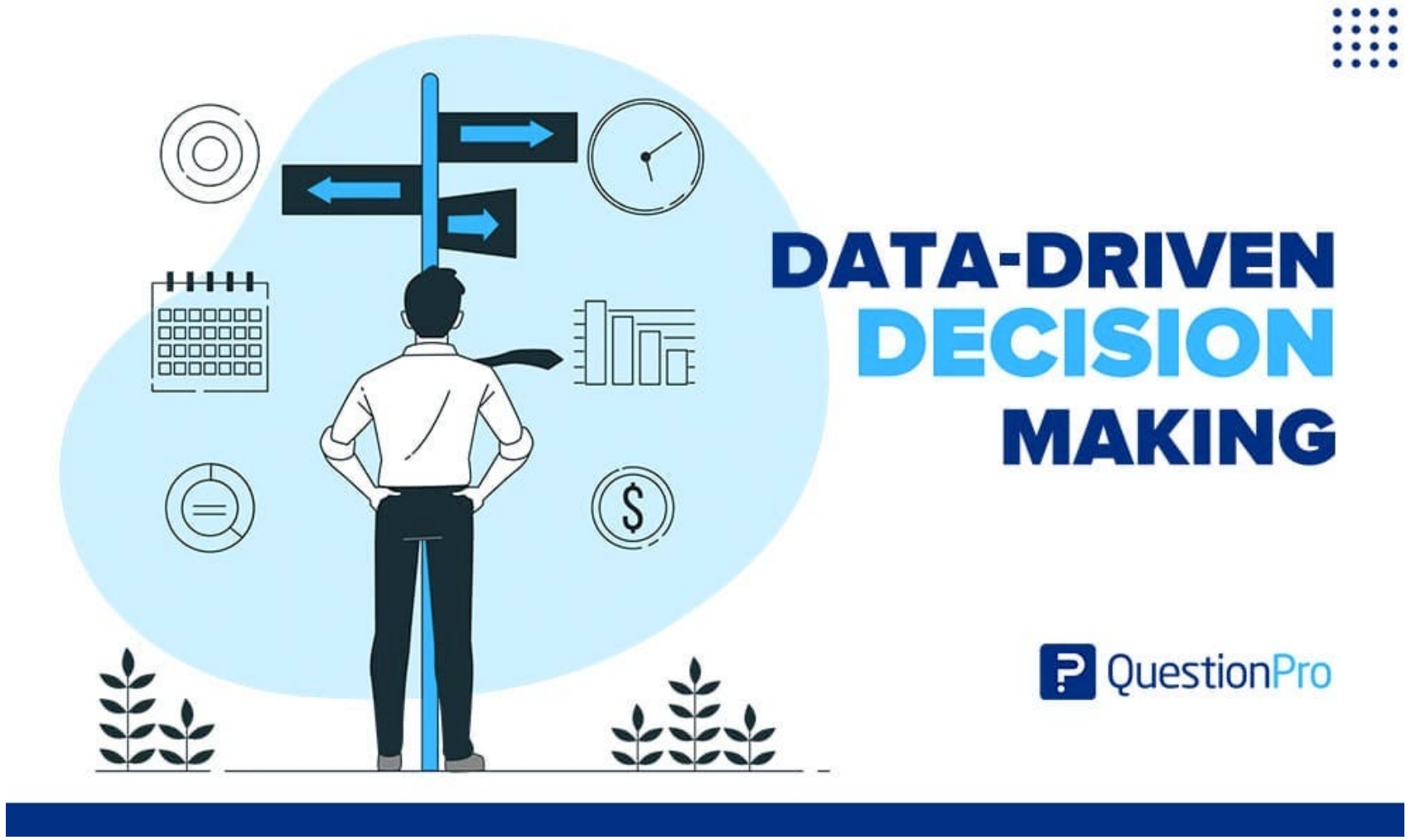
Machine Learning Pipeline



Overview

- Introduction
- Regression in ML
- Terminologies
- Regression types
- Simple Linear Regression
- Demo

Introduction



Introduction

- a statistical approach used to analyze the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables).
- The objective is to determine the most suitable function that characterizes the connection between these variables.
- It seeks to find the best-fitting model, which can be utilized to make predictions or draw conclusions.
- Correlation vs regression

Regression in Machine Learning

- It is a **supervised** machine learning technique, used to **predict** the value of the dependent variable for new, unseen data.
- It models the relationship between the input features and the target variable, allowing for the estimation or prediction of numerical values.
- Regression analysis problem works with if output variable is a real or continuous value, such as “salary” or “weight”.
- Many different models can be used, the simplest is the linear regression.
- It tries to fit data with the best hyper-plane which goes through the points.

Terminologies

- **Response Variable:** The primary factor to predict or understand in regression, also known as the dependent variable or target variable.
- **Predictor Variable:** Factors influencing the response variable, used to predict its values; also called independent variables.
- **Outliers:** Observations with significantly low or high values compared to others, potentially impacting results and best avoided.
- **Multicollinearity:** High correlation among independent variables, which can complicate the ranking of influential variables.
- **Underfitting and Overfitting:** Overfitting occurs when an algorithm performs well on training but poorly on testing, while underfitting indicates poor performance on both datasets.

Regression Types

- **Simple Regression**
  - Used to predict a continuous dependent variable based on a single independent variable.
  - Simple linear regression should be used when there is only a single independent variable.
- **Multiple Regression**
  - Used to predict a continuous dependent variable based on multiple independent variables.
  - Multiple linear regression should be used when there are multiple independent variables.
- **Non-Linear Regression**
  - Relationship between the dependent variable and independent variable(s) follows a nonlinear pattern.
  - Provides flexibility in modeling a wide range of functional forms.

Simple Linear Regression

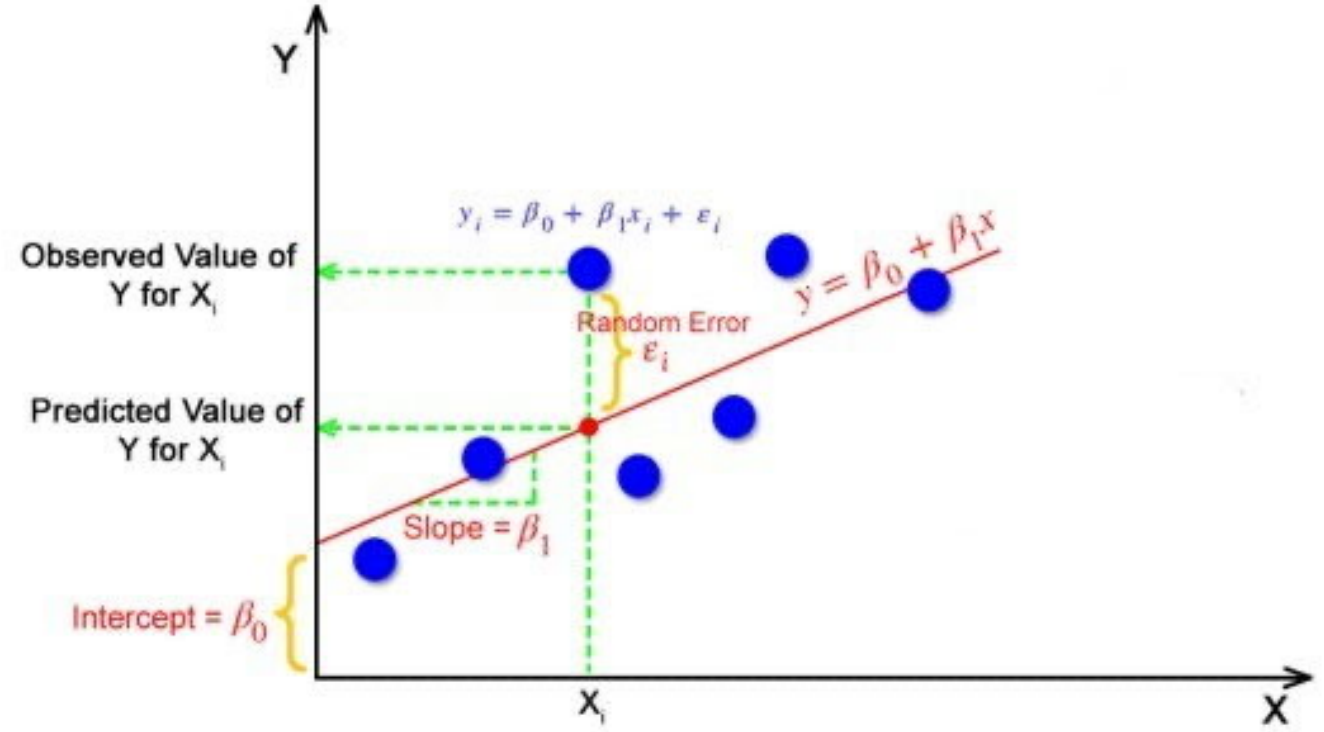
$$y = \alpha + \beta x$$

β = slope

α = y-intercept

y = y- coordinate

x = x-coordinate



Ordinary Least Squares

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

Residual (ε) = $y - \hat{y}$

Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

- we must find values of a and b that minimise

$$\Sigma (y - \hat{y})^2$$

Thank you