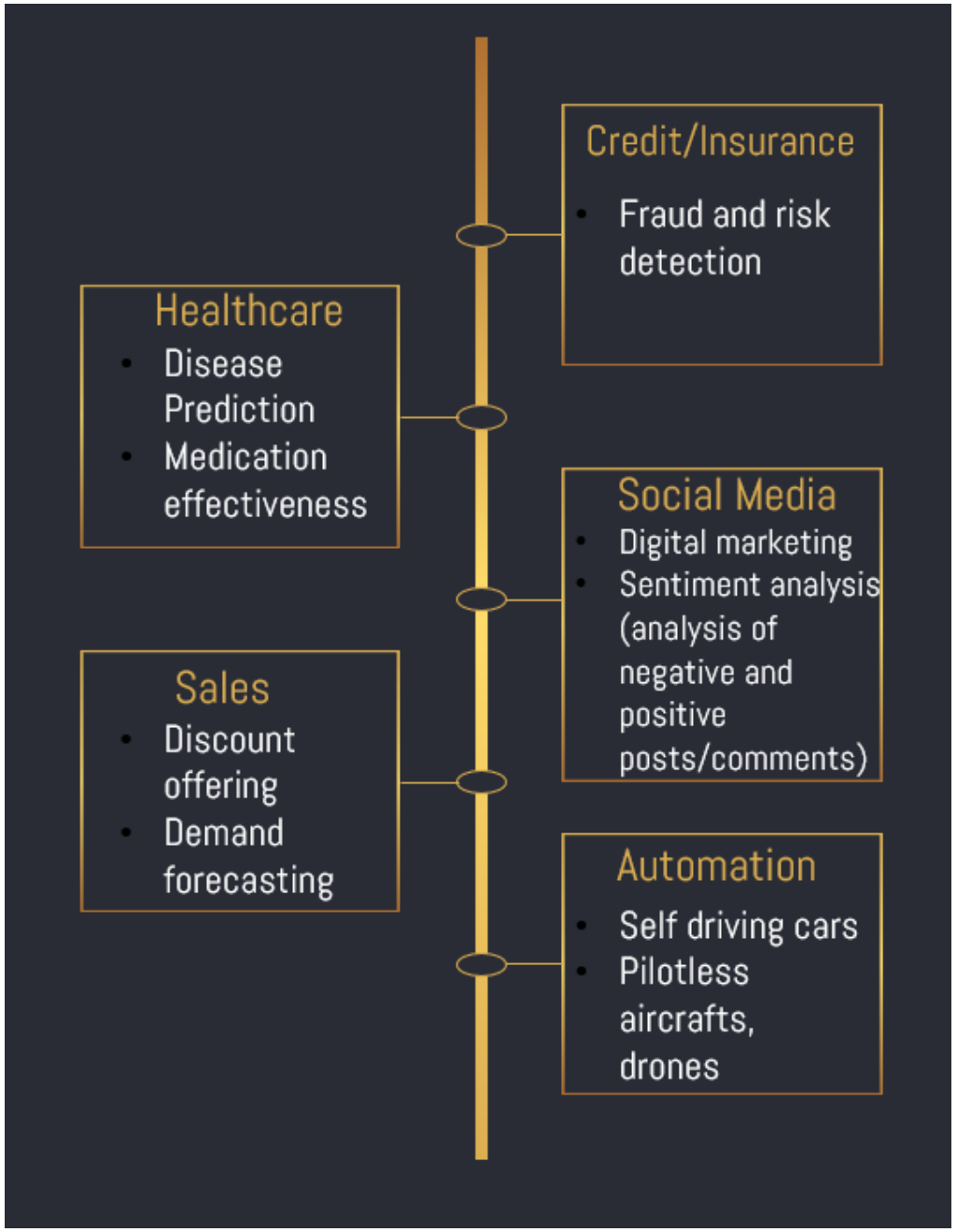
Data Science Track

DAY 1

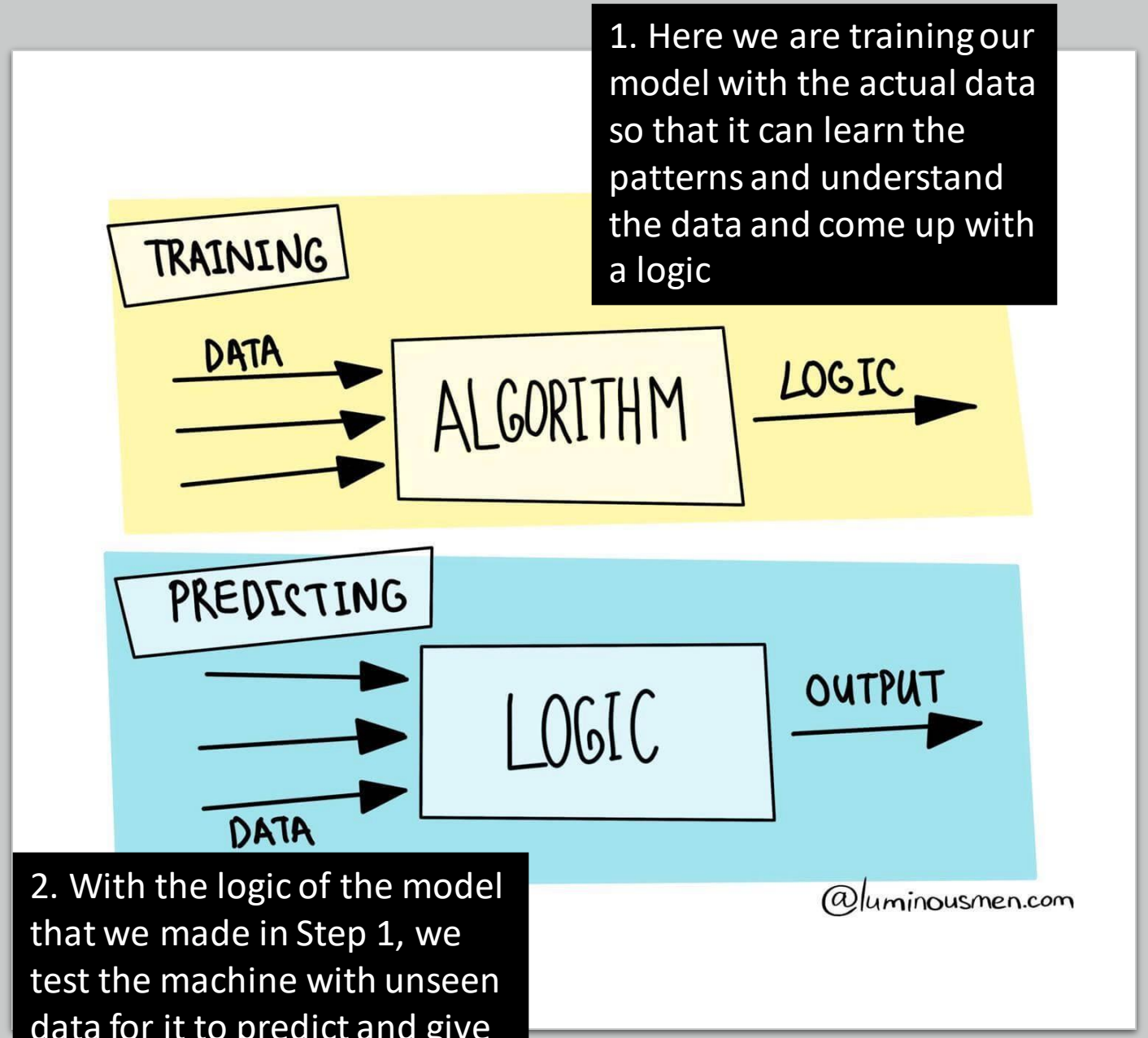
Applications of Data Science in the real world



Introduction to Machine Learning

What is Machine Learning?

- **It is the capability of a machine to imitate intelligent human behavior.**

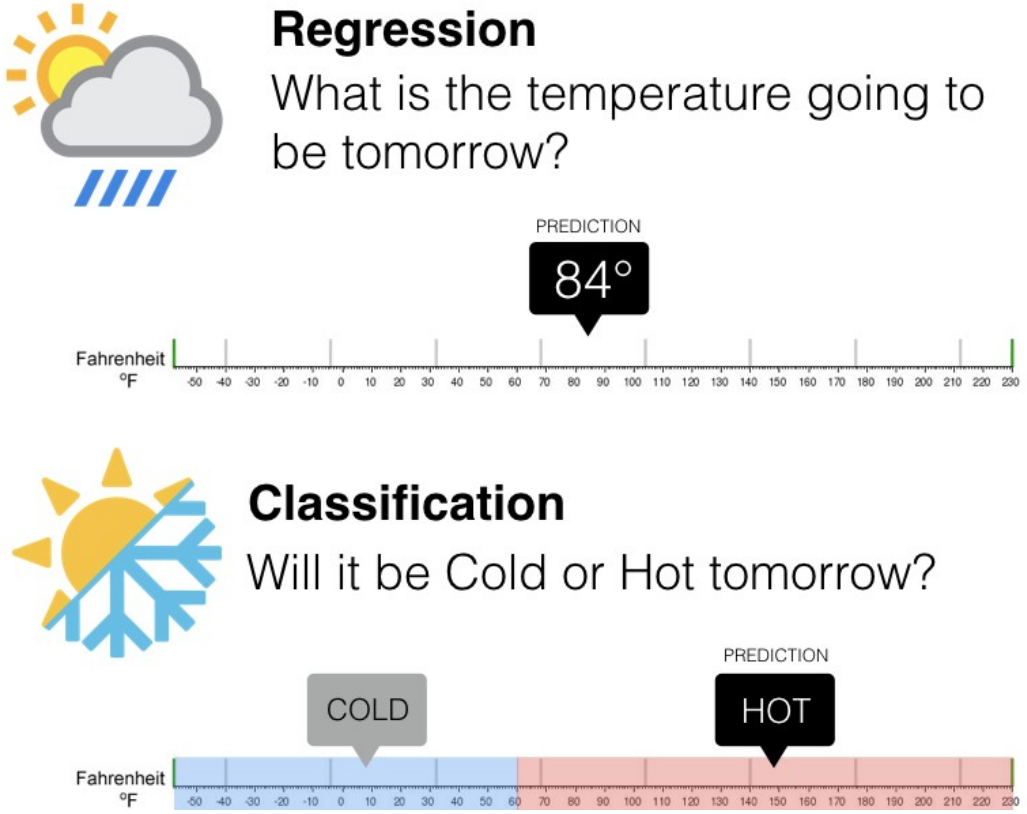


Types of Machine Learning algorithms

- https://www.youtube.com/watch?v=4dwsSz_fNSQ Watch this video for more insight

What is classification/regression?

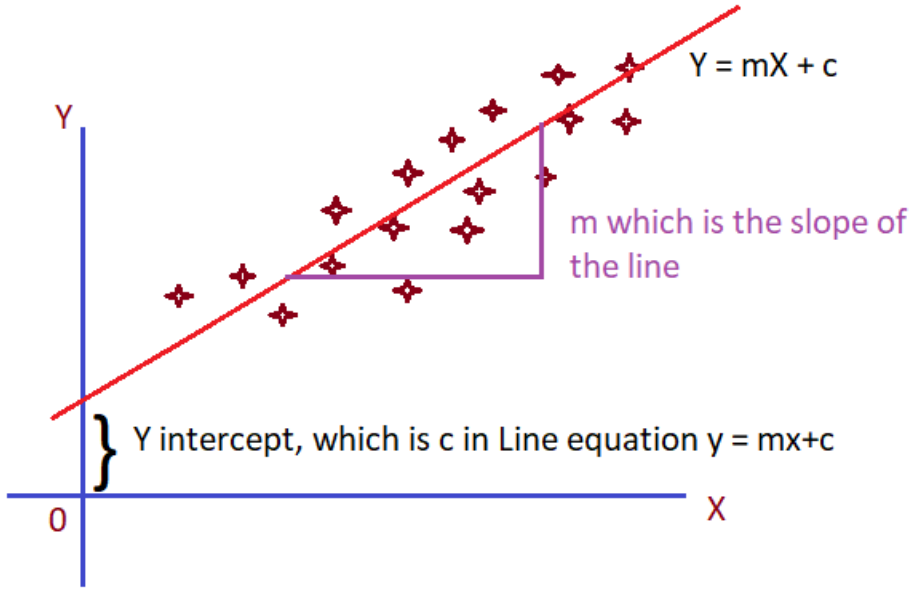
- In Regression, the output variable must be of **continuous nature** or real value
- In Classification, the output variable must be a **discrete value**.
- For more information - https://www.youtube.com/watch?v=9uHupa89LXE



DAY 2

What is Linear Regression?

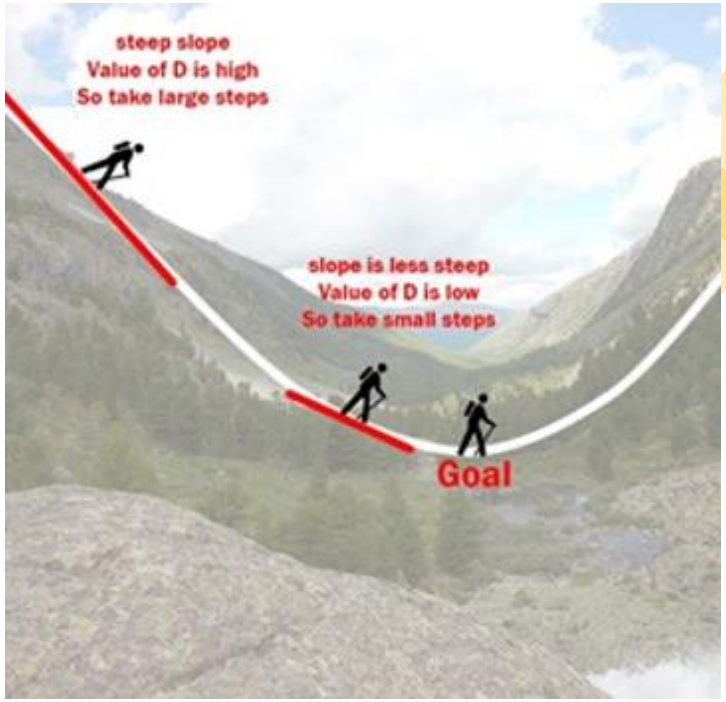
- In statistics, **linear regression** is a linear approach for modelling the relationship between dependent (x variables) and independent variables (y variable)
- As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable that we want to predict).Below is the equation to the linear regression model, we all are well aware of

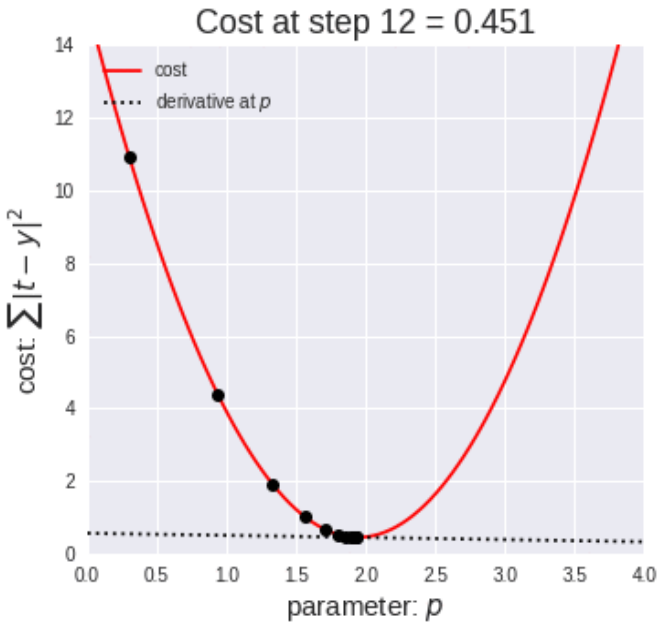


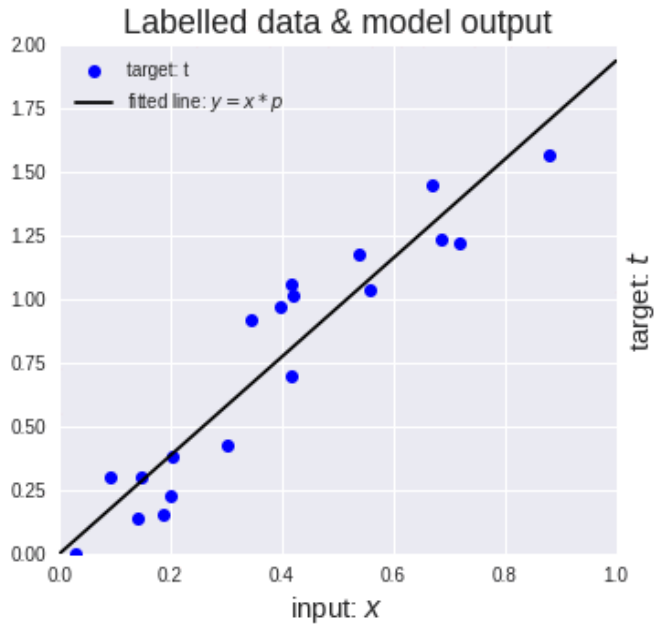
For more information: https://www.youtube.com/watch?v=ZkjP5RJLQF4

Basics of Gradient Descent

- The Gradient Descent algorithm determines the values of m and c, such that line corresponding to those values is best fitting line gives minimum error.
- Until the function is close to or equal to zero, the model will continue to adjust its parameters to yield the smallest possible error.







The left graph is the gradient descent that helps minimize the error in our prediction
The graph on the right is the linear regression line that shows the best fit line y (predicted points)

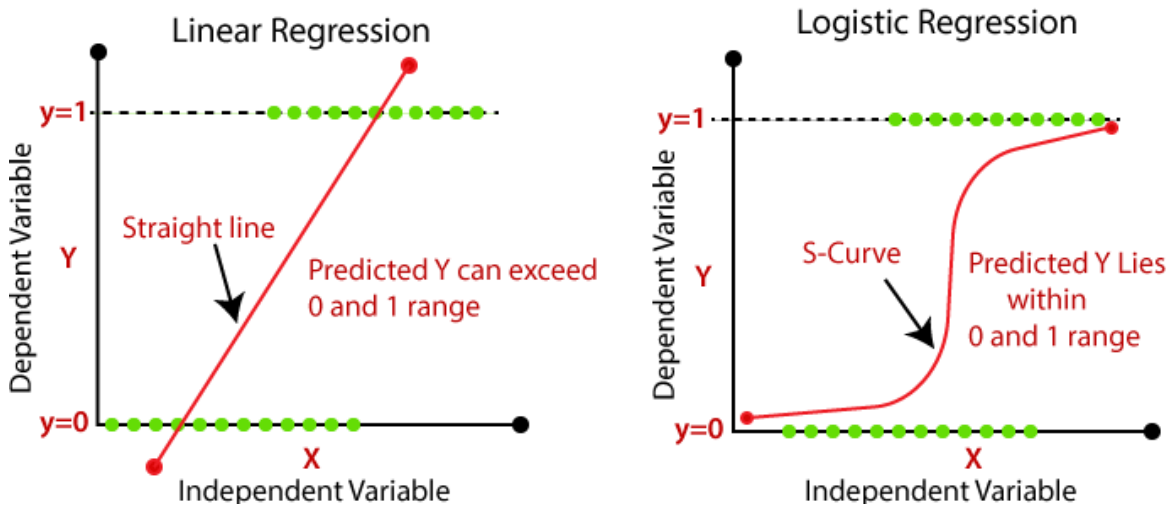
DAY 3

Logistic Regression

- Logistic regression is **a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a dataset.**
- Though it can also solve regression problems, it mostly suitable for solving **classification problems which brings linear regression to a disadvantage.**

Why linear regression is not suitable for classification?

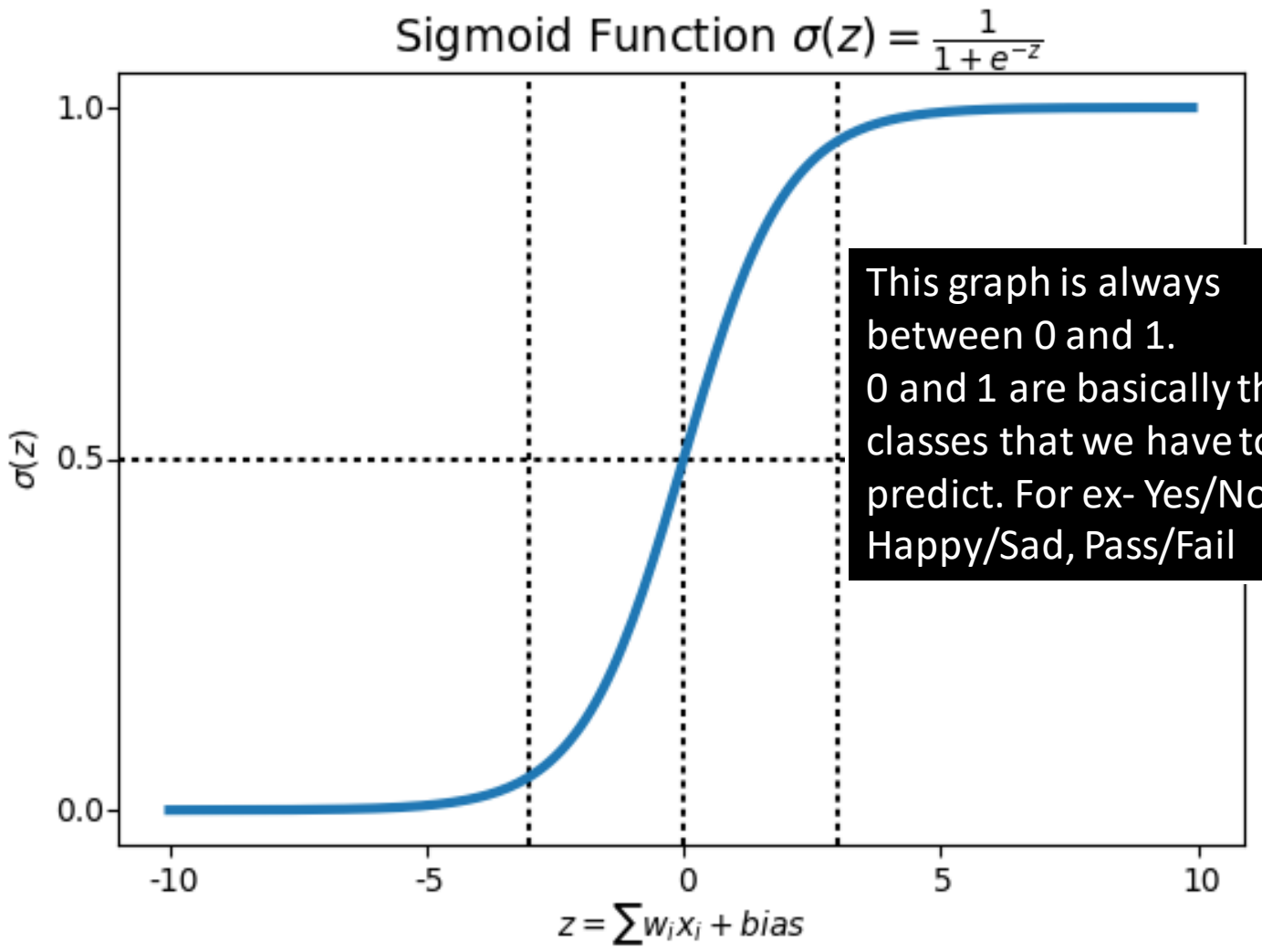
- Linear regression deals with continuous values while classification problems require **discrete** values.
- Linear regression is a great algorithm but it is **highly impacted by *outliers**. Hence we cannot use it to solve a classification problem.



*Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph

How does logistic regression help in solving classification problems?

- It makes the use of **sigmoid function** to find the final outcome.
- Sigmoid is a mathematical function that takes any real number and maps it to a probability between 1 and 0
- This helps in predicting 0 & 1 binary outcomes for classification



This graph is always between 0 and 1.
0 and 1 are basically the classes that we have to predict. For ex- Yes/No, Happy/Sad, Pass/Fail

For more clarity refer- https://www.youtube.com/watch?v=VImxF-9jk1E&t=3s

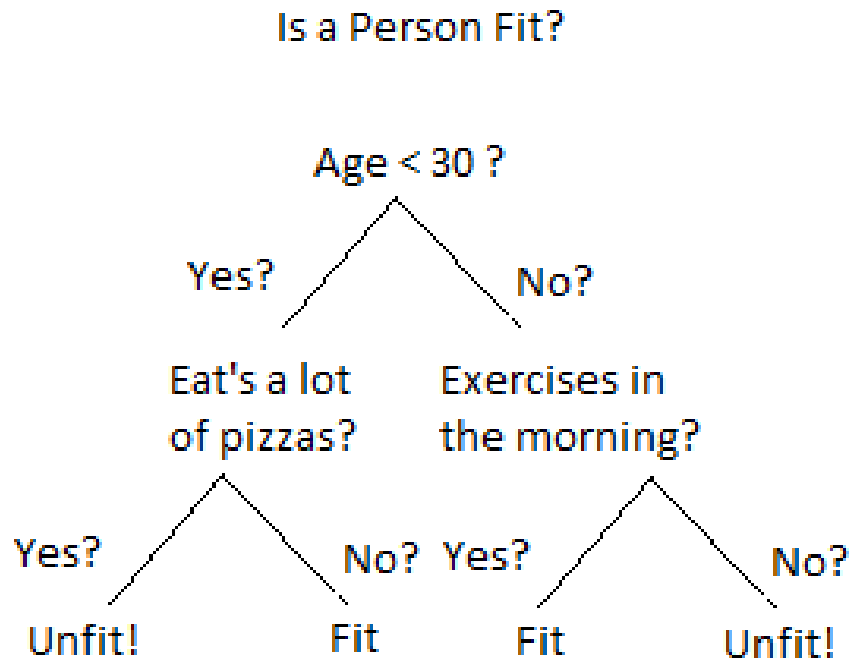
DAY 4

For in-depth information- https://www.youtube.com/watch?v=7VeUPuFGJHk

Decision Trees

- It is a supervised ML approach can be used for both classification and regression problems.
- Here the data is continuously split according to a certain parameter, like a tree with nodes.

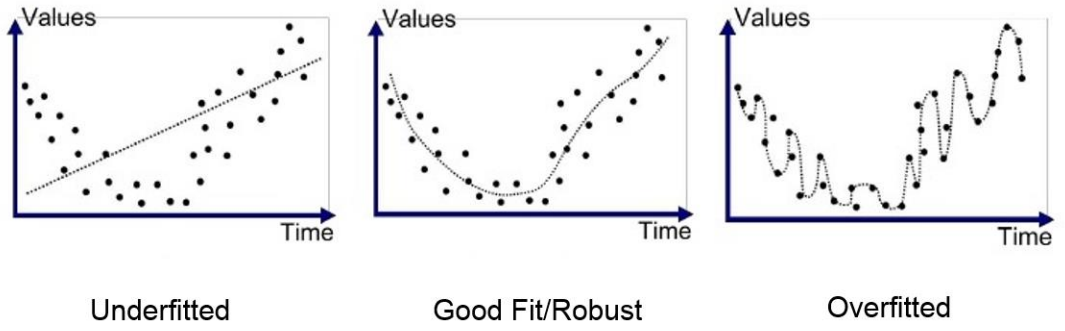
Here the model is trying to predict whether the person is fit or not according parameters like age, exercise, food intake. These values are given in the data

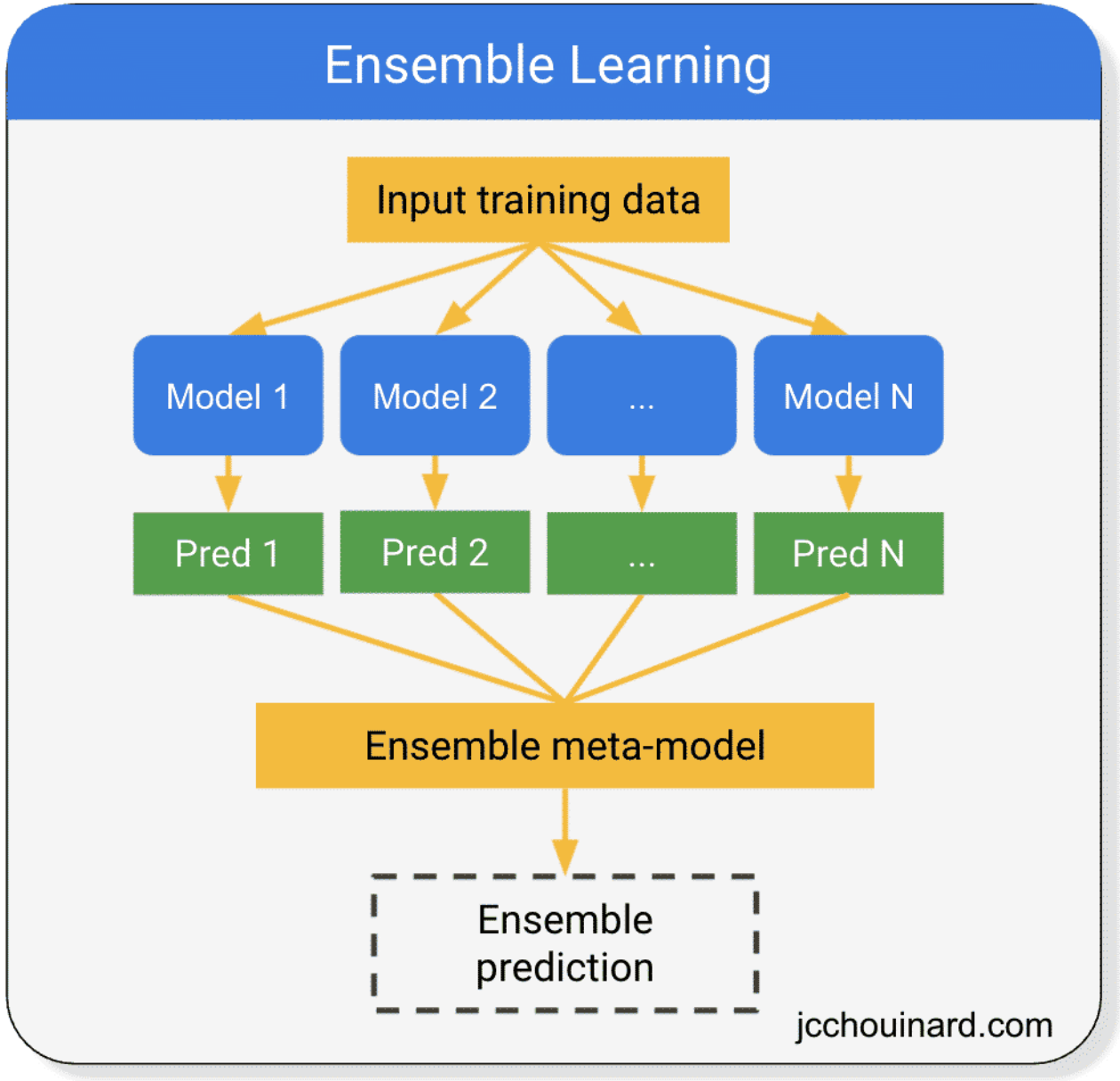


Disadvantages of Decision trees

Over-fitting occurs when the training model tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the new data (that are not a part of training data)

- ***Decision trees lead to overfitting.***
- ***Tree is designed so as to perfectly fit all samples in the training data set.***
- ***Thus this effects the accuracy when predicting samples that are not part of the training set.***





Ensemble Learning

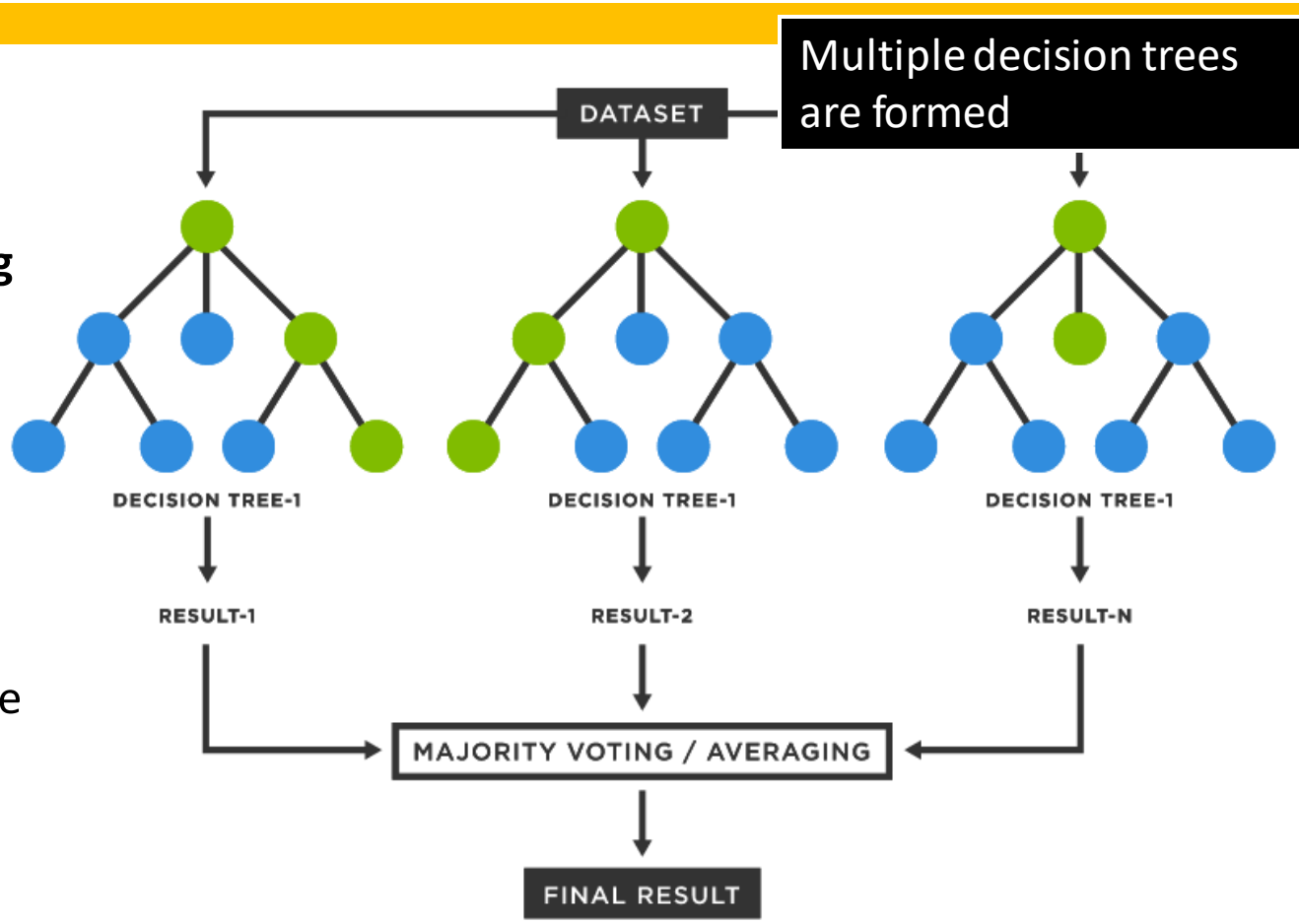
- **Ensemble learning** is a general approach to machine learning performs better by combining the predictions from multiple models so as to come to a final outcome.
- **It overcomes the problem of decision trees.**
- There are 2 types of ensemble learning: **Bagging** and **Boosting**.

For more information: https://www.youtube.com/watch?v=KIOeZ5cFZ50

Random Forest

- Random Forest comes under the **Bagging** category of ensemble learning.

- It is a ***Supervised Machine Learning Algorithm*** that is ***used in Classification and Regression problems***. It builds decision trees on different samples (of training data) and takes their **majority vote** for classification and **average** in case of regression.



DAY 5

UNSUPERVISED LEARNING

Clustering, K-Means

- Refer to the following ipynb file for content and implementation-
- https://colab.research.google.com/drive/1rLL5XA4DggeEOEqICFj80PDCCNFshLWu#scrollTo=HJhRX1rjR8MW