

Data-Driven Modelling Using Machine Learning: Hands-on with Orange Data Mining

Instructor:

Assoc Prof Dr Mahmud Iwan Solihin

AI workshop (2 July 2024 at GBS)



Objectives

- Understand various machine learning algorithms.
- Understand the types of problems and performance metrics.
- Learn the basics of data preparation and analysis.
- Learn how to implement the machine learning model training using Orange Data Mining (non-coding/visual programming) on various datasets.
- Gain fundamental knowledge for self-exploration and research.

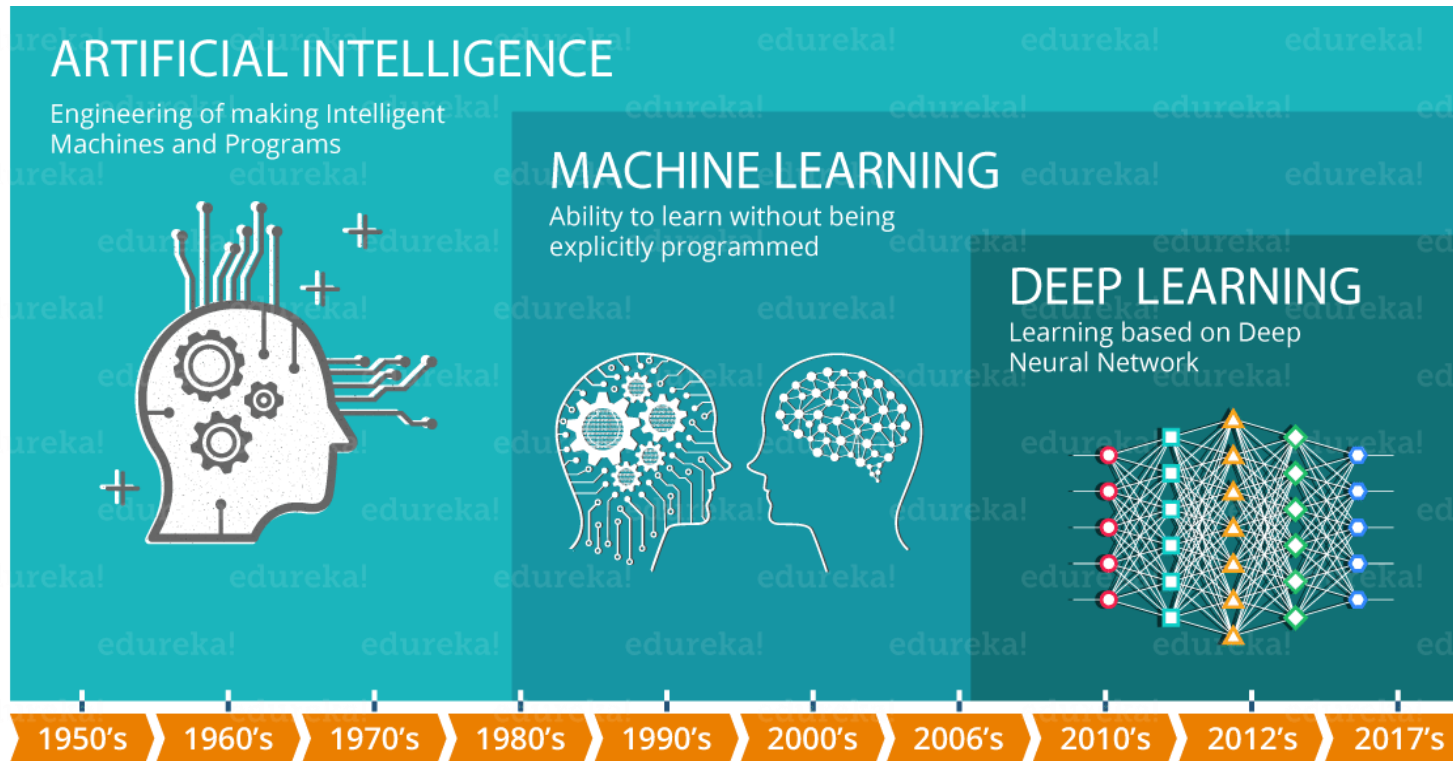
Workshop delivery

- We use Orange data mining software (quick and non-coding) when explaining the fundamental concept of Machine Learning. Free download:

<https://orangedatamining.com/download/>

- Machine Learning and applications with case studies in classification and regression.

Overview of ML and DL



- ❑ ML: A branch/subset of **AI**, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. → **data-driven decision model**
- ❑ As intelligence (like human) requires knowledge, it is necessary for the computers to acquire knowledge through learning. → **learning/training process**

Overview of ML and DL

Types of Learning Task

- **Supervised learning** (output/target values or labels are available)
 - Regression, including time-series (real values)
 - Classification (discrete labels)
- **Unsupervised learning** (NO output labels/values)
 - Clustering
 - Probability distribution estimation
 - Finding association (in features)
 - Dimension reduction
- **Reinforcement learning**
 - Decision making (robot, chess machine)

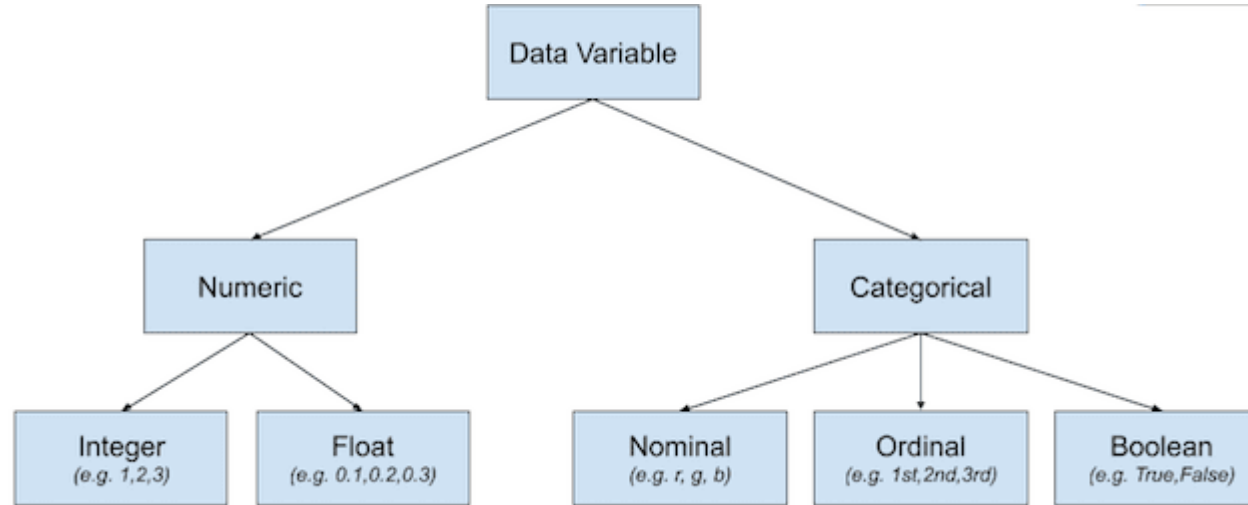


Main state-of-the-art

Overview of ML and DL

Data types

- The main ingredient of ML/DL modelling is data.



- Ordinal (ordered) → exam grade: A,B,C,D; quality grade: H,M,L
- Nominal (no order) → gender: M, F; blood type: A, B, AB, O
- Continuous/float: blood pressure, temperature
- Discrete/integer: accidents number, students number
- Boolean: whether animal has tail or not (YES, NO)

Overview of ML and DL

Types of learning task. Regression or classification?

- Estimating MPG (mile per gallon) of car's fuel consumption from car attributes (features) such as 'horse power', 'weight', 'number of piston', 'country of origin'. →
- Categorizing types of animal based on e.g.: 'number of leg', 'tailed or not', 'haired or not', 'milked of not' etc. →
- Identifying person emotions based on facial expression. →
- Predicting house price based on e.g. 'its location', 'number of room', 'sq ft of area', 'distance to station'. →

Overview of ML and DL

Types of learning task. Regression or classification?



Article

Modelling of River Flow Using Particle Swarm Optimized Cascade-Forward Neural Networks: A Case Study of Kelantan River in Malaysia

Gasim Hayder ^{1,2,*} , Mahmud Iwan Solihin ³ and Hauwa Mohammed Mustafa ^{4,5,*}

- ¹ Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor, Malaysia
- ² Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor, Malaysia
- ³ Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur 56000, Malaysia; mahmudi@ucsiuniversity.edu.my

Received: 16 January 2022 | Revised: 22 March 2022 | Accepted: 19 April 2022
DOI: 10.1111/jpp.16686

ORIGINAL ARTICLE



WILEY

Physicochemical properties and detection of glucose syrup adulterated Kelulut (*Heterotrigona itama*) honey using Near-Infrared spectroscopy

Venecia Woeng¹ | Lee Ying Lim¹ | Lejaniya Abdul Kalam Saleena¹ | Mahmud Iwan Solihin² | Liew Phing Pui¹

¹Department of Food Science and Nutrition, Faculty of Applied Sciences, UCSI University, Kuala Lumpur, Malaysia

²Department of Mechanical Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, Malaysia

Correspondence

Liew Phing Pui, Department of Food Science and Nutrition, Faculty of Applied Sciences, UCSI University, Kuala Lumpur 56000, Malaysia.
Email: pui@ucsiuniversity.edu.my

Funding Information

The author(s) received financial support for the research, authorship, and/or publication of this article from FRGS/1/2020/TK0/UCSI/02/4.

Abstract

In this study, the physicochemical properties and the precision of adulterated honey detection by NIRS were assessed. A total of 11 adulteration ratio samples were produced with glucose syrup. The ash and moisture content of the honey samples were found to be in accordance with the Malaysian Standard, including all adulterated ones. The pH value and HMF content complied with the standard except for the pH of 100% adulterated honey and HMF content of 30–100% adulteration. NIR spectra were scanned at 900–1700 nm, 1848 spectra were recorded, and the data were pre-processed to reduce the unwanted spectral interference. Chemometric techniques and machine learning were used as prediction tools for adulteration levels. This includes principal component analysis (PCA), k-nearest neighbors (kNN) and random forest. The accuracy of the prediction samples was above 90%. Hence, this was satisfactory for rapid detection of Kelulut honey adulteration.

SPRINGER LINK

Find a journal | Publish with us | Track your research | Search

Home > Waste and Biomass Valorization > Article

Intelligent Kitchen Waste Composting System via Deep Learning and Internet-of-Things (IoT)

Original Paper | Published: 07 December 2023

Volume 15, pages 3133–3146, (2024) | [Cite this article](#)

Teh Boon Hong, Sarah Atifah Saruchi , Ain Atiqah Mustapha, Jonathan Lam Lit Seng, Ahmad Nor Alifa A. Razap, Nico Halisno, Mahmud Iwan Solihin & Nor Aziyatul Izni

157 Accesses | [Explore all metrics](#) →

Abstract

Kitchen waste is listed among the top global sustainability issue as it contributes to global warming and climate change. Composting is one of the solutions to tackle the issue of kitchen waste increment. However, a manual composting system has led to several problems for the waste management authorities to invest more in human labor, cost, and

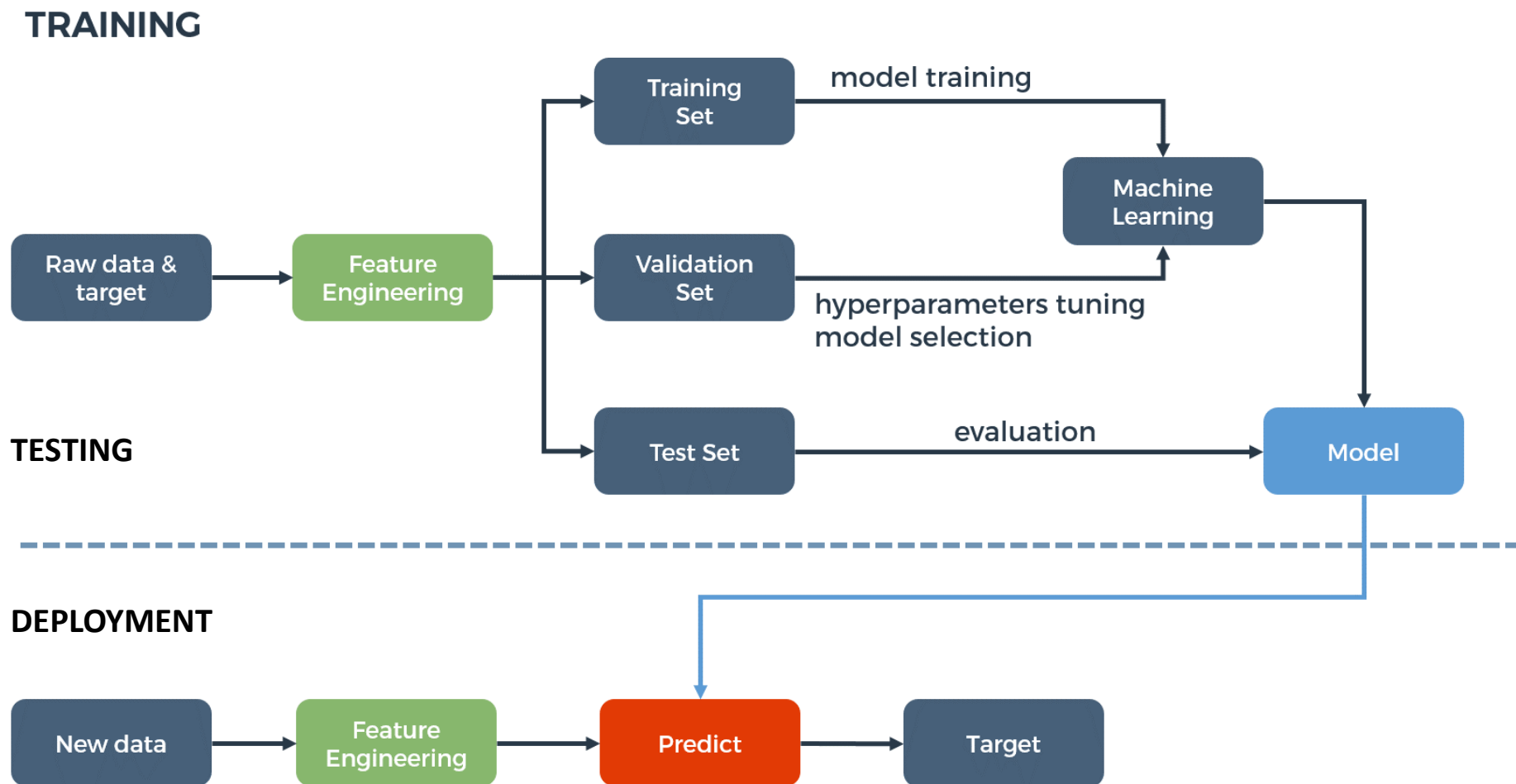
Overview of ML and DL

Tabular data representation

| Data No. | Meta data (if any) | Input/feature-1 | Input/feature-2 | Input/feature-3 | | Input/feature-n | Output/target/g round truth |
|----------|--------------------|-----------------|-----------------|-----------------|-------|-----------------|--------------------------------|
| Sample-1 | | | | | | | |
| Sample-2 | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| Sample-n | | | | | | | |

Overview of ML and DL

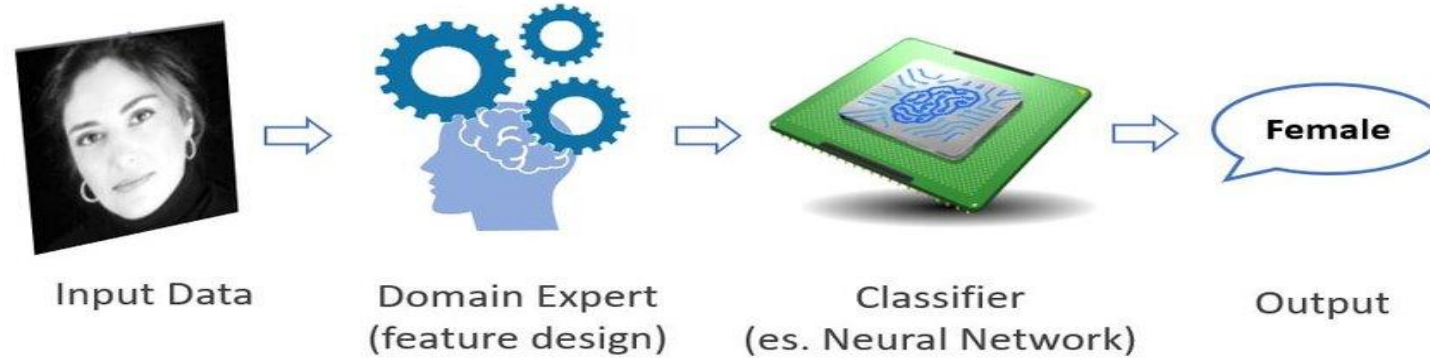
ML/DL modeling workflow



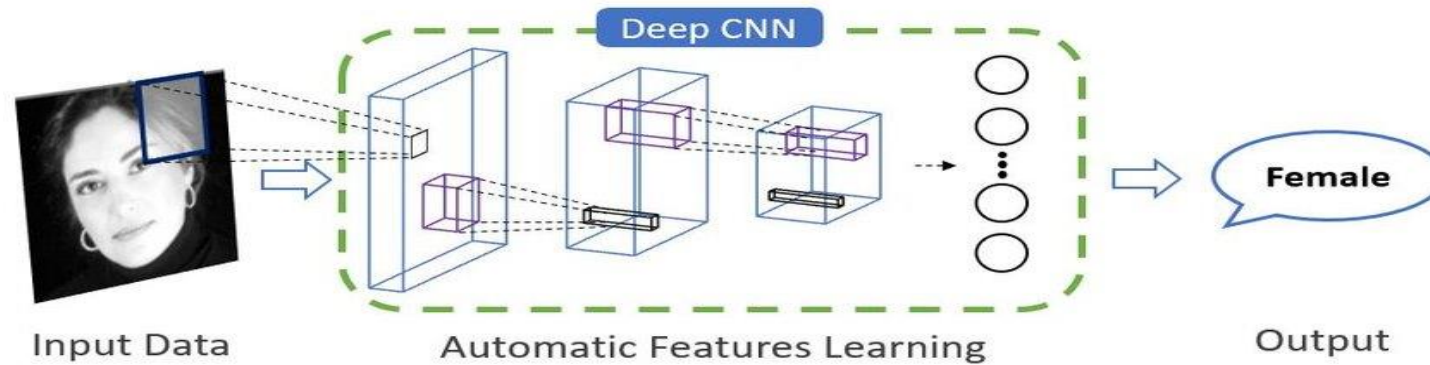
Overview of ML and DL

ML vs DL

Machine Learning



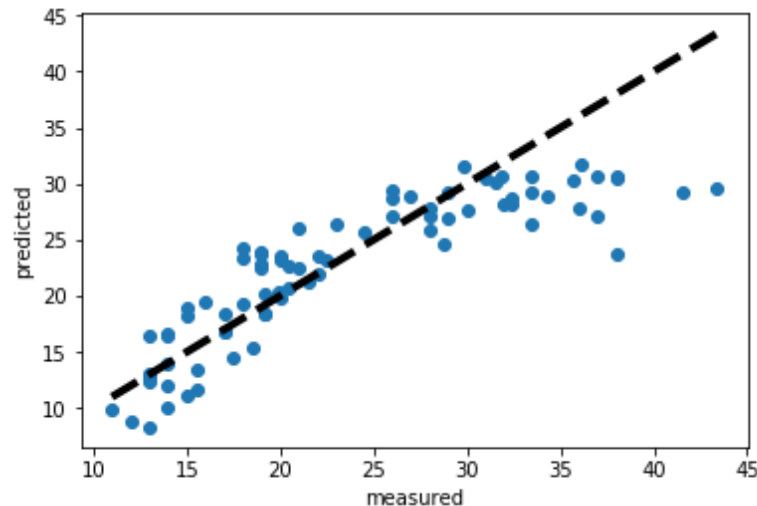
Deep Learning



Feature extractor+Neural Networks

Performance metric: regression

- The regression model is evaluated by error of prediction (smaller is better) with metrics e.g. RMSE (root mean squared error), MSE and MAE (mean absolute error).
- R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.



$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

\hat{y} – predicted value of y

\bar{y} – mean value of y

Performance metric: classification





- The classifier model is evaluated by using **confusion matrix** and the metrics derived from it, i.e. Accuracy, Precision, Recall (Sensitivity) and F-1 Score .

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

| | | Predicted class | |
|--------------|---|--|---|
| | | P | N |
| Actual Class | P | True Positives (TP) <i>Sensitivity Recall</i> | False Negatives (FN) |
| | N | False Positives (FP) <i>Precision</i> | True Negatives (TN) <i>Specificity</i> |

| | | Predicted Class | | |
|--------------|----------|--|--|--|
| | | Positive | Negative | |
| Actual Class | Positive | True Positive (TP) | False Negative (FN) Type II Error | Sensitivity $\frac{TP}{(TP + FN)}$ |
| | Negative | False Positive (FP) Type I Error | True Negative (TN) | Specificity $\frac{TN}{(TN + FP)}$ |
| | | Precision $\frac{TP}{(TP + FP)}$ | Negative Predictive Value $\frac{TN}{(TN + FN)}$ | Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$ |

Performance metric: classification

| | | PREDICTIVE VALUES | |
|---------------|----------------|---|--|
| | | POSITIVE (CAT) | NEGATIVE (DOG) |
| ACTUAL VALUES | POSITIVE (CAT) | <p>TRUE POSITIVE</p>  <p>3</p> | <p>FALSE NEGATIVE</p>  <p>1</p> <p>TYPE II ERROR</p> |
| | NEGATIVE (DOG) | <p>FALSE POSITIVE</p>  <p>2</p> <p>TYPE I ERROR</p> | <p>TRUE NEGATIVE</p>  <p>4</p> |

Accuracy:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) = (3+4)/(3+4+2+1) = 0.70$$

Recall: Recall gives us an idea about when it's actually yes, how often does it predict yes.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 3/(3+1) = 0.75$$

Precision: Precision tells us about when it predicts yes, how often is it correct.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 3/(3+2) = 0.60$$

F1-score: F1 score is a measure of the harmonic mean of precision and recall. It will be useful when we have imbalanced data in class distribution.

$$\text{F1-score} = (2 * \text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) = (2 * 0.75 * 0.60) / (0.75 + 0.60) = 0.67$$

Imbalanced data in classification

- Accuracy is **not a good metric** when dealing with imbalanced data.

| Real \ Prediction | positive | negative |
|-------------------|----------|----------|
| | positive | negative |
| positive | 5 | 5 |
| negative | 10 | 990 |

Accuracy = $995 / 1010 = 0.98$ (**isn't that high?**)

Precision = $5 / 15 = 0.33$

Recall/Sensitivity = $5 / 10 = 0.5$

Specificity = $990 / 1000 = 0.99$

F1-score = $2 * (0.5 * 0.33) / (0.5 + 0.33) = 0.4$
(but F-1 score is low, model is bad)

- Beside F1-score, Balanced Accuracy is better metric in imbalanced data.

$$\text{Balanced Accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

Balanced Accuracy = $(0.5 + 0.99) / 2 = 0.745$ (telling us that the model is not that good)

Multi-class classification

| | | Actual Classes | | | |
|-------------------|---|----------------|---|---|---|
| | | a | b | c | d |
| Predicted Classes | a | 50 | 3 | 0 | 0 |
| | b | 26 | 8 | 0 | 1 |
| | c | 20 | 2 | 4 | 0 |
| | d | 12 | 0 | 0 | 1 |

In the multi-class classification problem, TP, TN, FP, and FN are calculated for each class. Then average of metrics over all classes can be obtained. There is micro, macro and weighted averaging.

$$\text{Precision}(\text{class} = a) = \frac{TP(\text{class} = a)}{TP(\text{class} = a) + FP(\text{class} = a)} = \frac{50}{53} = 0.943$$

$$\text{Recall}(\text{class} = a) = \frac{TP(\text{class} = a)}{TP(\text{class} = a) + FN(\text{class} = a)} = \frac{50}{108} = 0.463$$

$$\text{F-1 Score}(\text{class} = a) = \frac{2 \times \text{Precision}(\text{class} = a) \times \text{Recall}(\text{class} = a)}{\text{Precision}(\text{class} = a) + \text{Recall}(\text{class} = a)} = \frac{2 \times 0.943 \times 0.463}{0.943 + 0.463} = 0.621$$

$$\text{F-1 Score}(\text{class} = b) = \frac{2 \times 0.228 \times 0.615}{0.228 + 0.615} = 0.333$$

$$\text{F-1 Score}(\text{class} = c) = \frac{2 \times 0.154 \times 1.000}{0.154 + 1.000} = 0.267$$

$$\text{F-1 Score}(\text{class} = d) = \frac{2 \times 0.077 \times 0.500}{0.077 + 0.500} = 0.133$$

<https://www.baeldung.com/cs/multi-class-f1-score>

<https://vitalflux.com/micro-average-macro-average-scoring-metrics-multi-class-classification-python/>

CASE STUDIES WITH MACHINE LEARNING

Case study: Regression ('carsmall' data)

- Common initial analysis: statistical description, visualization and feature selection (e.g. using correlation analysis)
- We will see correlation analysis as feature selection.

Case study: Classification

- Statistical description (boxplot), visualization (histogram distribution)
- Common dataset is usually perfectly balance. What if the data is extremely imbalance?
We may use up-sampling or down-sampling to balance the class distribution.

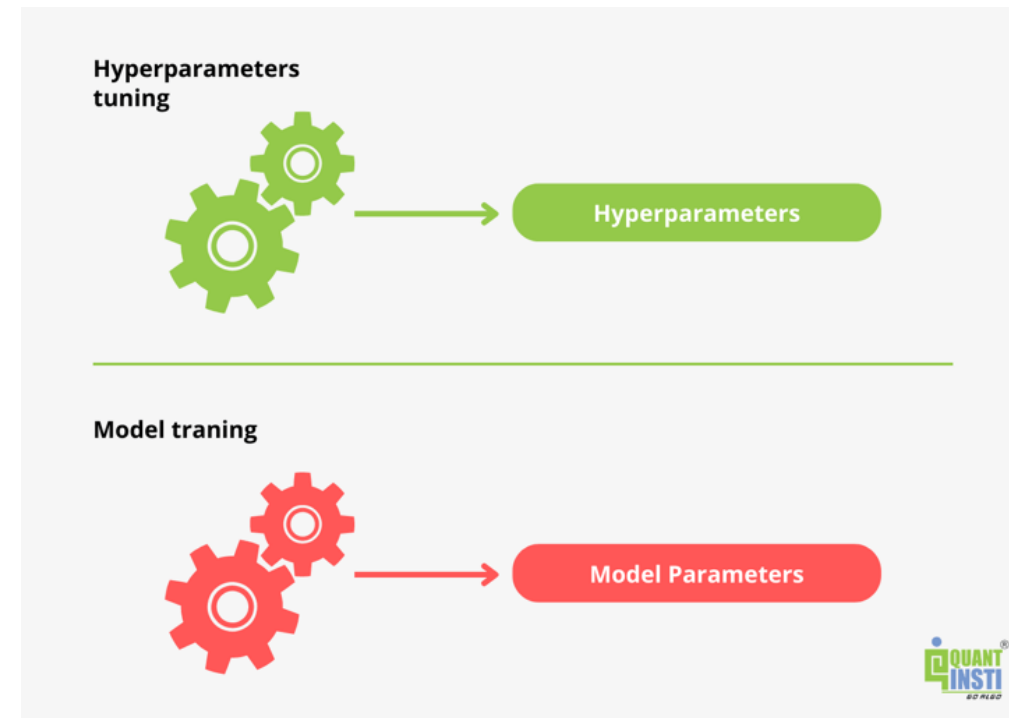
Case study: Landslide Susceptibility Mapping using ML

<https://pubs.aip.org/aip/acp/article-abstract/2482/1/050006/2867377/The-effect-of-spatial-scales-and-imbalanced-data> (ICETIR paper)

We will look at MCC rather than Accuracy, F1 score:

[The advantages of the Matthews correlation coefficient \(MCC\) over F1 score and accuracy in binary classification evaluation | BMC Genomics | Full Text \(biomedcentral.com\)](#)

ML Parameters (Trainable) vs Hyperparameters



- <https://doi.org/10.1016/j.neucom.2020.07.061>

Hyperparameters Tuning Example

- We can use Python-keras tuner

Lets use: NIR Spectroscopy data (Honey adulteration)

link: [iwanmahmud77/NIR-spectroscopy \(github.com\)](https://github.com/iwanmahmud77/NIR-spectroscopy)

- Sample of python coding (MLRegression_CKP_AH_optim.ipynb)

THANK YOU