A decision tree approach for sugarcane area classification

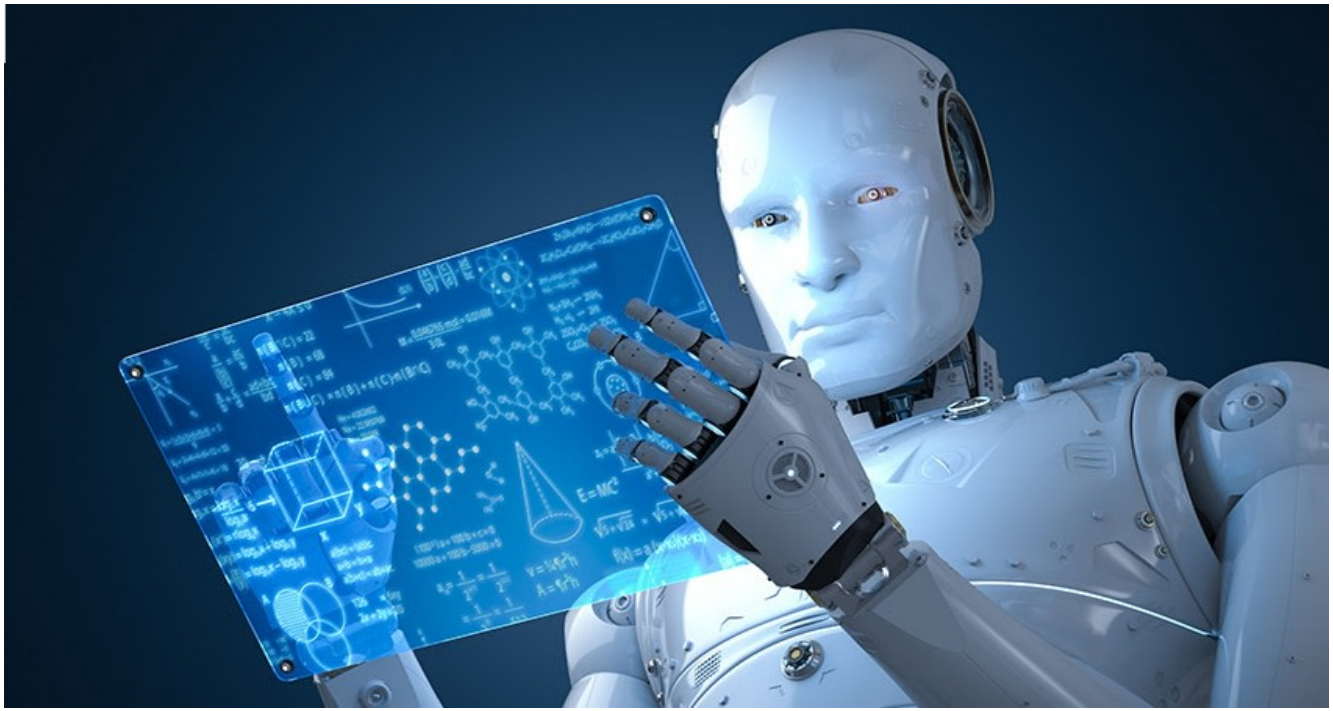
Clarice Demétrio Cristian Villegas Marcelo da Silva

Department of Exact Sciences, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba, Brazil

ESALQ

What is Artificial Intelligence? [1]

It is technology that enables computers and machines to simulate human intelligence and problem-solving capabilities.



Source: https://engineerera.home.blog/2019/08/01/artificial-intelligence/.

[1] https://www.ibm.com/topics/artificial-intelligence

ESALQ

Artificial Intelligence

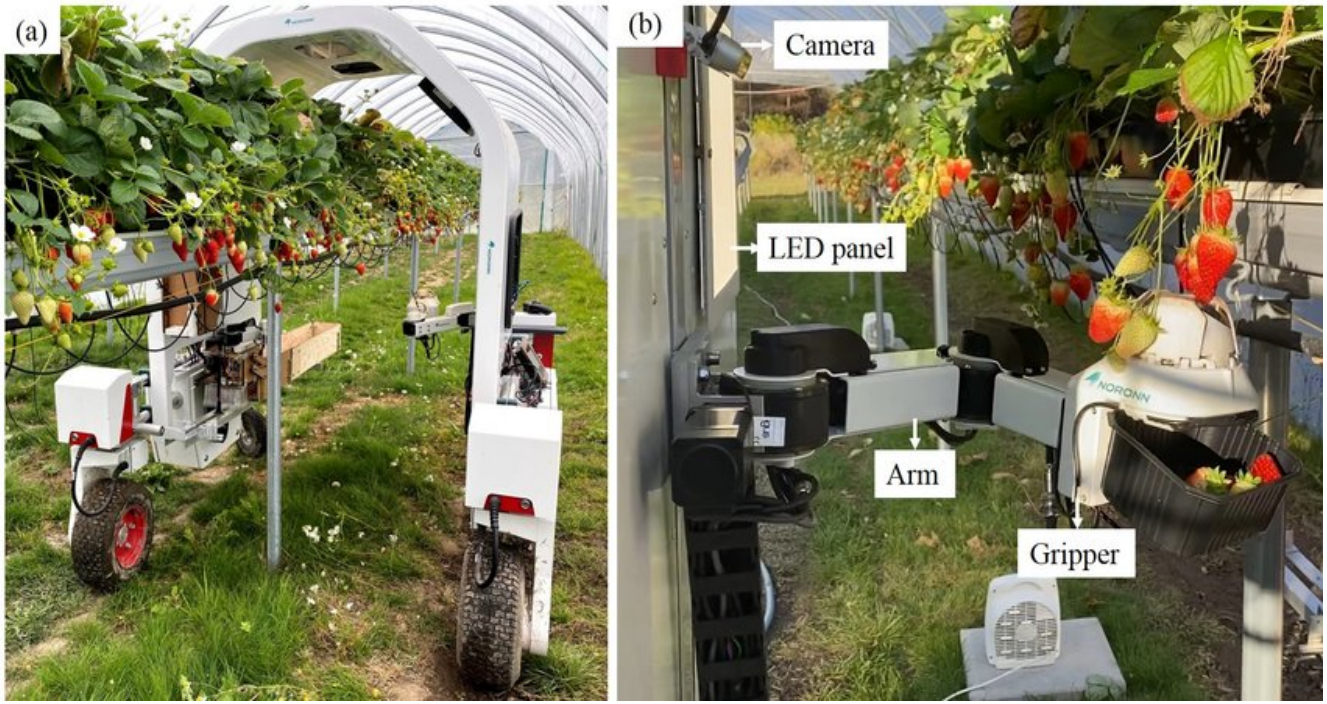
Recommendation systems, autonomous cars, virtual assistants

Source: https://caradas.com/adas-vs-autonomous-driving/.

ESALQ

Artificial Intelligence

Smart drones, harvesting robots, crop monitoring systems.



Source: https://images.app.goo.gl/9X23Z3XTMD3L2FYaA.

ESALQ

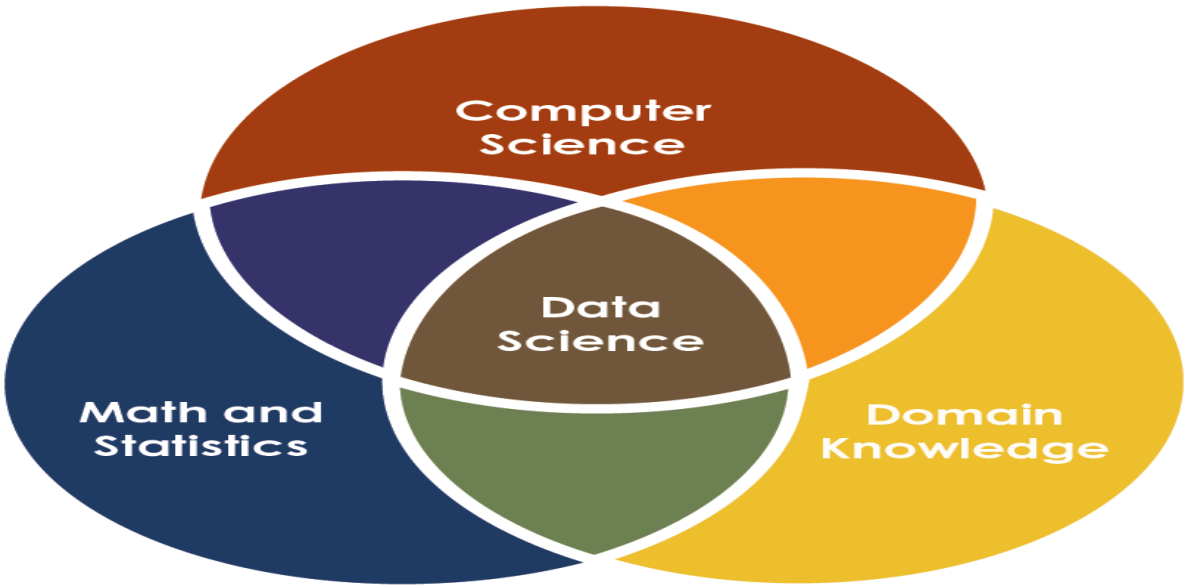
What is Machine Learning?

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy [2]

[2] https://www.ibm.com/topics/machine-learning

What is Data Science?

Multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data in various forms.



Source: https://www.datascience-pm.com/data-science-vs-software-engineering/.

ESALQ

What is Data Science?

Data collection, cleaning, analysis, modeling, and interpretation. Relevance to Agronomy: Transforming raw data (sensors, satellite images, climate data) into actionable information.

Source: https://ankageo.com/en/satellite-imagery-technology/.

ESALQ

Essentials for Data Science

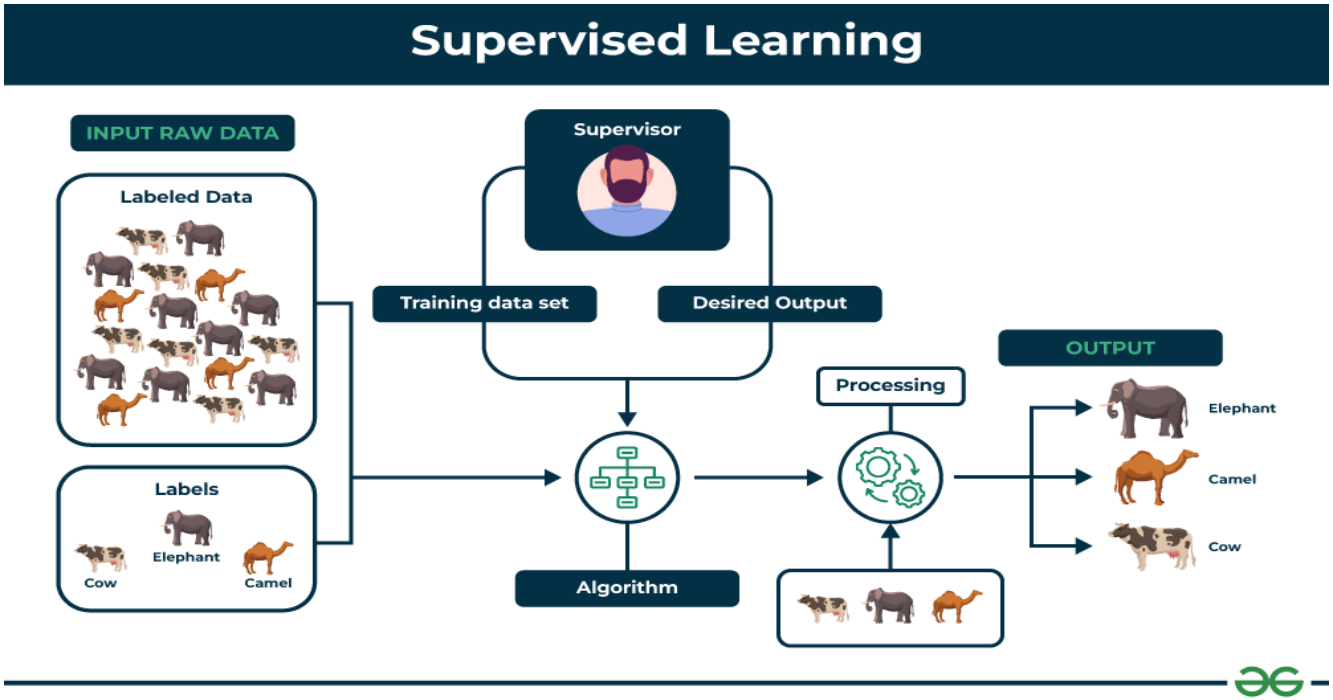
Skills:

1. Mathematics (linear algebra and calculus)
2. Statistics (EDA, probability and inference)
3. Programming (R, Python, etc)

ESALQ

Types of ML: Supervised Learning

Training with labeled data (input + expected output).

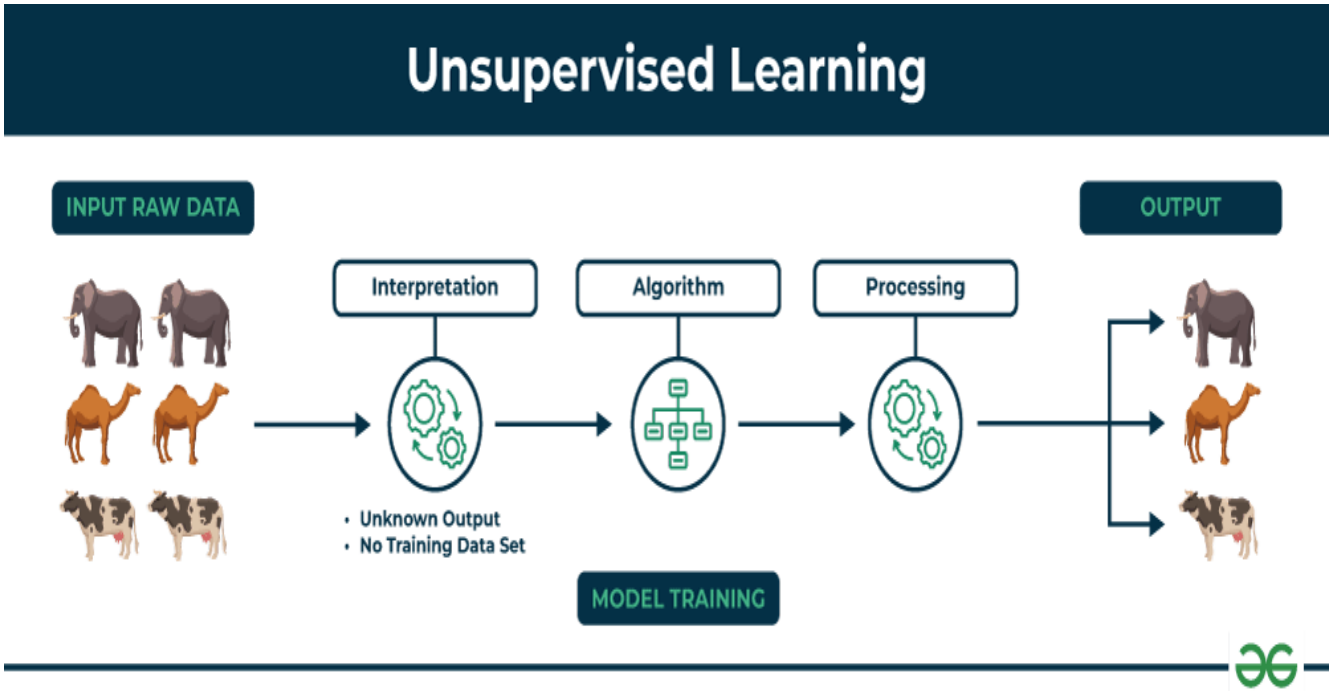


Source:https://www.geeksforgeeks.org/machine-learning/supervised-unsupervised-learning/.

ESALQ

Types of ML: Unsupervised Learning

Training with unlabeled data, searching for patterns and structures.

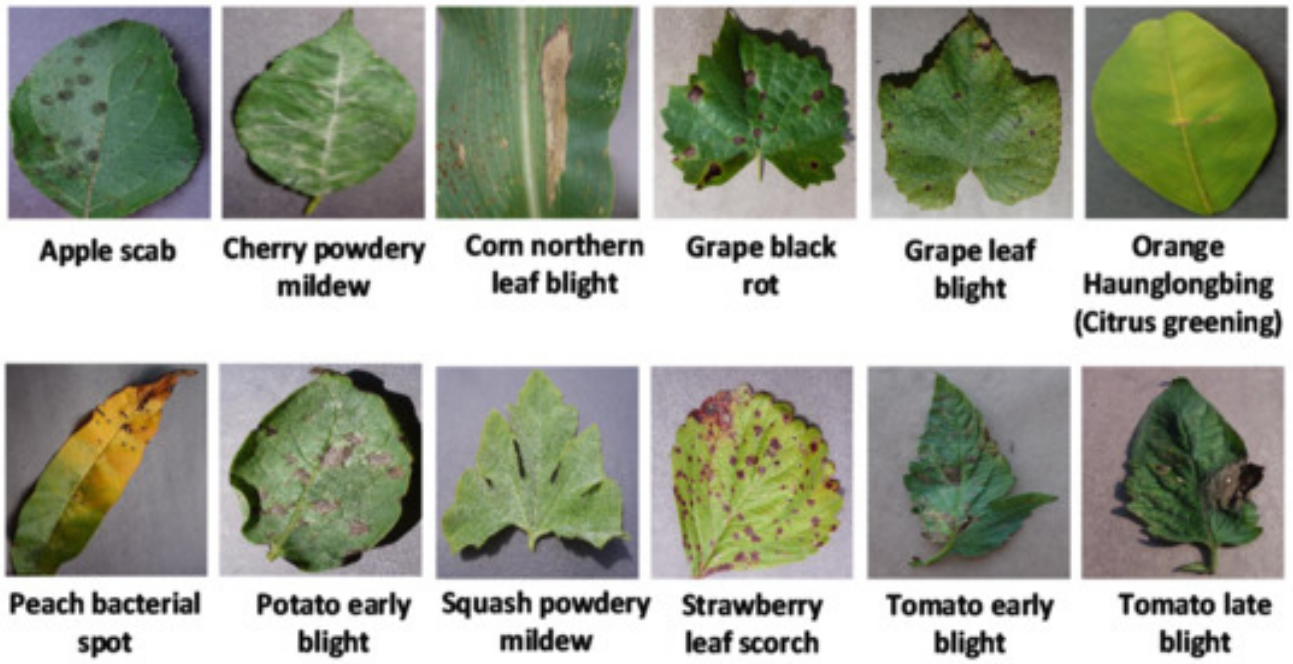


Source: https://www.geeksforgeeks.org/machine-learning/unsupervised-learning/.

ESALQ

Supervised Learning in Agronomy

- Plant disease classification: Training a model with images of healthy and diseased leaves to identify pathologies.

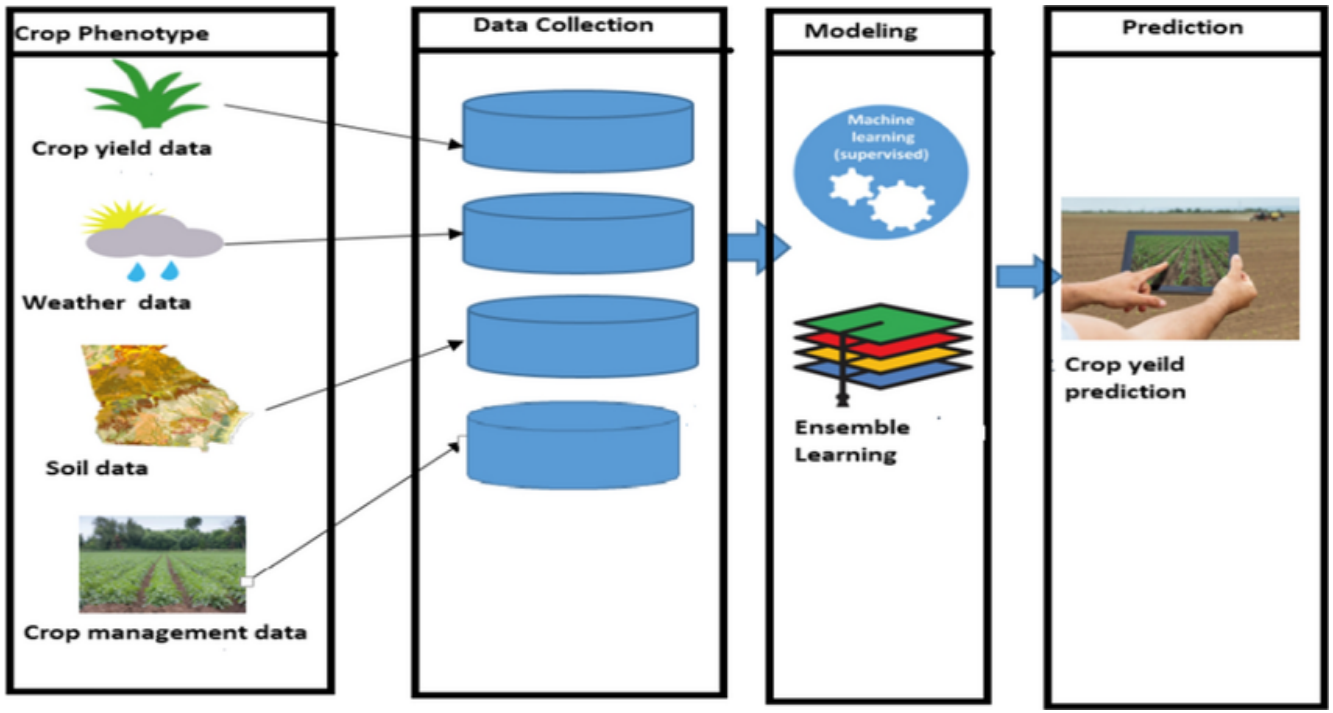


Source: https://www.mdpi.com/2223-7747/9/10/1319.

ESALQ

Supervised Learning in Agronomy

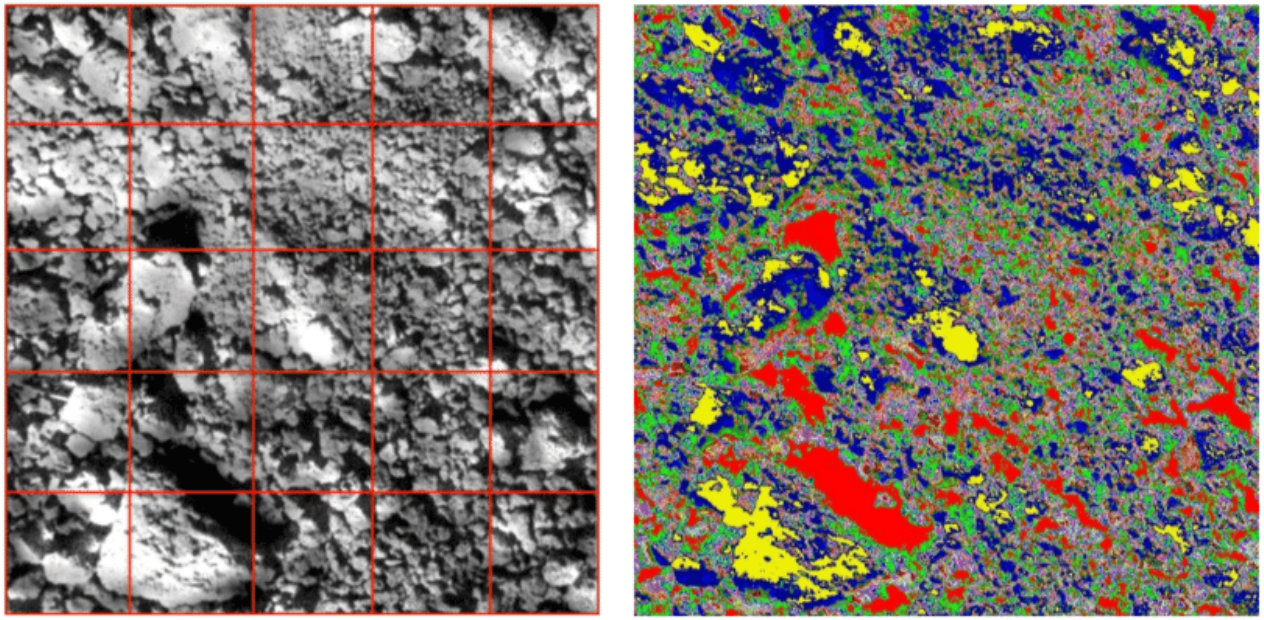
- Crop yield prediction: Using historical data on climate, soil type, and management to predict future yields.



Source: https://www.researchgate.net/publication/352866597/figure/fig6/AS:1143128108960069 diagram-of-crop-yield-prediction.png.

Unsupervised Learning in Agronomy

- Soil segmentation: Grouping different areas of a field based on soil characteristics (moisture, nutrients) for optimized management.

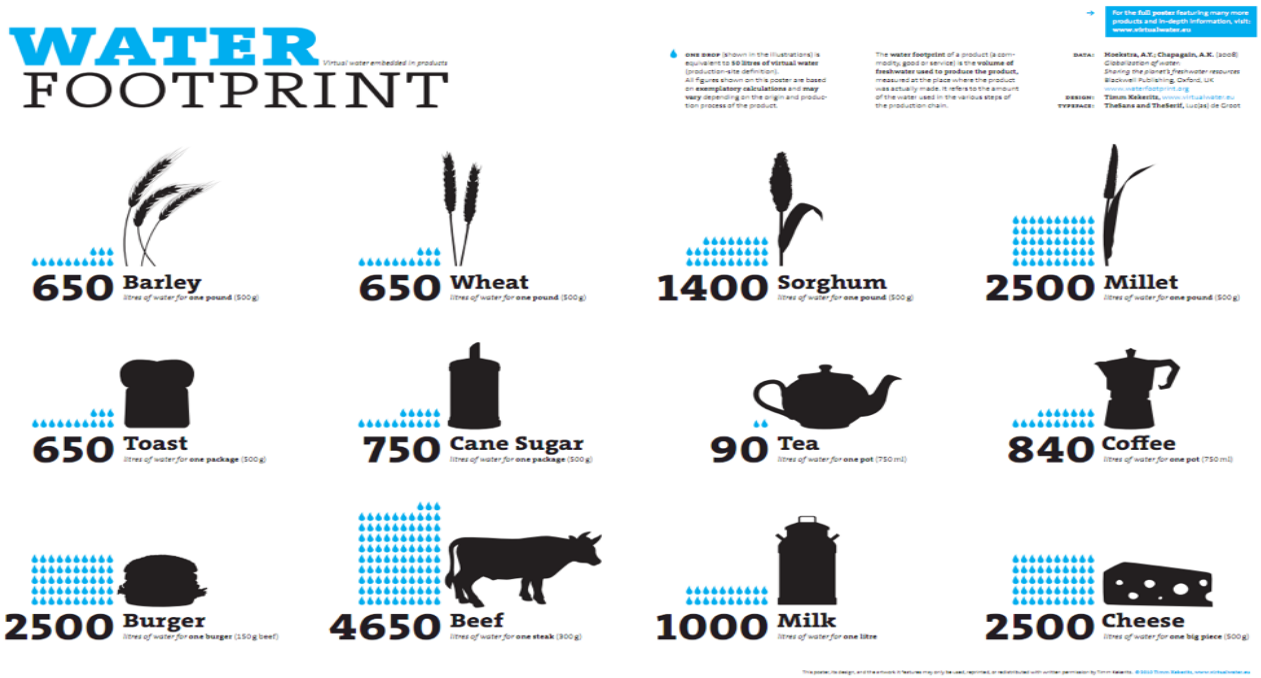


(a) (b)

Source: https://www.researchgate.net/publication/281711514/figure/fig6/AS:668581306060810
matrix-a-and-k-means-clusters-b-of-a-raw-soil-image.png.

Unsupervised Learning in Agronomy

- Identification of water consumption patterns: Grouping plants or areas with similar water consumption for irrigation optimization.



Source: https://www.jkgeography.com/global-patterns-and-trends-in-the-availability-and-consumption-of-water.html.

Supervised learning algorithms

Here are some of the most common supervised learning methods:

- Linear Regression
- Logistic Regression
- **Decision Trees, Random Forests**
- Support Vector Machines (SVM),
- K-Nearest Neighbors (KNN)
- Naive Bayes

ESALQ

Unsupervised learning algorithms

The goal is to discover hidden patterns, structures, or relationships within the data.

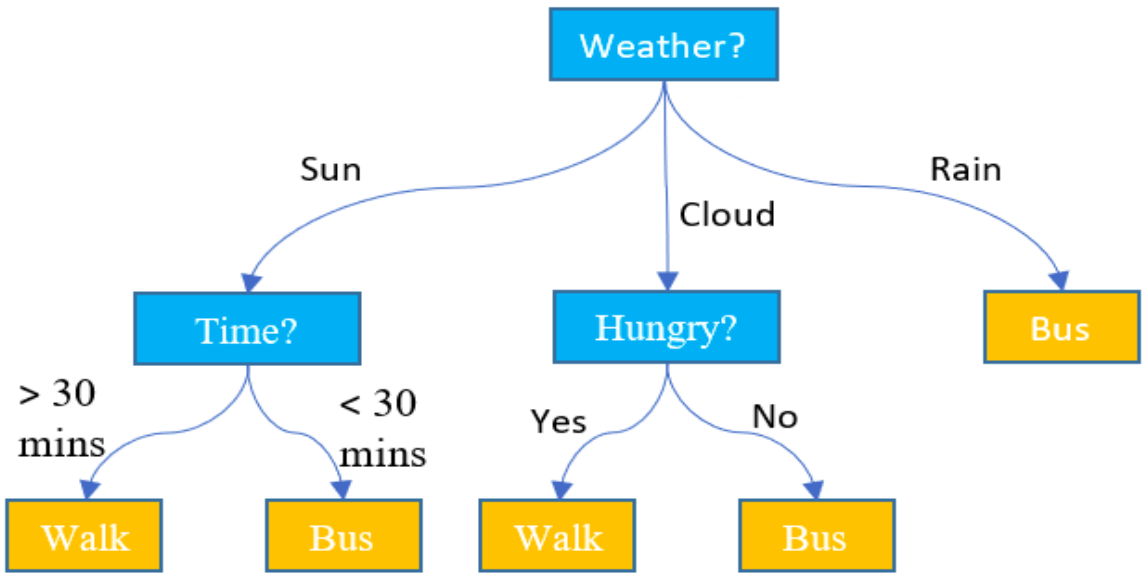
Here are some of the most common unsupervised learning methods:

- K-Means Clustering
- Hierarchical Clustering
- Principal Component Analysis (PCA)

ESALQ

Decision Tree [3]

- Seek to find the best split to subset the data,
- Two kinds of trees:
  - Classification Trees: factor responses and
  - Regression Trees: continuous response.
- Metrics, such as Gini impurity, information gain, or mean square error, can be used to evaluate the quality of the split.



Source: https://scikit-learn.org/stable/modules/tree.html.

Random Forest [4]

- It is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result;
- Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

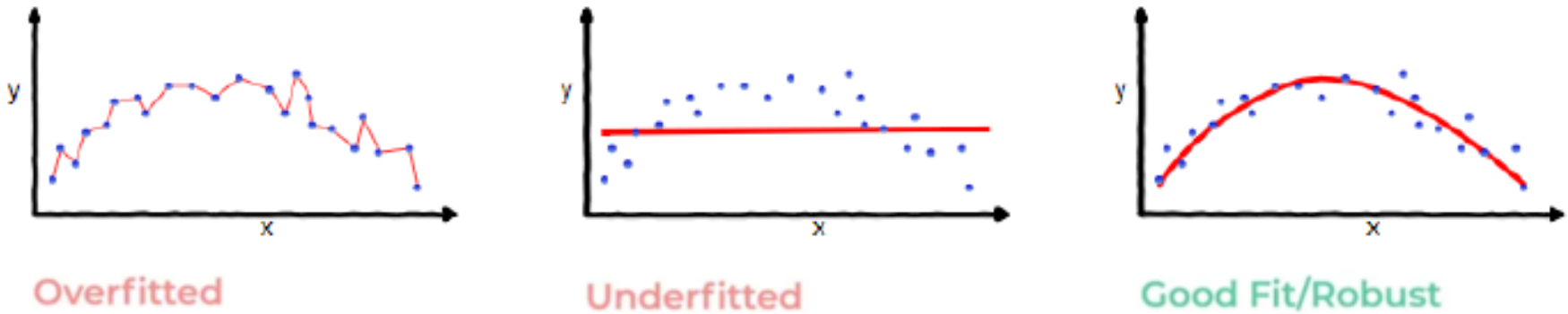
Source: https://onestopdataanalysis.com/python-random-forest-regression/.

[4] www.ibm.com/topics/random-forest

Model validation

In real life, the dataset is used to train a machine learning model/algorithm, and this is used to predict outputs that are unknown on new datasets;

To ensure predictive capability for new datasets, the model fit must be balanced (parsimonious):

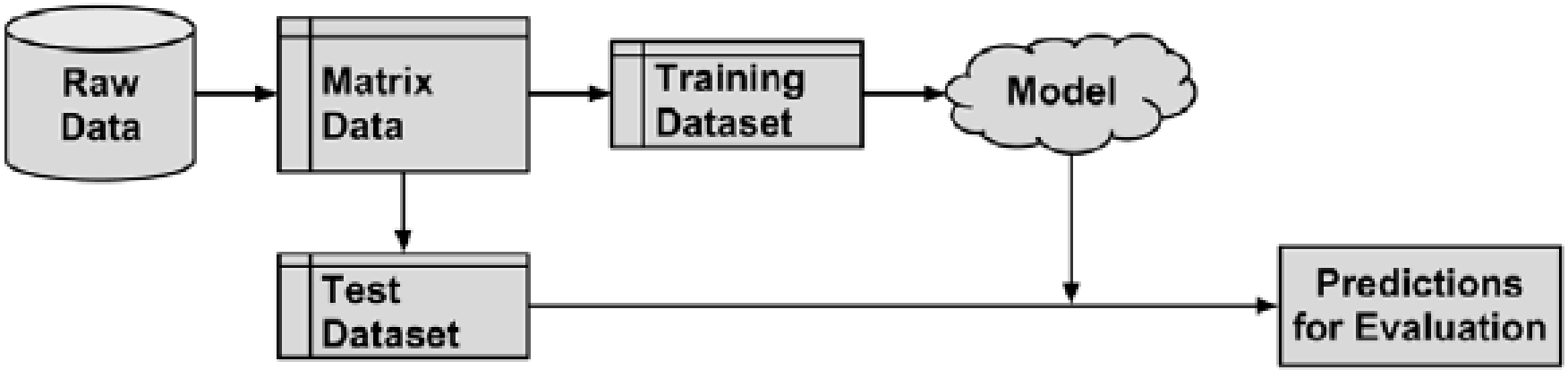


Source: Lecture Notes by Prof. Rafael A. Moral.

Model validation

To check the predictive capability of the model for new data, holdout method or k-fold cross-validation method can be used;

Holdout method: data is split into training and testing sets.



Source: Book "Machine Learning with R".

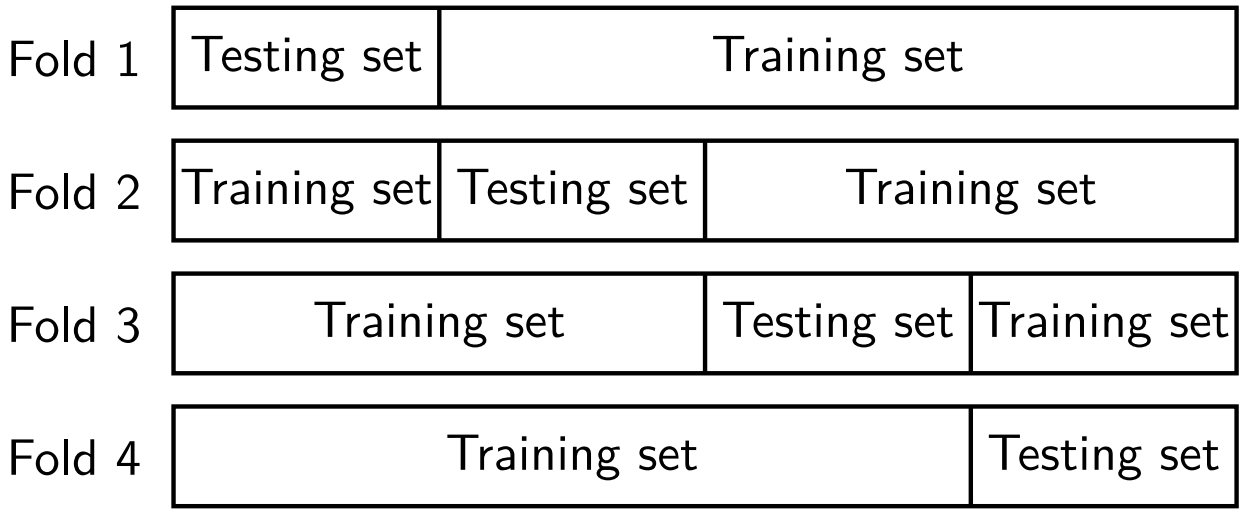
Model validation

k-fold cross-validation method:

The dataset is divided into k sets of approximately equal size;

There are k situations in which exactly what was done in the holdout procedure is done;

Each observation will be in the test set once, making the evaluation of the model have greater coverage.



Performance of classification algorithms

Confusion Matrix (Error matrix) - supervised learning
Matching Matrix - unsupervised learning
https://en.wikipedia.org/wiki/Confusion_matrix

- For the analysis of learning methods in classification problems, the confusion matrix is widely used:

Confusion matrix comparing predicted values and observed values.

		Prediction outcome	
		Positive	Negative
Actual value	Positive (1)	**True positive (TP)**	**False negative (FN)**
	Negative (0)	**False positive (FP)**	**True negative (TN)**

Source: own authors.

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Performance of classification algorithms

To define the indices that describe the degree of reliability of an algorithm, it is necessary to work with the following events:

- True positive (TP): cases where the model correctly classified an instance as positive.
- False positive (FP): cases where the model incorrectly classified an instance as positive.
- True negative (TN): cases where the model correctly classified an instance as negative.
- False negative (FN): cases where the model incorrectly classified an instance as negative.

ESALQ

Performance of classification algorithms - Reliability indices

- **Sensitivity (Recall):** measures how well the model can correctly identify instances of the positive class. It is defined as:

$$Sensitivity = \frac{TP}{TP + FN},$$

i.e., it is the proportion of positive results given that the actual value is positive (true positives).

Confusion matrix comparing predicted values and observed values.

		Prediction outcome	
		Positive	Negative
Actual value	Positive	**True positive (TP)**	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Source: own authors.

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

ESALQ

Performance of classification algorithms - Reliability indices

- **Specificity:** measures how well the model can correctly identify instances of the negative class. It is defined as:

$$Specificity = \frac{TN}{TN + FP},$$

i.e., it is the proportion of negative results given that the actual value is negative (true negatives).

Confusion matrix comparing predicted values and observed values.

		Prediction outcome	
		Positive	Negative
Actual value	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	**True negative (TN)**

Source: own authors.

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Performance of classification algorithms - Reliability indices

- **Precision (positive predictive value):** it is the number of true positive results divided by the number of all samples predicted to be positive, including those not identified correctly. It is defined as:

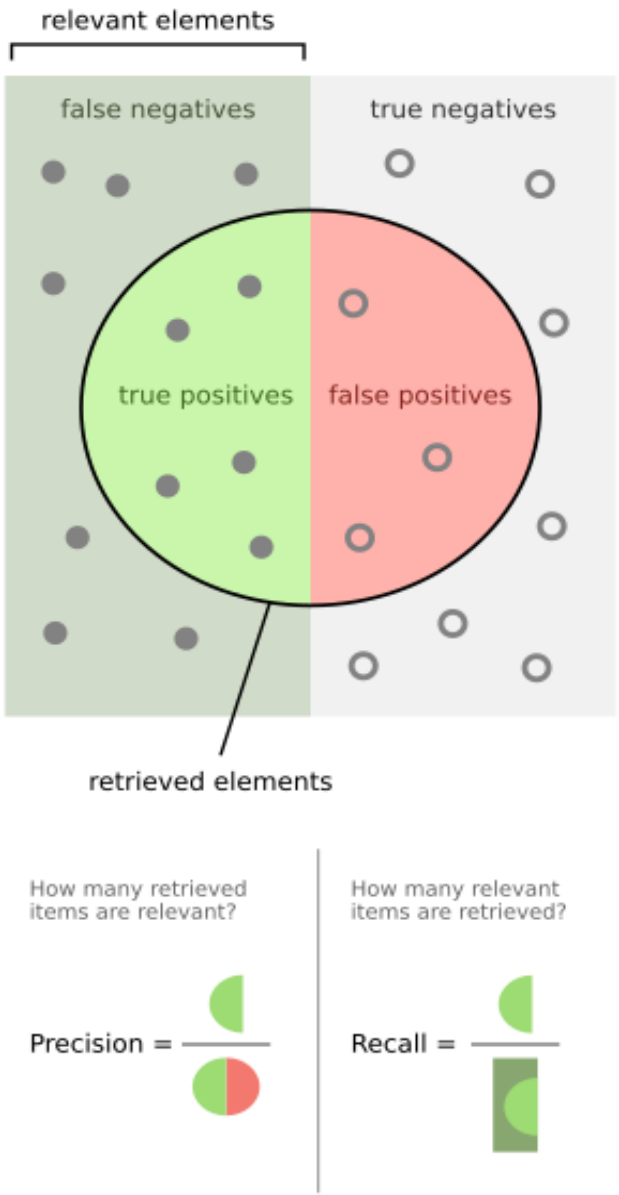
$$Precision = \frac{TP}{TP + FP},$$

Confusion matrix comparing predicted values and observed values.

		Prediction outcome	
		Positive	Negative
Actual value	Positive	**True positive (TP)**	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Source: own authors.

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN



Source: https://en.wikipedia.org/wiki/F-score

Performance of classification algorithms - Reliability indices

- **Accuracy:** proportion of correctly classified instances in relation to the total number of instances. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Confusion matrix comparing predicted values and observed values.

		Prediction outcome	
		Positive	Negative
Actual value	Positive	**True positive (TP)**	False negative (FN)
	Negative	False positive (FP)	**True negative (TN)**

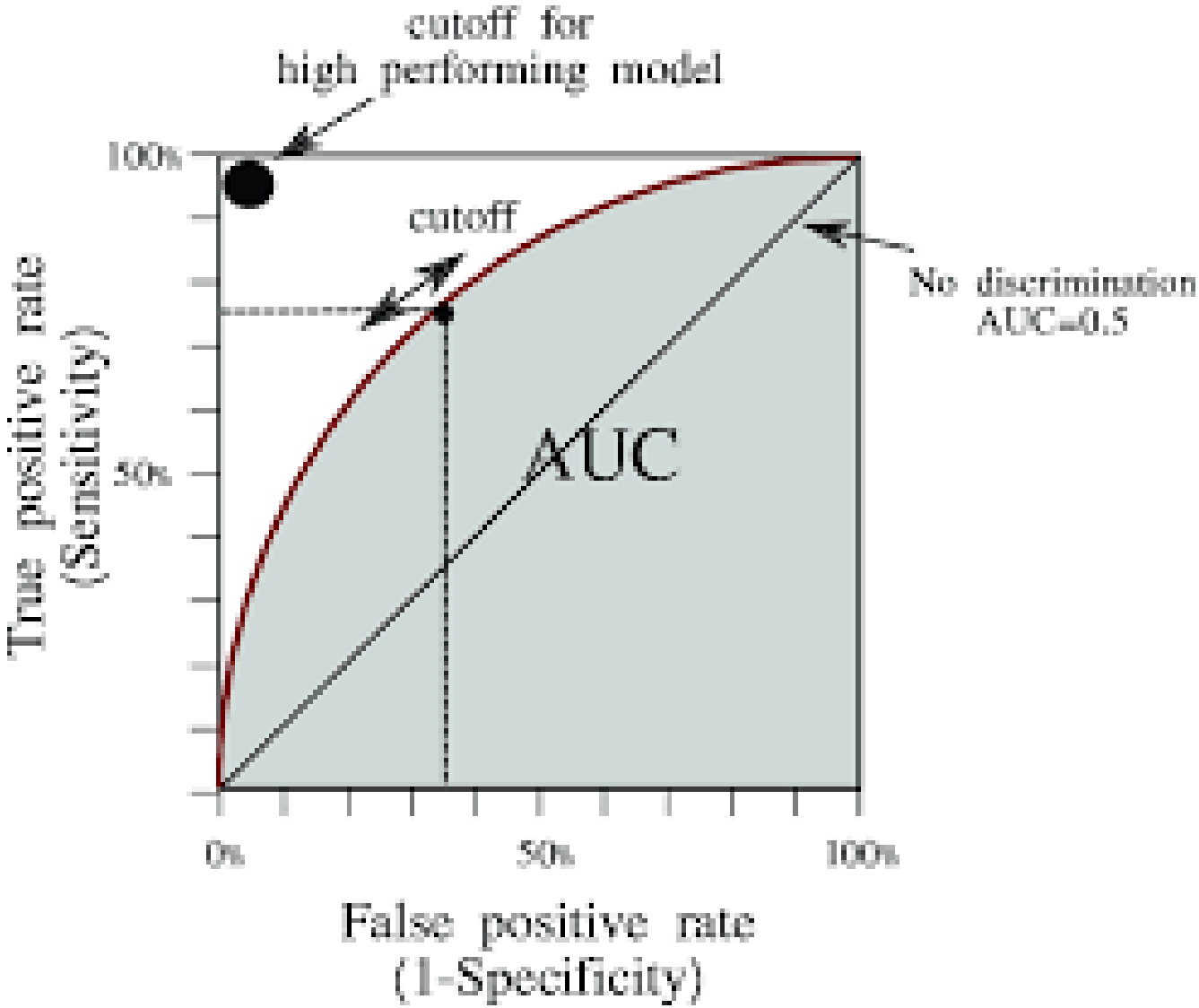
Source: own authors.

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Receiver Operating Characteristic (ROC) Curve

- It is the probability curve displaying the classifier's performance, showing how true positive rate (sensitivity) behaves under changing false positive rate (inverse specificity) .
- Inverse specificity is given by (1 - specificity).
- Area Under the ROC Curve (AUC): represents how well the algorithm is able to separate the different classes.
- The higher the AUC, the better the model is able to predict the classes.
- A well- performing algorithm must have an AUC value close to 1.

ESALQ


cutoff for
high performing model
100%
cutoff
No discrimination
AUC=0.5
True positive rate
(Sensitivity)
50%
AUC
0%
50%
100%
False positive rate
(1-Specificity)

Application

Source: https://globalplantcouncil.org/.

- Sugarcane cultivation occupies an important position in Brazilian agribusiness, covering about 2.4% of arable land;
- In the 2022/2023 harvest, SP surpassed the mark of 4.147 million hectares, reaching a production of around 312 million tons;
- Use technology (remote sensing) to enhance and optimize applications.

ESALQ

Application

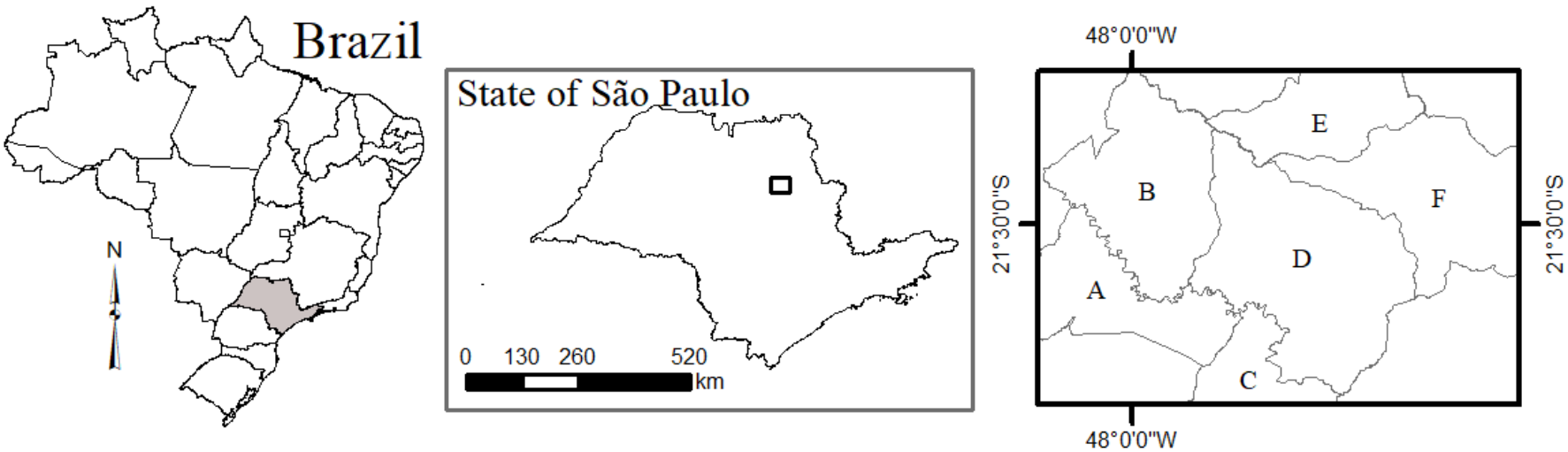
Purpose

The objective of this study was to develop and test classifiers for mapping sugarcane plantations, aiming to produce a rapid and accurate monitoring of the sugarcane area at a regional scale.

Application

Dataset

- A test area with 306 thousand hectares located in the north-central of the state of São Paulo was selected;
- It comprises part of the cities of A: Rincão; B: Guatapará; C: São Carlos; D: Luís Antônio; E: Cravinhos; F: São Simão (figure below);



Source: own authors.

Application

Dataset

- 6 features were measured over 17 months:

Description of vegetation indices.

Indices	Identification
Normalized difference vegetation index	NDVI
Enhanced vegetation index	EVI
Normalized difference water index	NDWI
Normalized difference moisture index	NDMI
Short-wave infrared 1	SWIR1
Short-wave infrared 2	SWIR2

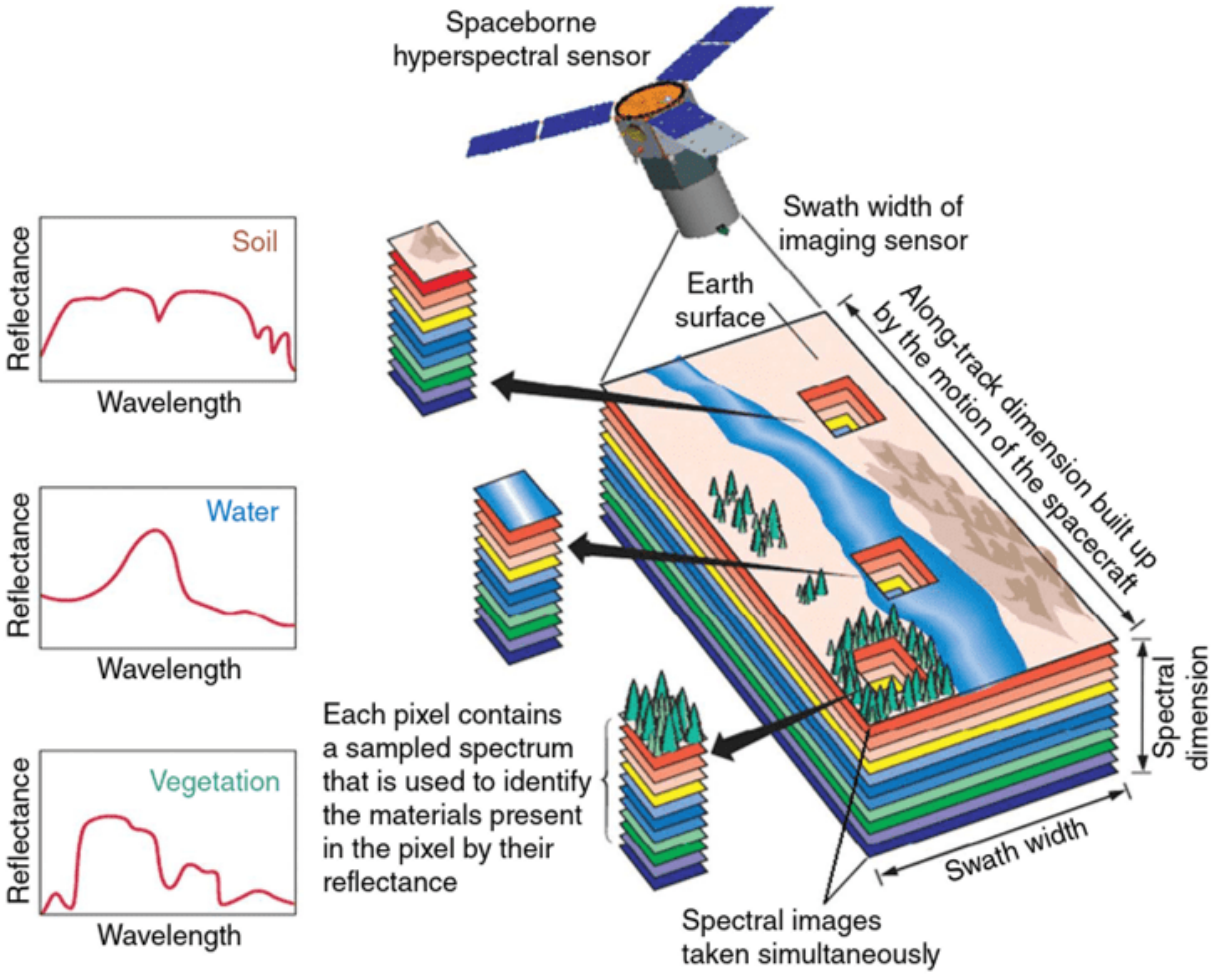
Source: own authors.

- Total: 102 features (6×17);
- The region was segmented into approximately 46,000 polygons.

ESALQ

Application

Dataset

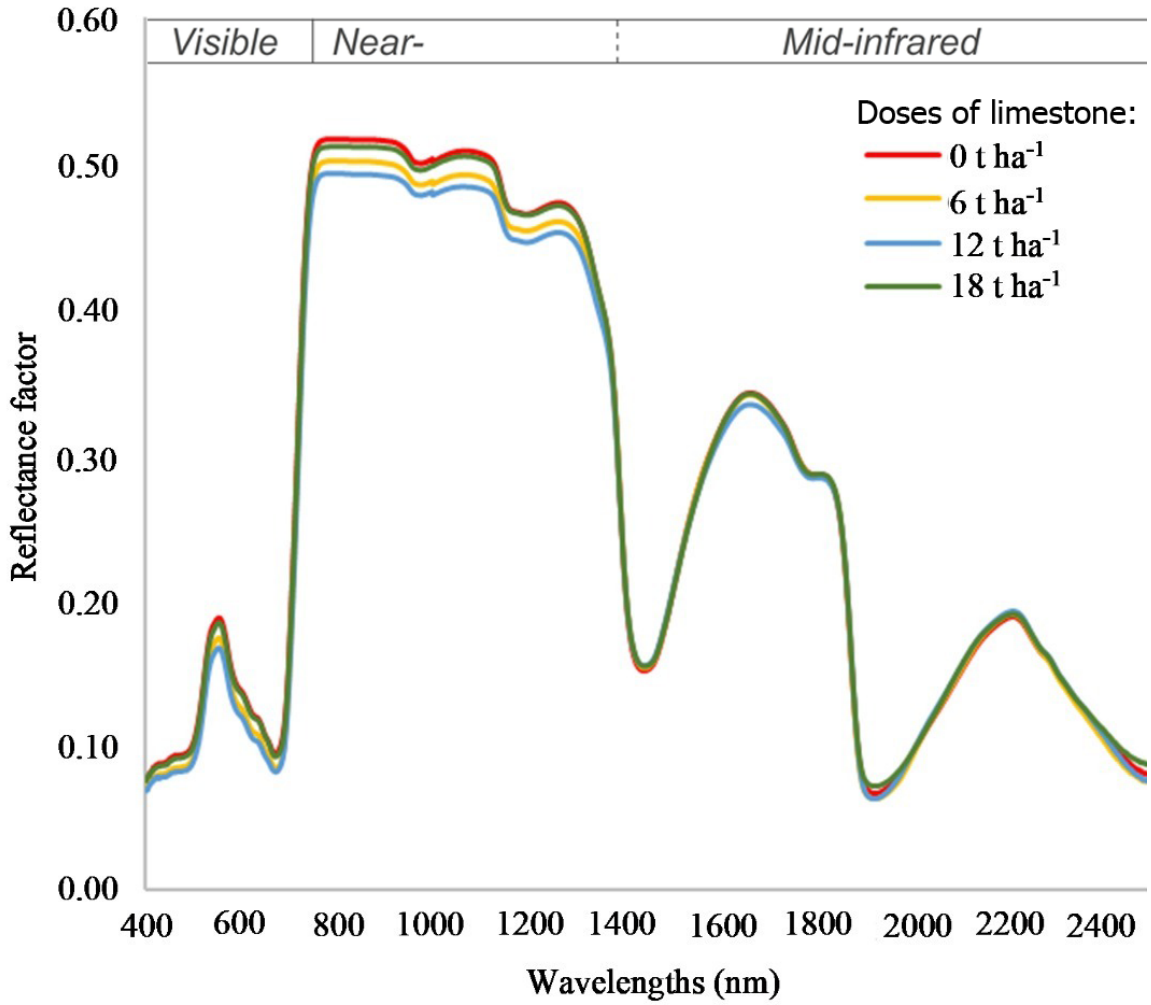


Source: https://ebrary.net/205101/engineering/introduction_hyperspectral_satellites

ESALQ

Application

Dataset



Source: "Detection of nutritional stress in sugarcane by VIS-NIR-SWIR reflectance spectroscopy".

ESALQ

Application

Outcome (Response variable): classification (sugarcane or not);

Machine learning methods

- Algorithms:
  - Logistic regression;
  - Decision tree;
  - Randon forest.
- Model validation: Holdout Method was used with 80% of the sample for training and 20% for testing.

ESALQ

Application

Softwares and packages

- QGIS;
- R (caret, ggplot2, randomForest, rpart).

Application

Results

To evaluate the models, we compared their accuracy, sensitivity, specificity and their confusion matrices.

Confusion matrices comparing predicted values and observed values.

		Prediction outcome					
		Logistic Regression		Decision tree		Random forest	
		1	0	1	0	1	0
Actual value	1	4341	1476	4754	1063	5265	552
	0	1310	2013	1469	4754	1412	1911

Source: own authors.

Application

Results

Accuracy, sensitivity, and specificity of the 3 models considered in the study.

Modelo	Accuracy	Sensitivity	Specificity
Logistic regression	69.52%	74.63%	60.58%
Decision tree	72.30%	81.73%	55.79%
Random forest	78.51%	90.51%	57.51%

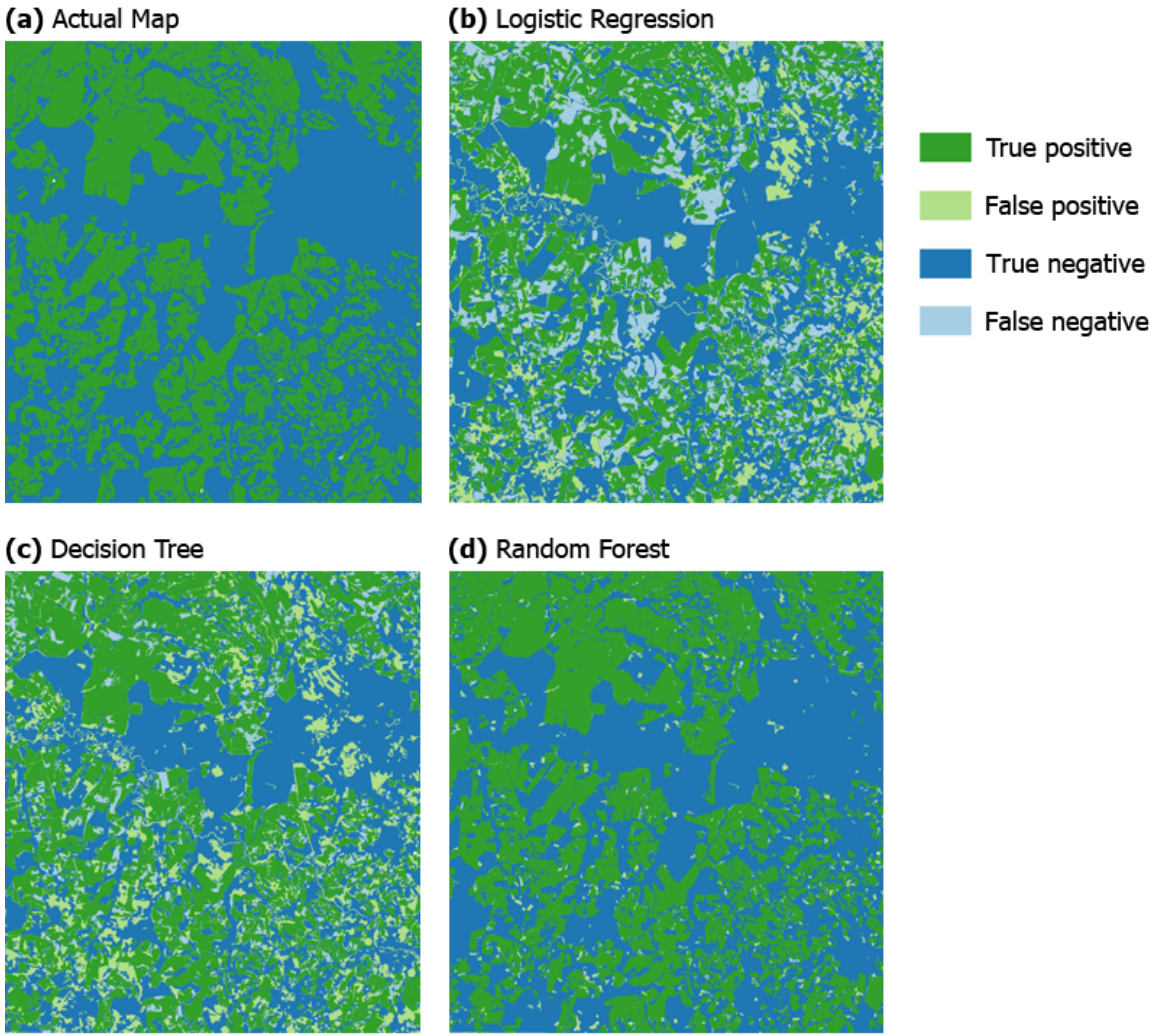
Source: own authors.

ESALQ

Application

Results

Maps of the study region with the comparison between the predictions of each method.



Source: own authors.

Application

Conclusion

- We reached an accuracy above 69%, considering the appropriate methods for classifying areas with sugarcane;
- A limitation found is the high number of false positive predictions in the classification of polygons;
- To solve this problem, we suggest, as a future work, the implementation of a post-classification method that considers the classification of neighboring polygons to confirm the prediction of each segment.

ESALQ

Obrigado


рахмат
danke
謝謝
ngiyabonga
spas
Баярлалаа
mersi
barka
welalin
tack
teşekkür ederim
спасибо
faafetai lava
vinaka
kia ora
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
спасибі
blagodaram
kiitos
dankie
thank you
gracias
хвала
nandri
nami
dhanyavad
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràcie
hvala
mauruuru
köszönöm
dankon
aciū
djiere dieuf
tau
mochchakkeram
murakoze
enkosi
akun
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dėkuji
taiku
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
obrigado
mèsi
sagolun
ありがとう
tanemirt
grazie
didi madloba
najis tuke
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercè
мерси
kam sah hamnida
rahmat
xièxie
ευχαριστώ
merci
তোমাকে ধন্যবাদ
감사합니다