

PGCAfrica AI Workshop

Workshop Theme: Artificial Intelligence in Human Genetics

7th February 2025

Plenary 1: General Introduction to Artificial Intelligence

Itunuoluwa Isewon (Ph.D)
Department of Computer & Information Sciences
Covenant University, Ota, Nigeria

Course Manual:
bit.ly/PGCAfricaAI_2025

An Introduction to AI and its Applications in Human Genetics Research

Introduction to AI

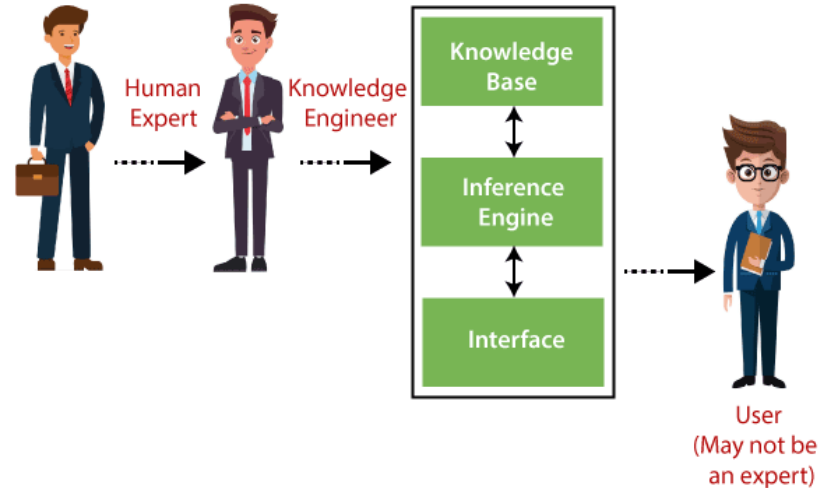
- Definition of AI: The simulation of human intelligence in machines.
- AI vs. Machine Learning: AI is a broad field, while ML is a subset.
- Major branches of AI: Expert systems, robotics, computer vision, natural language processing (NLP), and more.



<https://ats.net/en/focus-ai-artificial-intelligence-and-the-new-doomsday-machine/>

Knowledge-Based AI (Expert Systems)

- Rule-based reasoning and decision-making.
- Example: Medical diagnosis expert systems.
- Strengths: Explainability, domain knowledge integration



<https://www.jaroeeducation.com/blog/what-are-expert-systems-in/>

Robotics and AI

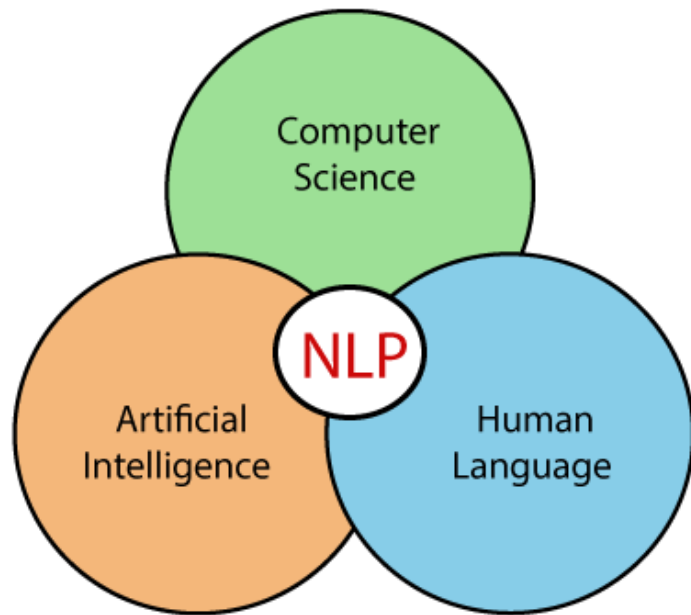
- AI-driven automation in robotics.
- Key applications: Manufacturing, autonomous vehicles, healthcare.
- AI techniques: Sensor fusion, reinforcement learning, motion planning



<https://www.coop.co.za/navigating-the-future-the-convergence-of-ai-and-robotics/>

Natural Language Processing (NLP) & AI

- How AI enables understanding and generating human language.
- Applications: Chatbots, translation, sentiment analysis.
- Beyond ML: Symbolic AI and rule-based NLP.



<https://speakai.co/what-is-natural-language-processing-the-definitive-guide/>

Generative AI & Creativity

- AI for art, music, and content generation. Example: GPT for text, DALL·E for images. Ethical considerations: Bias, deepfakes, originality.

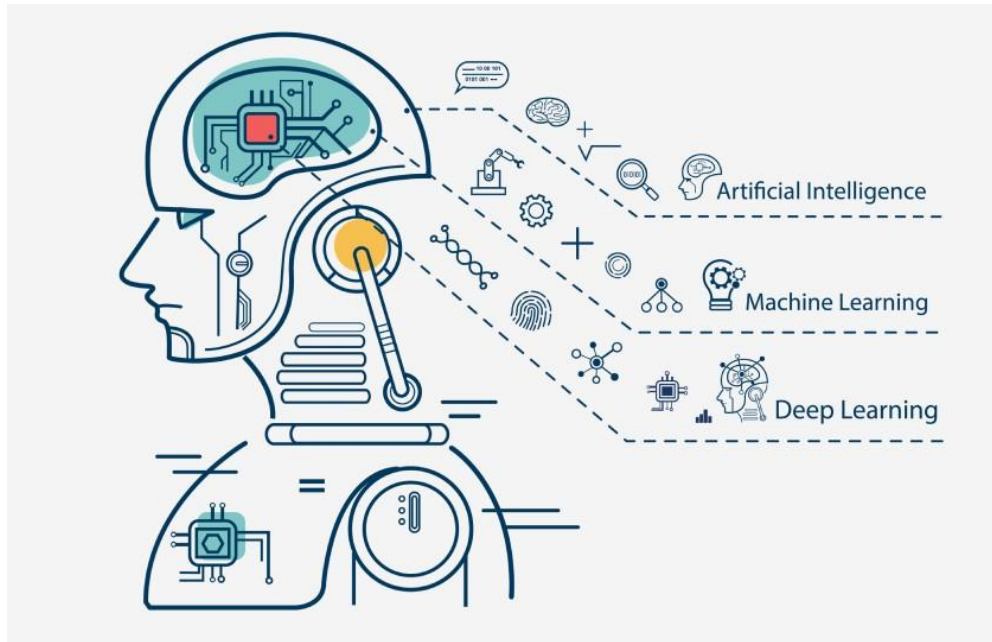


<https://news.uga.edu/generative-ai-and-creativity/>

What is Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) that enables computers to learn and improve from experience without being explicitly programmed.

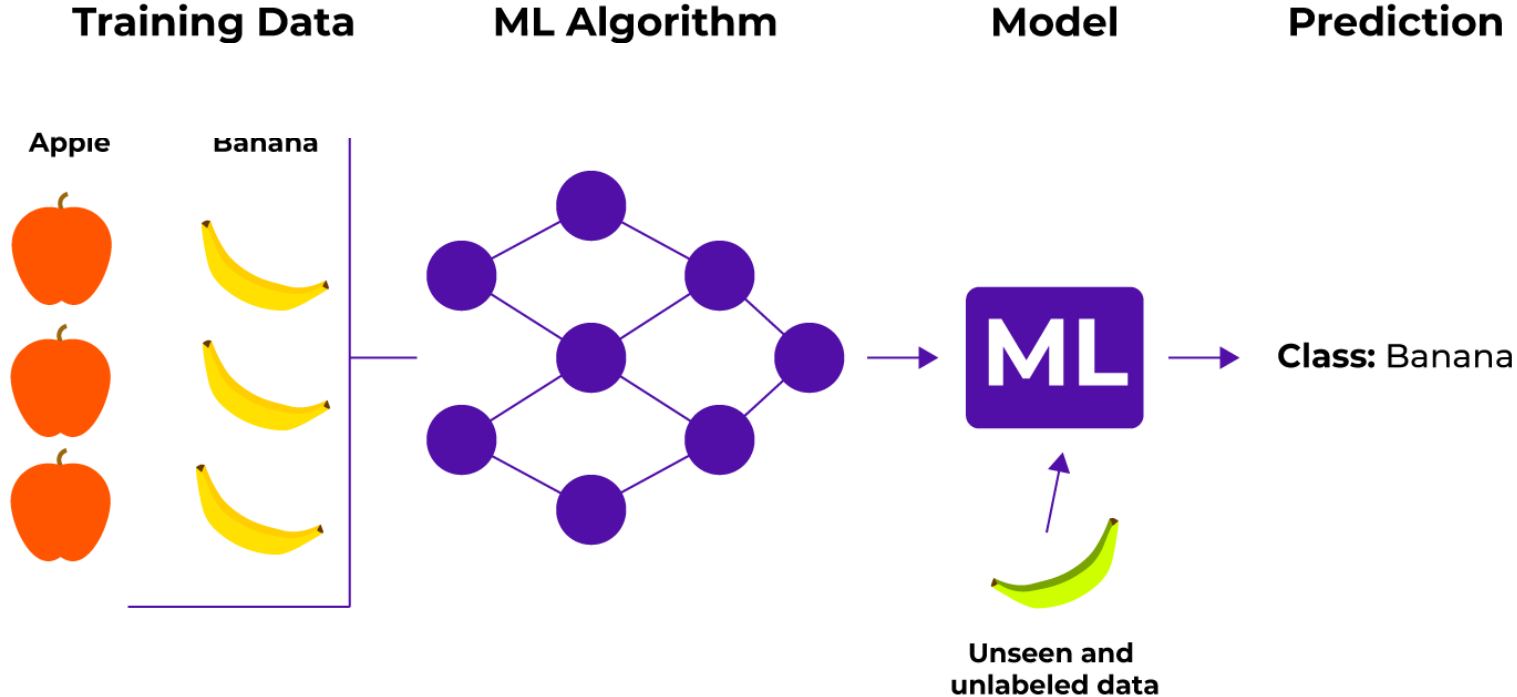
The primary goal of machine learning is to develop algorithms that can automatically learn patterns and relationships from data to make predictions for new, previously unknown data .



<https://www.insiris.com/blog/machine-learning-in-the-real-world>

Categories of Machine learning

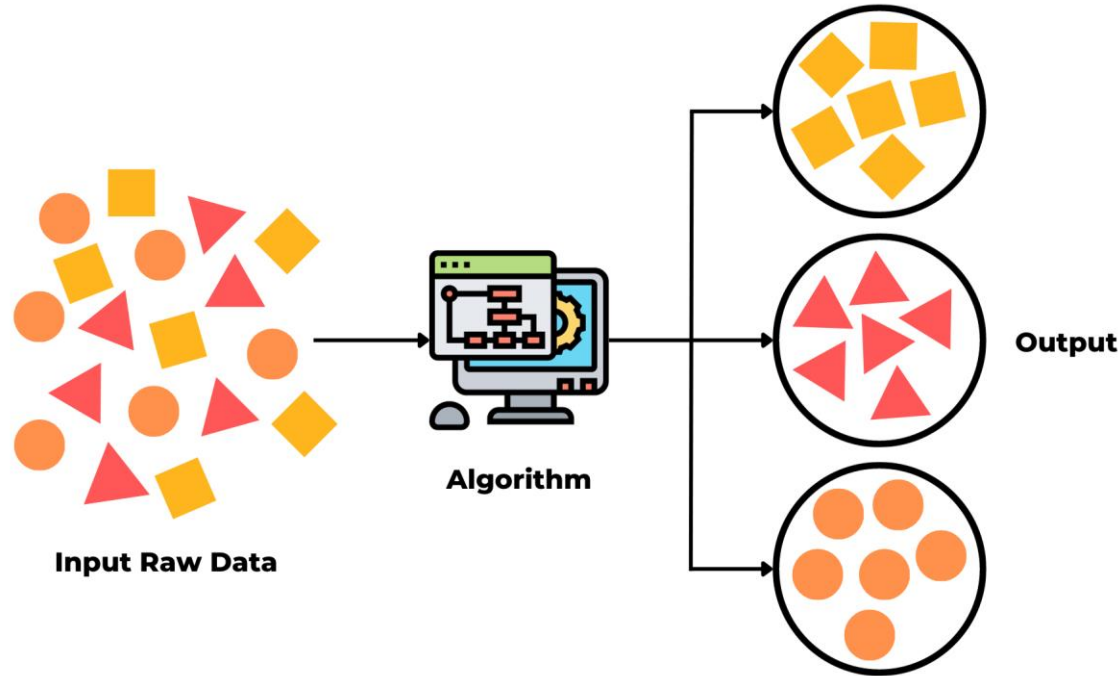
Supervised Learning:



<https://medium.com/@george.felobes/overview-of-supervised-learning-the-art-of-function-approximation-5a508bbd66de>

Categories of Machine learning

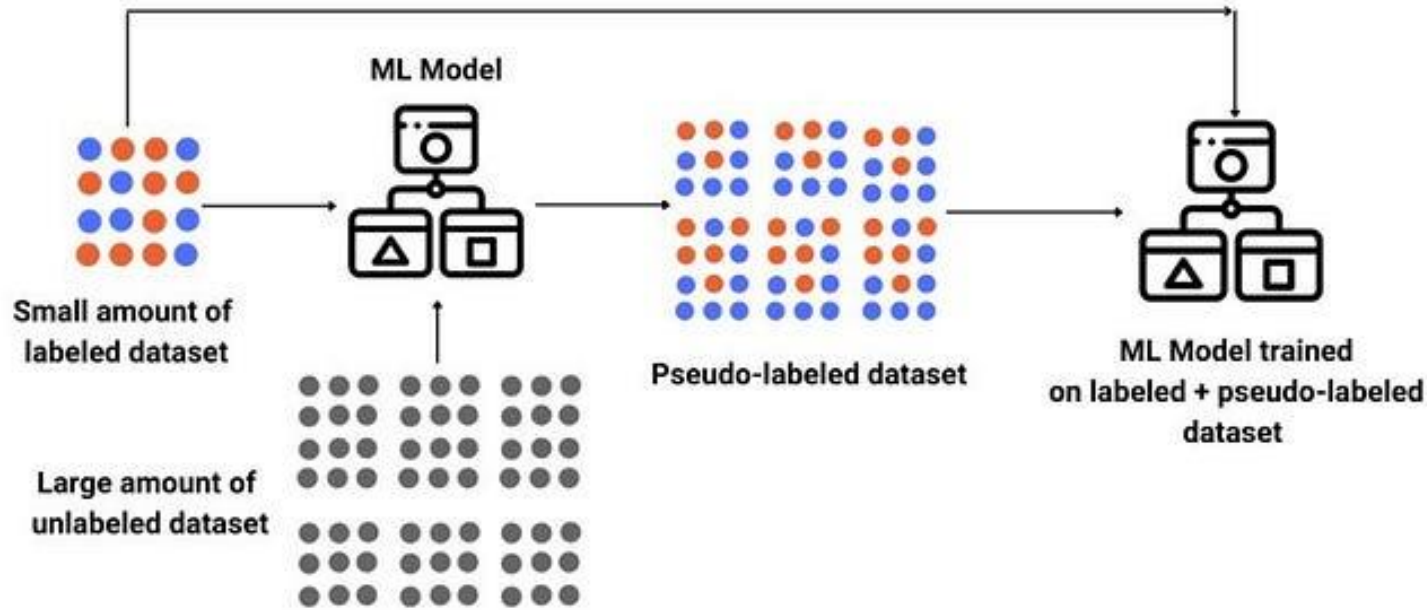
Unsupervised Learning:



<https://dev.to/samarpitnandanwar/unsupervised-learning-a-comprehensive-guide-2bn0>

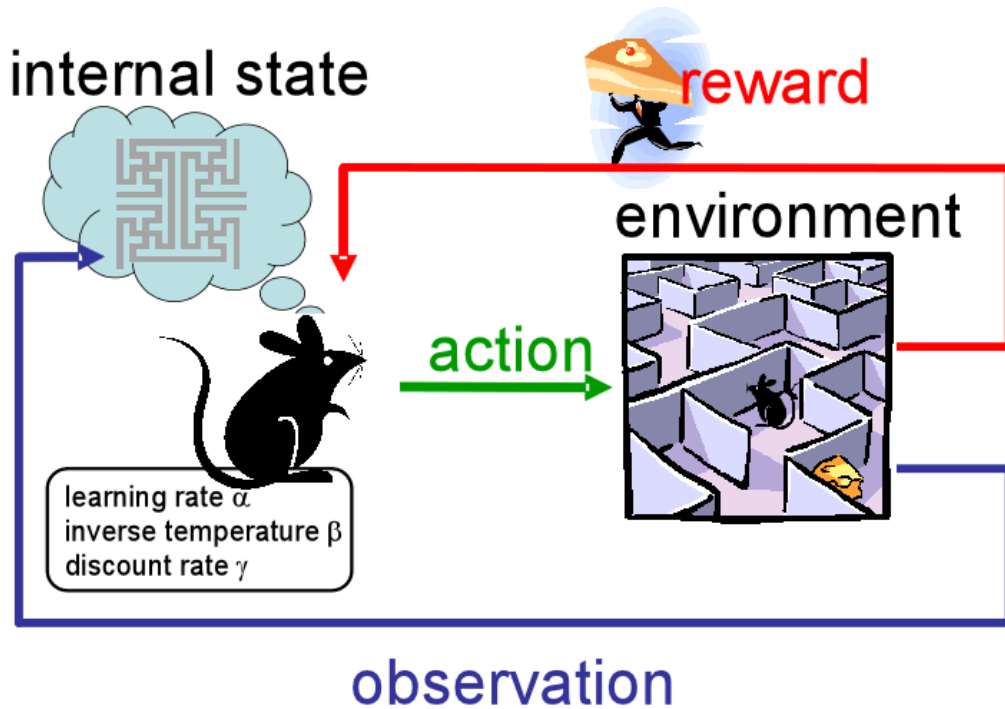
Categories of Machine learning

Semi-supervised learning :



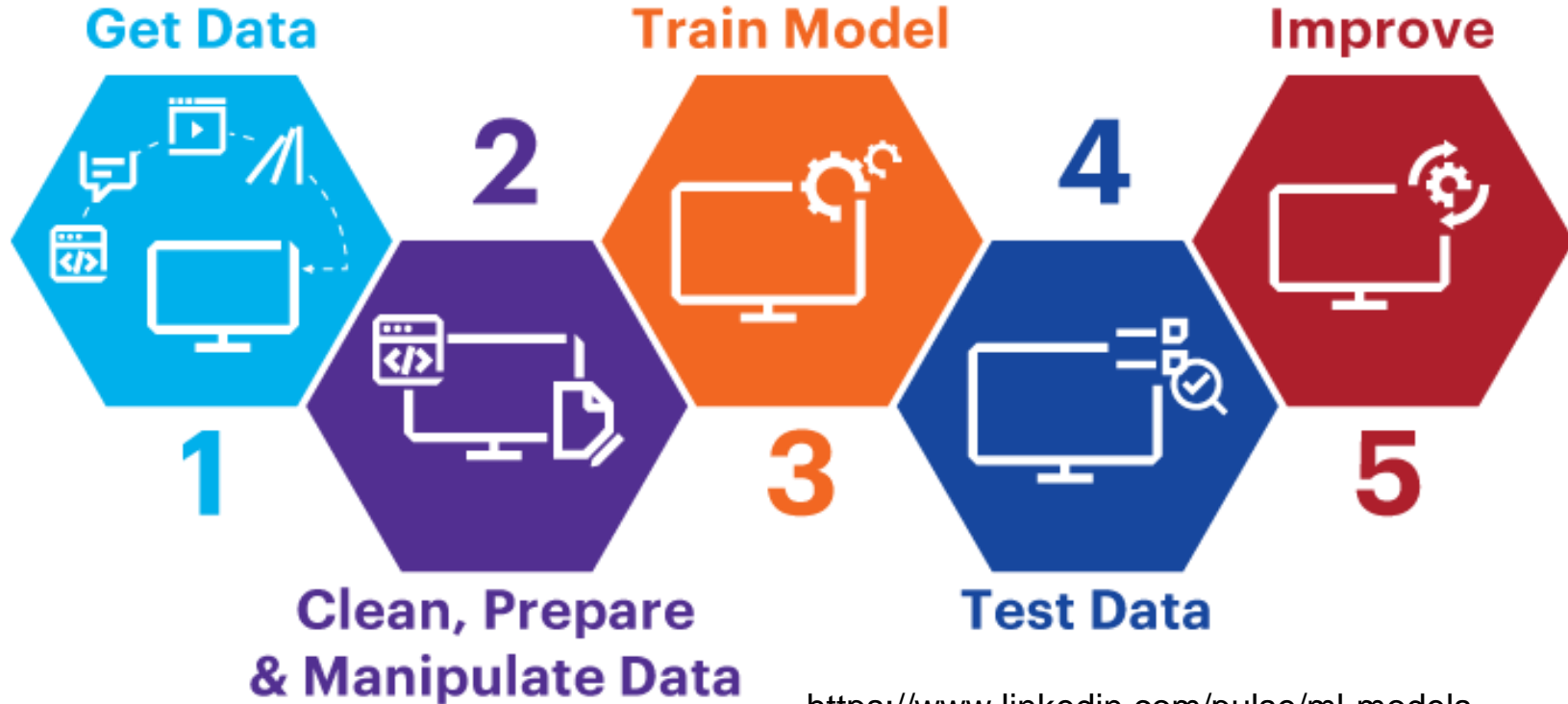
Categories of Machine learning

Reinforcement Learning:



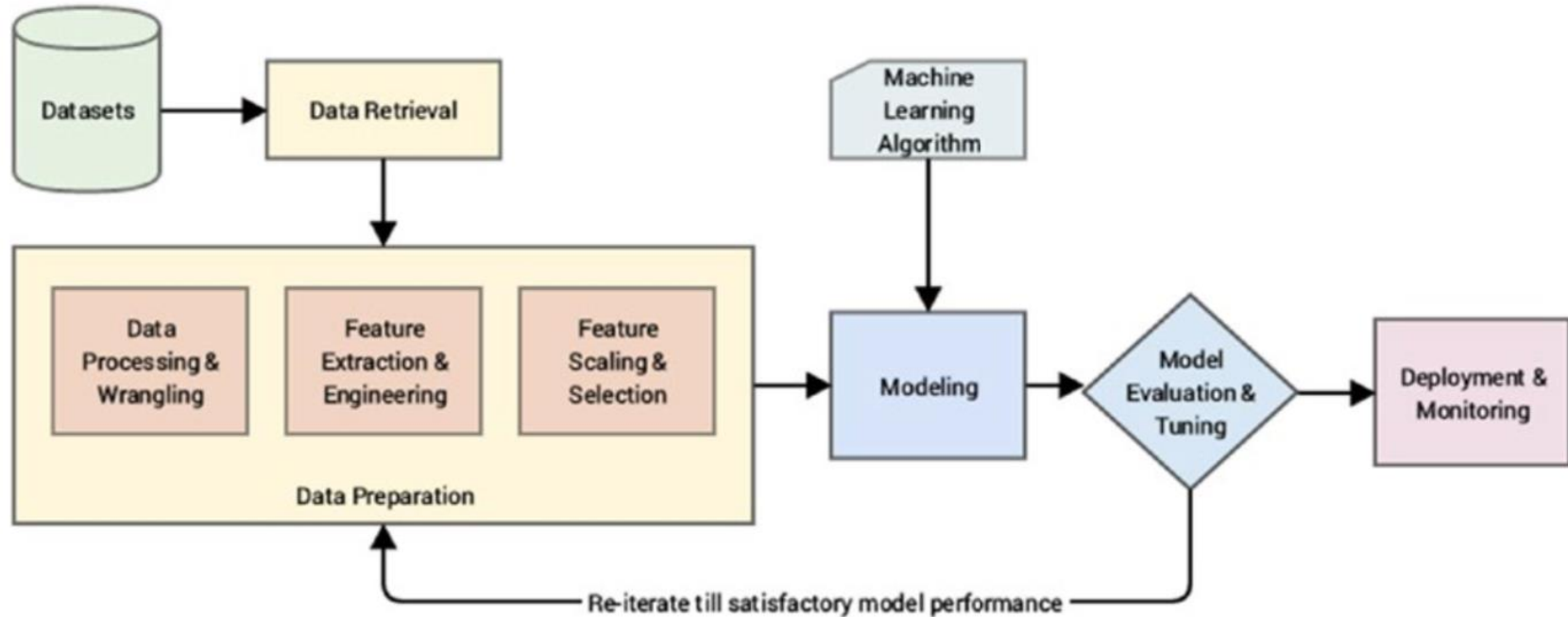
<https://towardsdatascience.com/introduction-to-q-learning-88d1c4f2b49c>

Process of Machine learning Experiments



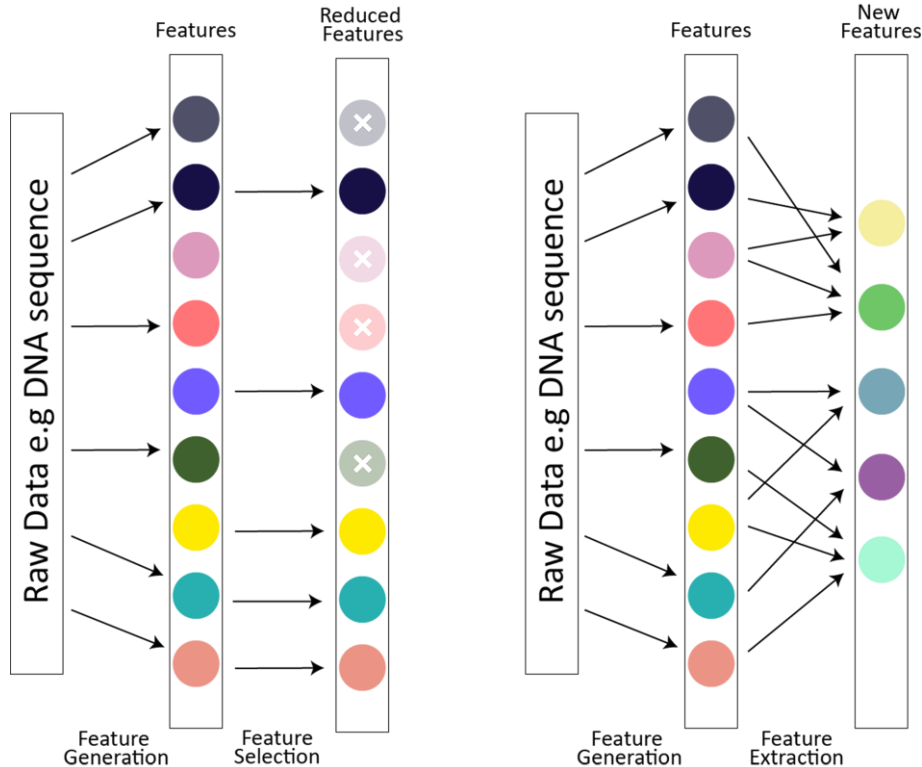
<https://www.linkedin.com/pulse/ml-models-darshika-srivastava-tuhkc/>

Process of Machine learning Experiments



<https://web2.qatar.cmu.edu/~gdicaro/15288/>

Feature Generation, Selection and extraction



Feature Selection

What is Feature Selection?

Choosing the most important features for a model.

Why is it important?

- Reduces overfitting
- Improves model interpretability
- Enhances computational efficiency

Types of Feature Selection:

- Filter Methods (e.g., correlation, mutual information)
- Wrapper Methods (e.g., Recursive Feature Elimination)
- Embedded Methods (e.g., LASSO, ElasticNet)

All Features



Feature Selection



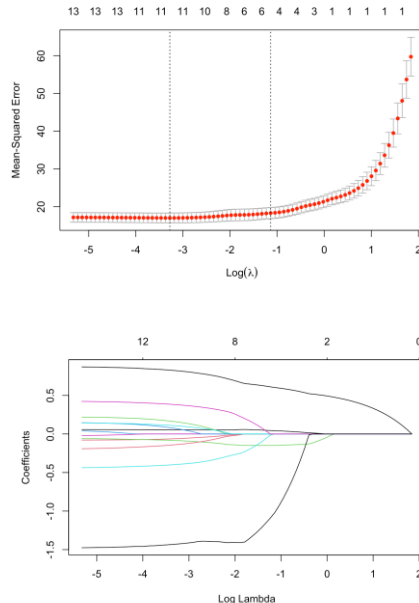
Final Features



<https://vitalflux.com/machine-learning-feature-selection-feature-extraction/>

LASSO (Least Absolute Shrinkage and Selection Operator)

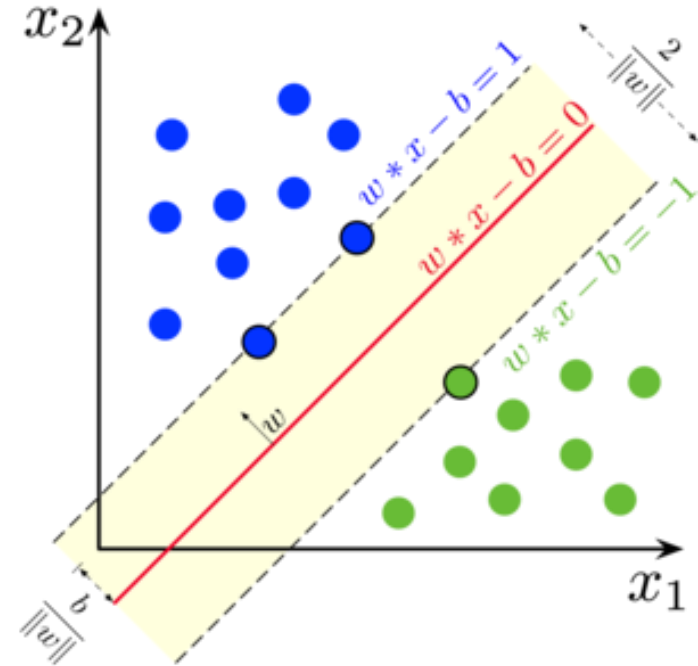
- **How it works:** Adds **L1 penalty** to the loss function, forcing some coefficients to become zero.
- Helps in automatic feature selection.



https://bookdown.org/tpinto_home/Rregularisation/lasso-regression.html

SVM

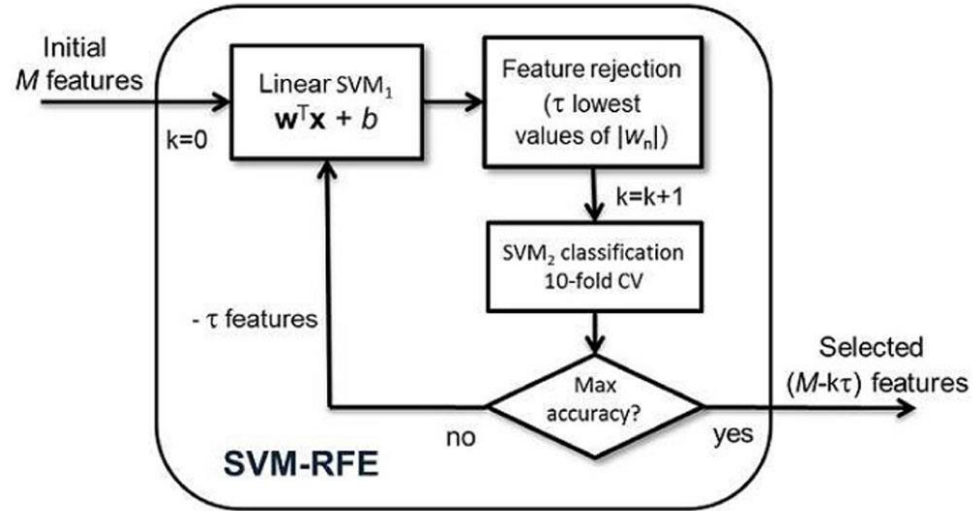
SVM (Support Vector Machine) is a supervised learning algorithm for classification and regression. It finds an optimal hyperplane, maximizes margins, and uses kernels to handle non-linearly separable data. Useful in bioinformatics, cancer research, and text classification.



SVM-RFE (Support Vector Machine Recursive Feature Elimination)

- How it works:** Trains an SVM model, ranks features by importance, and removes the least important iteratively.

- Advantages:** Works well for high-dimensional data (e.g., genetics, imaging)

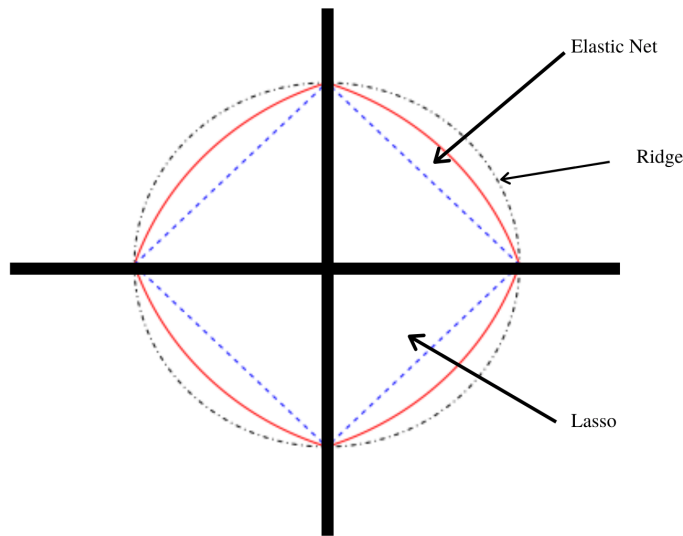


<https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2014.00020/full>

ElasticNet (Combination of LASSO & Ridge Regression)

- **How it works:** Uses both L1 (LASSO) and L2 (Ridge) penalties balancing feature selection and regularization.

- **When to use?** When features are correlated (LASSO may randomly drop one).





<https://corporatefinanceinstitute.com/resources/data-science/elastic-net/>

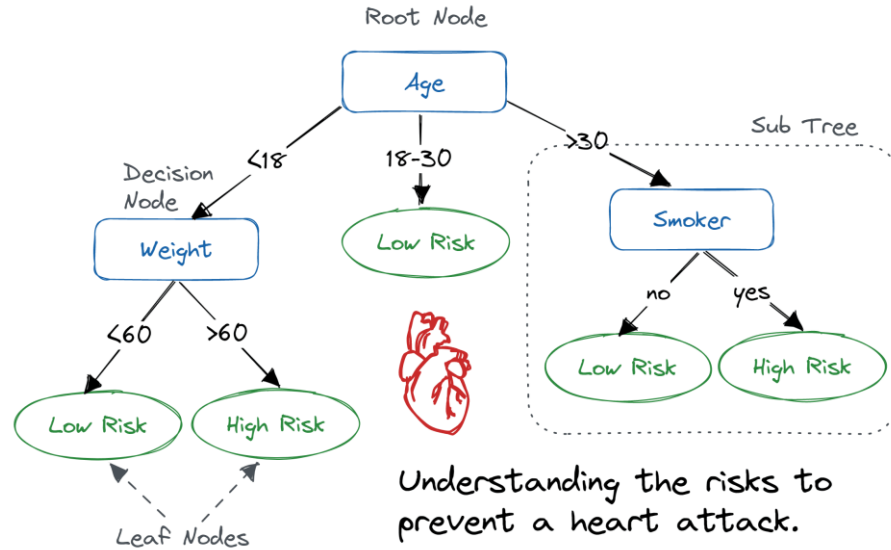
What is a Classification Task

Binary , Multiclass, Multi-label Classification Tasks

Binary

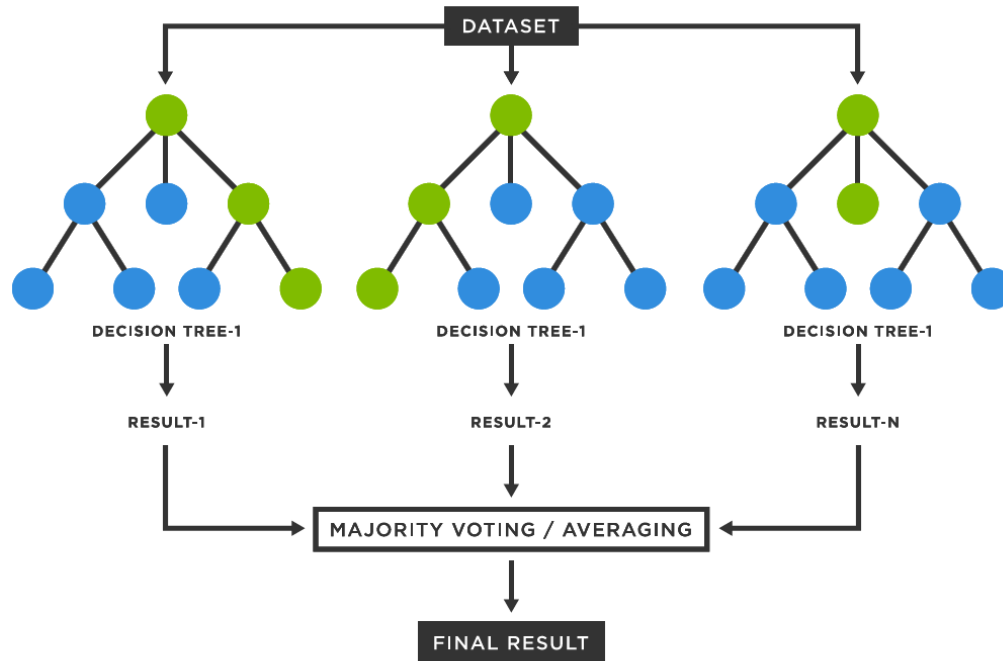
Binary		Multi-Class	Multi-Label
C = 2	C = 3	<p>Samples</p>  <p>Labels</p> <p>[100] [010] [001]</p>	<p>Samples</p>  <p>Labels</p> <p>[110] [011] [111]</p>

Machine learning algorithms for classification tasks – Shallow models



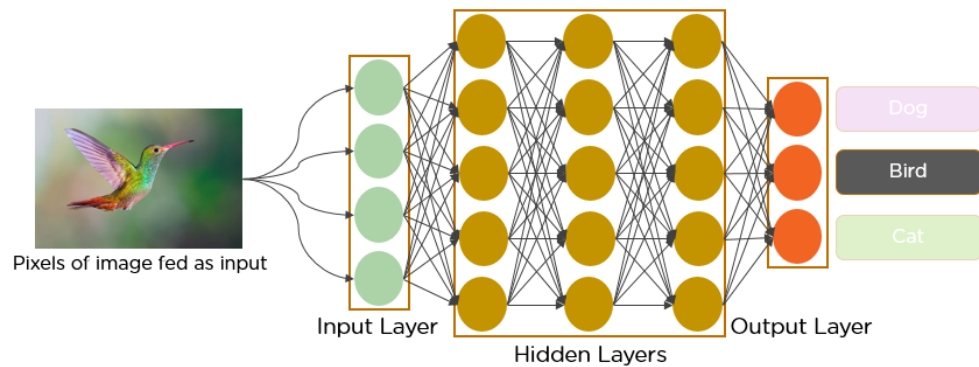
- Logistic Regression
- Decision Trees
- Support Vector Machines (SVM)
- k-Nearest Neighbors (k-NN)

Machine learning algorithms for classification tasks – Ensembles



Random Forest
Gradient Boosting Machines
AdaBoost
Bagging Classifier

Machine learning algorithms for classification tasks – Deep Learning



<https://medium.com/@soumyaranjanmishra.in/how-does-cnn-recognize-images-4ebcf9a3d9d1>

ChatGPT



Convolutional Neural Networks (CNN)
Recurrent Neural Networks (RNN)
Transformer models (e.g., BERT, GPT)
Long Short-Term Memory (LSTM)
Autoencoders

How do I rate my ML algorithm

Performance metrics provide insights into how well the model is performing and can help in comparing different models. Examples include

Accuracy: Percentage of correct predictions out of total predictions. Measures overall correctness of the model.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

How do I rate my ML algorithm

Precision: Measures how many predicted positives are actually positive

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall (Sensitivity): Measures how many actual positives are predicted correctly.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

How do I rate my ML algorithm

Specificity: Measures how many actual negatives are predicted correctly.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

F1 Score: A robust performance metric that balances between precision and recall

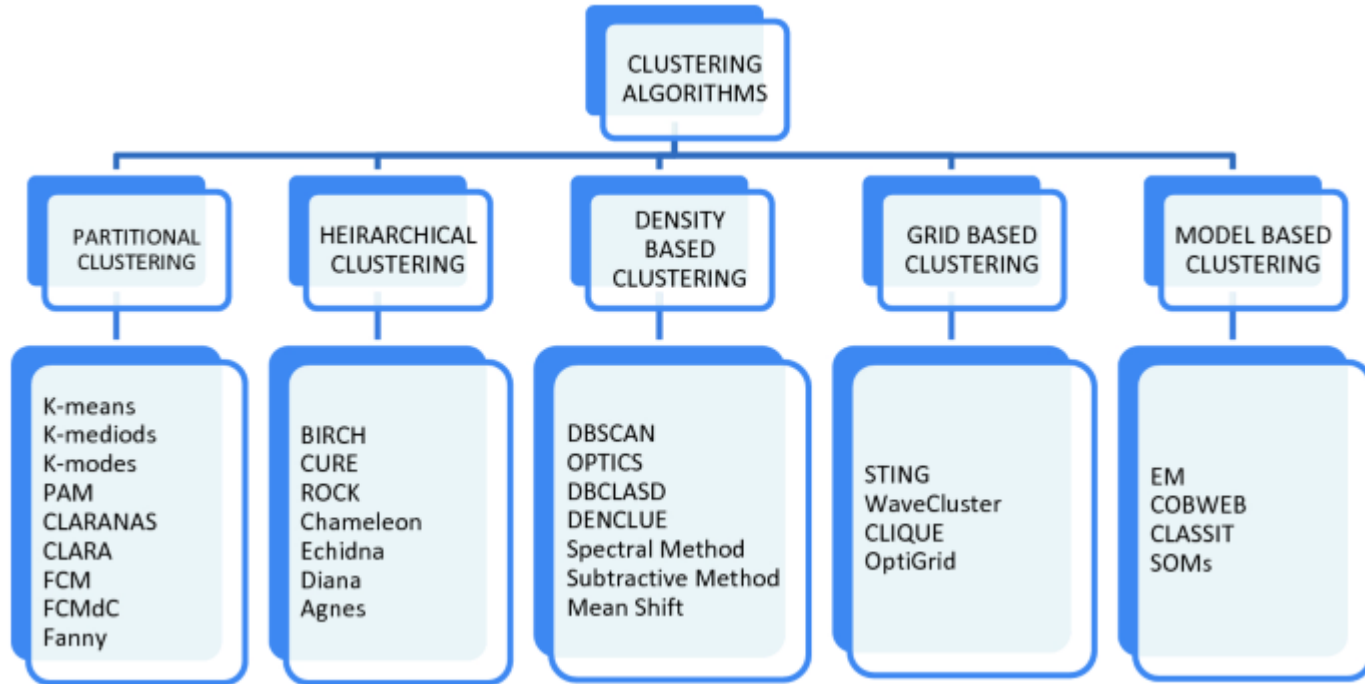
$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

How do I rate my ML algorithm

	precision	recall	f1-score	support
0	0.77	0.86	0.81	37584
1	0.84	0.75	0.79	37577
accuracy			0.80	75161
macro avg	0.81	0.80	0.80	75161
weighted avg	0.81	0.80	0.80	75161

What is Clustering

Categories of Clustering algorithms



https://link.springer.com/chapter/10.1007/978-981-13-7403-6_9

How do I rate my clustering algorithm

Internal Evaluation Metrics

Internal evaluation metrics assess the quality of clustering without reference to external labels or ground truth.

1. **Silhouette Score:** measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).

$$Silhouette(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

$a(i)$: average distance from point i to other points within the same cluster.

$b(i)$: Average distance from point i to points in the nearest neighboring cluster

How do I rate my clustering algorithm

Overall Silhouette Score :

$$\text{Silhouette Score} = \frac{1}{n} \sum_{i=1}^n \text{Silhouette}(i)$$

where n is the total number of data points

How do I rate my clustering algorithm

2. **Davies-Bouldin Index:** measures the average similarity between each cluster and its most similar cluster, where a lower value indicates better clustering.

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right)$$

where:

k : Number of clusters.

c_i, c_j : Centroids of clusters i and j respectively

s_i, s_j : Average distance from points in cluster i and j to their respective centroids

$d(c_i, c_j)$: Distance between centroids c_i and c_j

How do I rate my clustering algorithm

External Evaluation Metrics

External evaluation metrics compare clustering results with ground truth labels (if available)

1. **Adjusted Rand Index:** measures the similarity between two clusterings, considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.

$$ARI = \frac{RI - \mathcal{E}[RI]}{\max(RI) - \mathcal{E}[RI]}$$

where:

RI : Rand index, which is the proportion of agreements between two clusterings.

$\mathcal{E}[RI]$: Expected Rand index for random clusterings.

How do I rate my clustering algorithm

2. **Mutual Information (MI):** measures the agreement between two clusterings, considering the entropy (uncertainty) of the true labels and predicted clusters.

$$MI(U, V) = \sum_{u \in U} \sum_{v \in V} p(u, v) \log \left(\frac{p(u, v)}{p(u)p(v)} \right)$$

where:

U, V : Sets of true labels and predicted clusters, respectively.

$p(u, v)$: Joint probability distribution of U and V

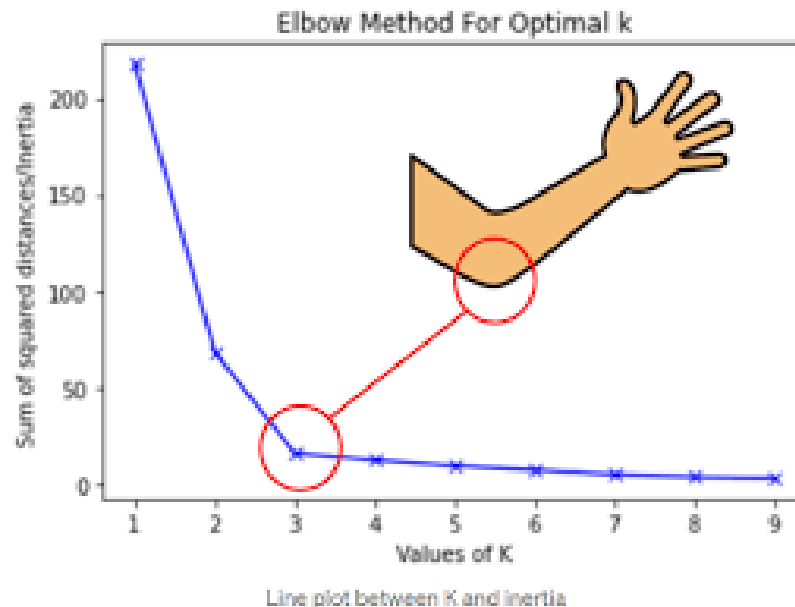
$p(u)p(v)$: Marginal probabilities probability distribution of U and V

Choosing Optimal 'K'

Choosing the optimal number of clusters K is a crucial step in clustering analysis, especially for algorithms like K-means where the number of clusters K needs to be specified in advance. Here are several methods commonly used to determine the optimal:

Choosing Optimal 'K'

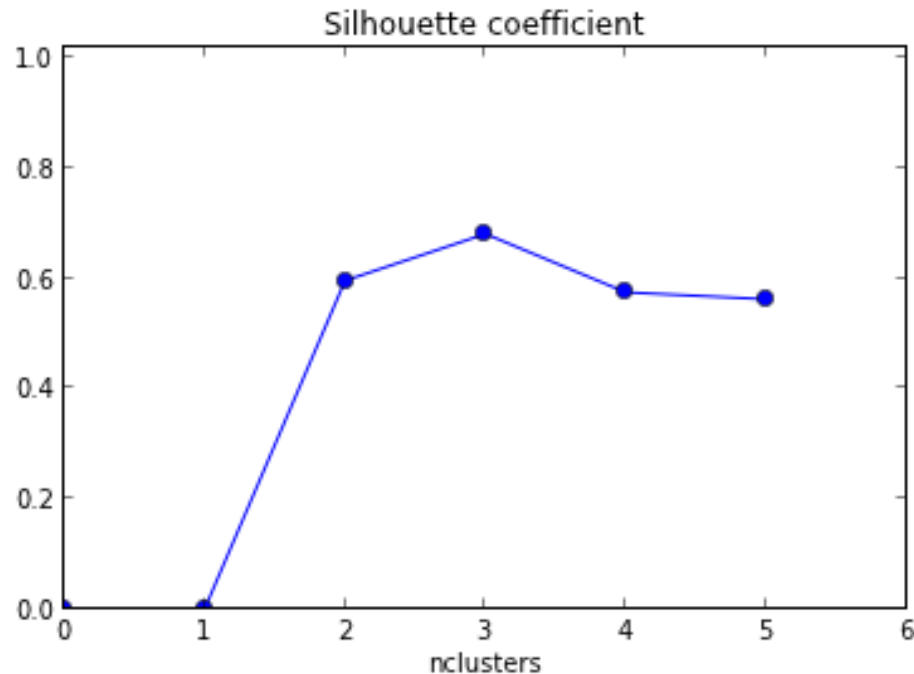
Elbow Method: involves plotting the sum of squared distances (SSD) or within-cluster sum of squares (WCSS) against different values of K . The idea is to look for the "elbow point" in the plot, where the rate of decrease in SSD/WCSS slows down significantly. This point can be a good estimate of the optimal K



<https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/>

Choosing Optimal 'K'

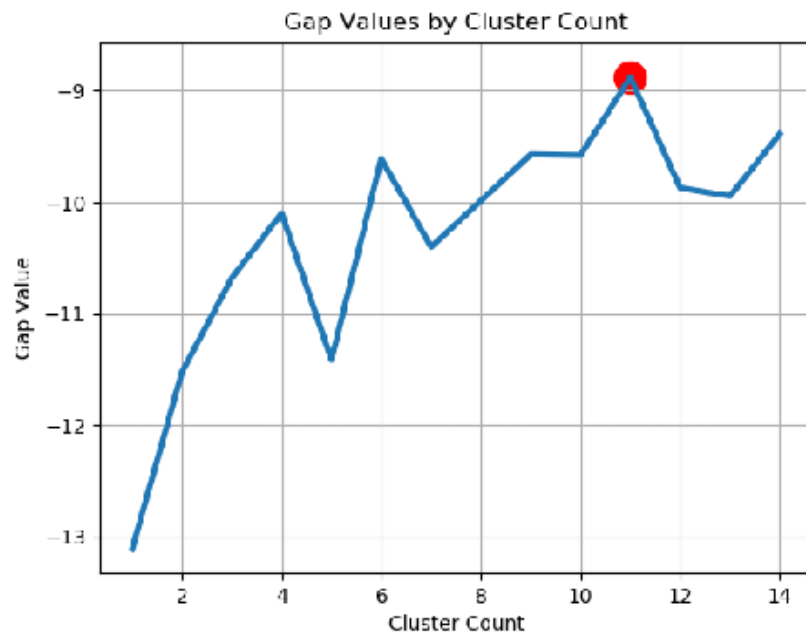
Silhouette Score: The Silhouette Score measures the quality of clustering by evaluating how well-separated the clusters are. It ranges from -1 to 1, where a higher score indicates better-defined clusters. The optimal K is often associated with the highest average silhouette score across different K values



<https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/>

Choosing Optimal 'K'

Gap Statistic: The Gap Statistic compares the total intra-cluster variation for different values of K with its expected variation under a reference null distribution of the data (typically generated by random sampling). The optimal K value is determined based on maximizing the gap statistic.



<https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/>

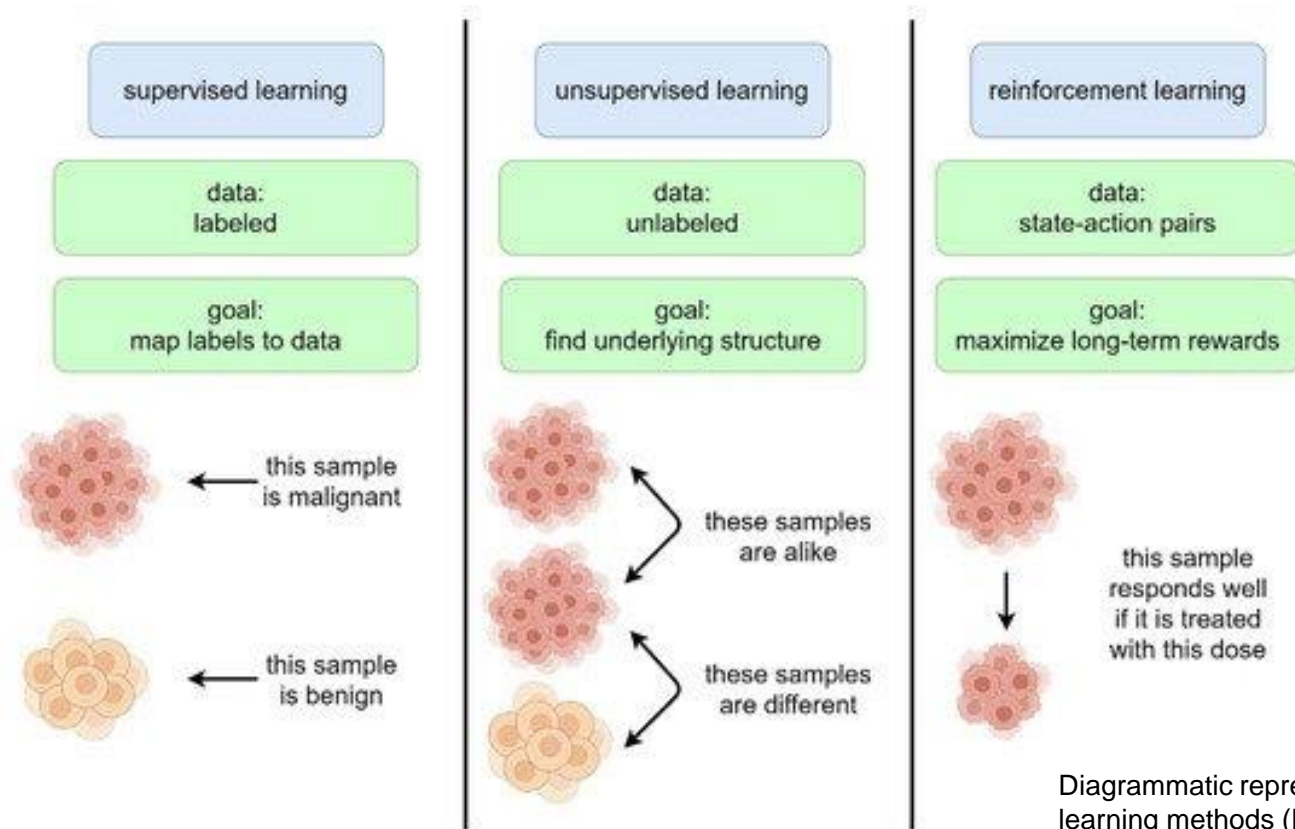
Choosing Optimal 'K'

Domain Knowledge and Interpretability:

In some cases, the optimal K can also be determined based on domain knowledge or the interpretability of the clustering results. For example, if clustering is used for customer segmentation, the number of customer segments may align with known customer types or business requirements



Refresher



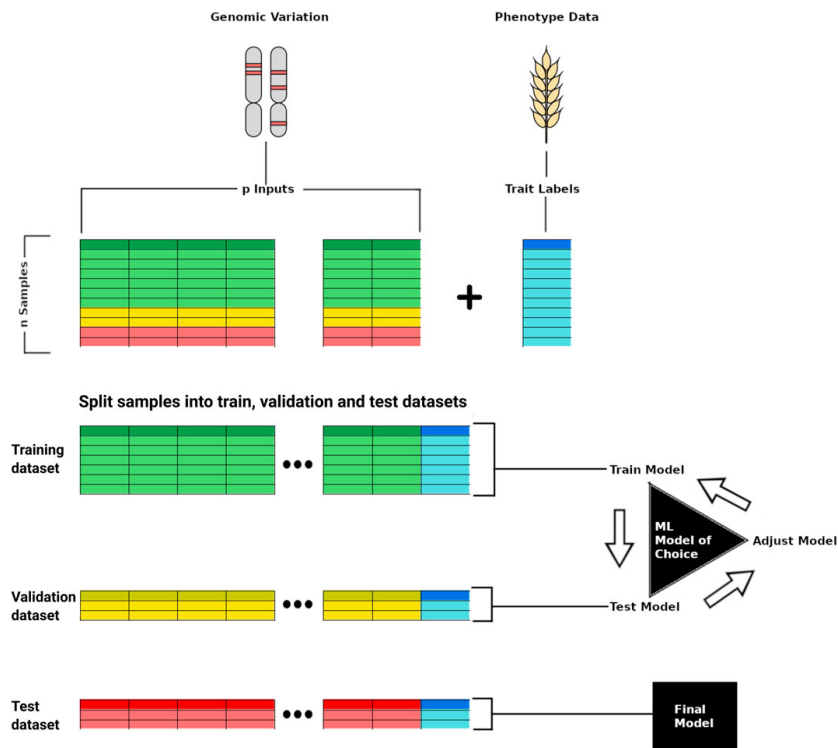
Diagrammatic representation of different learning methods (Eckardt et al., 2021).

Now the interesting Part

Applications Applications Applications



Genotype to Phenotype Prediction



Article

Deep Learning Framework for Complex Disease Risk Prediction Using Genomic Variations

Hadeel Alzoubi ^{1,*}, Raid Alzubi ¹ and Naeem Ramzan ²



RESEARCH ARTICLE

Machine learning approach to single nucleotide polymorphism-based asthma prediction

Joverlyn Gaudillo^{1,5*}, Jae Joseph Russell Rodriguez^{2*}, Allen Nazareno^{1*}, Lei Rigi Baltazar^{1,5*}, Julianne Vilela^{3*}, Rommel Bulalacao^{4*}, Mario Domingo^{4*}, Jason Albia^{1,5*}

1 Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, Philippines, **2** Genetics and Molecular Biology Division, Institute of Biological Sciences, University of the Philippines Los Baños, Philippines, **3** Philippine Genome Center Program for Agriculture, Office of the Vice Chancellor for Research and Extension, University of the Philippines Los Baños, Philippines, **4** Domingo Artificial Intelligence Research Center, Los Baños, Philippines, **5** Computational Interdisciplinary Research Laboratories (CINTERLabs), University of the Philippines Los Baños, Philippines

* These authors contributed equally to this work.

* jralbia@up.edu.ph



<https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.822173/full>

Biomarker Identification

scientific reports

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [scientific reports](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 13 September 2023


Machine learning approaches for biomarker discovery to predict large-artery atherosclerosis

[Ting-Hsuan Sun](#), [Chia-Chun Wang](#), [Ya-Lun Wu](#), [Kai-Cheng Hsu](#)  & [Tsong-Hai Lee](#) 

[Scientific Reports](#) **13**, Article number: 15139 (2023) | [Cite this article](#)

BMC Bioinformatics

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#) [Collections](#) [Join The Board](#)

[Submit manuscript](#) 

Research | [Open access](#) | Published: 15 January 2024

Methodology for biomarker discovery with reproducibility in microbiome data using machine learning

[David Rojas-Velazquez](#) , [Sarah Kidwai](#), [Aletta D. Kraneveld](#), [Alberto Tonda](#), [Daniel Oberski](#), [Johan Garssen](#) & [Alejandro Lopez-Rincon](#)



Multi-omics Integration:



MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis

Xiao Li^{1†}, Jie Ma^{1†}, Ling Leng², Mingfei Han¹, Mansheng Li¹, Fuchu He^{1*} and Yunping Zhu^{1*}

¹State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing, China, ²Stem Cell and Regenerative Medicine Lab, Department of Medical Science Research Center, State Key Laboratory of Complex Severe and Rare Diseases, Translational Medicine Center, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

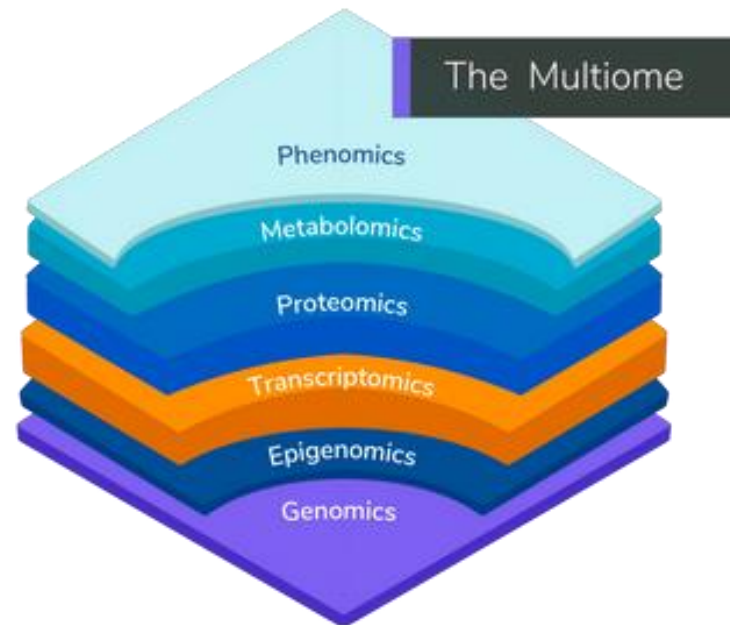
Genetics and Genomics

Graph machine learning for integrated multi-omics analysis

Nektarios A. Valous , Ferdinand Popp, Inka Zörnig, Dirk Jäger & Pornpimol Charoentong

British Journal of Cancer **131**, 205–211 (2024) | [Cite this article](#)

9403 Accesses | 1 Altmetric | [Metrics](#)



<https://synergymag.co/multiomic-age/>

Drug Target Identification and Drug Repurposing




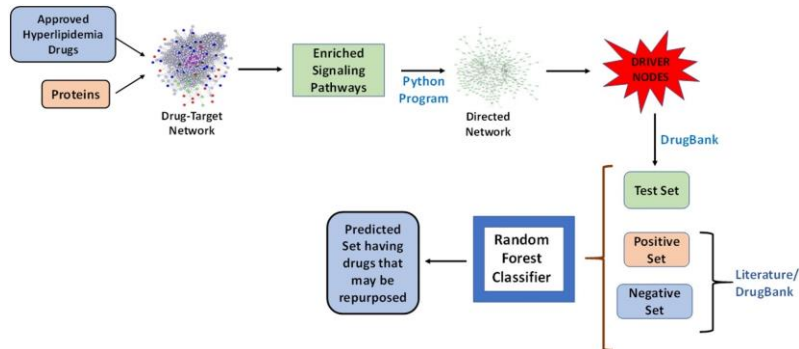
Computational Biology and Chemistry

Volume 92, June 2021, 107505



Drug repurposing for hyperlipidemia associated disorders: An integrative network biology and machine learning approach

Sneha Rai ^{a b 1}, Venuggopal Bhatia ^{a 1}, Sonika Bhatnagar ^{a c}  



Radiogenomics



Academic Radiology

Volume 31, Issue 6, June 2024, Pages 2464-2475



Neuroradiology

Machine Learning-Based MRI Radiogenomics for Evaluation of Response to Induction Chemotherapy in Head and Neck Squamous Cell Carcinoma

Zheng Li MD ^{a 1}✉, Ru Wang MD ^{b 1}✉, Lingwa Wang MD ^b✉, Chen Tan MD ^b✉, Jiaqi Xu MD ^b✉, Jugao Fang MD ^b✉, Junfang Xian MD, PhD ^a✉



Informatics in Medicine Unlocked

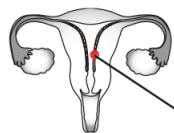
Volume 33, 2022, 101062



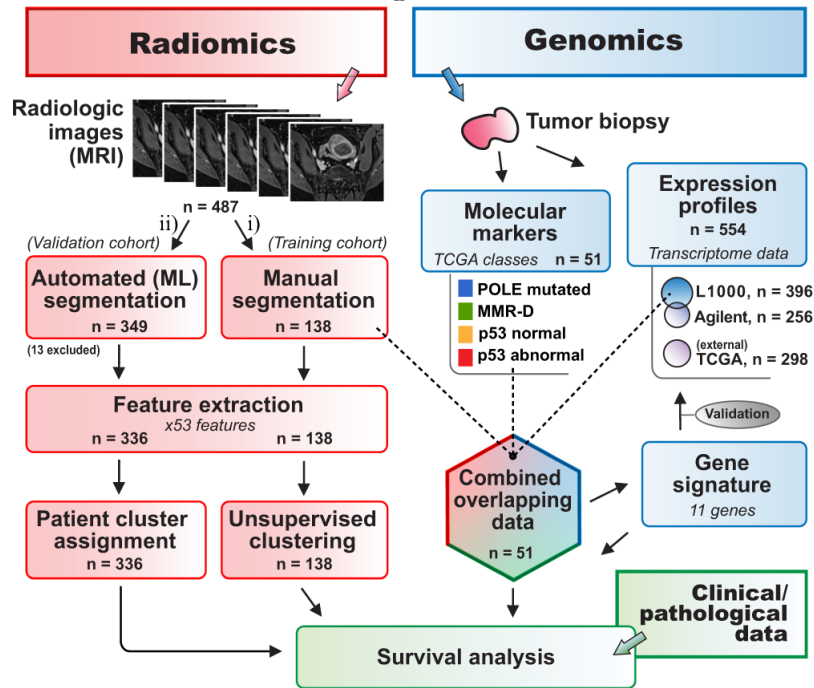
RadGenNets: Deep learning-based radiogenomics model for gene mutation prediction in lung cancer

Satvik Tripathi ^{a b}✉, Ethan Jacob Moyer ^c✉, Alisha Isabelle Augustin ^d✉, Alex Zavalny ^a✉, Suhani Dheer ^b✉, Rithvik Sukumaran ^a✉, Daniel Schwartz ^a✉, Brandon Gorski ^a✉, Farouk Dako ^e✉, Edward Kim ^a✉

Endometrial cancer

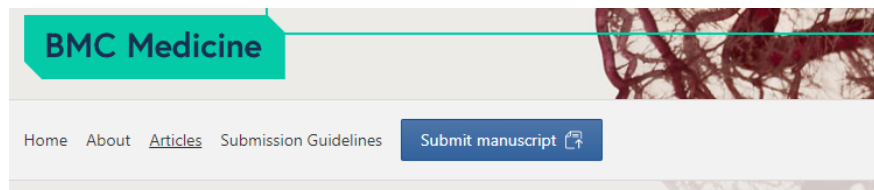


Primary tumor, n = 866 patients



<https://www.nature.com/articles/s42003-021-02894-5>

EHR (Electronic Health Records)



Commentary | [Open access](#) | Published: 16 July 2020

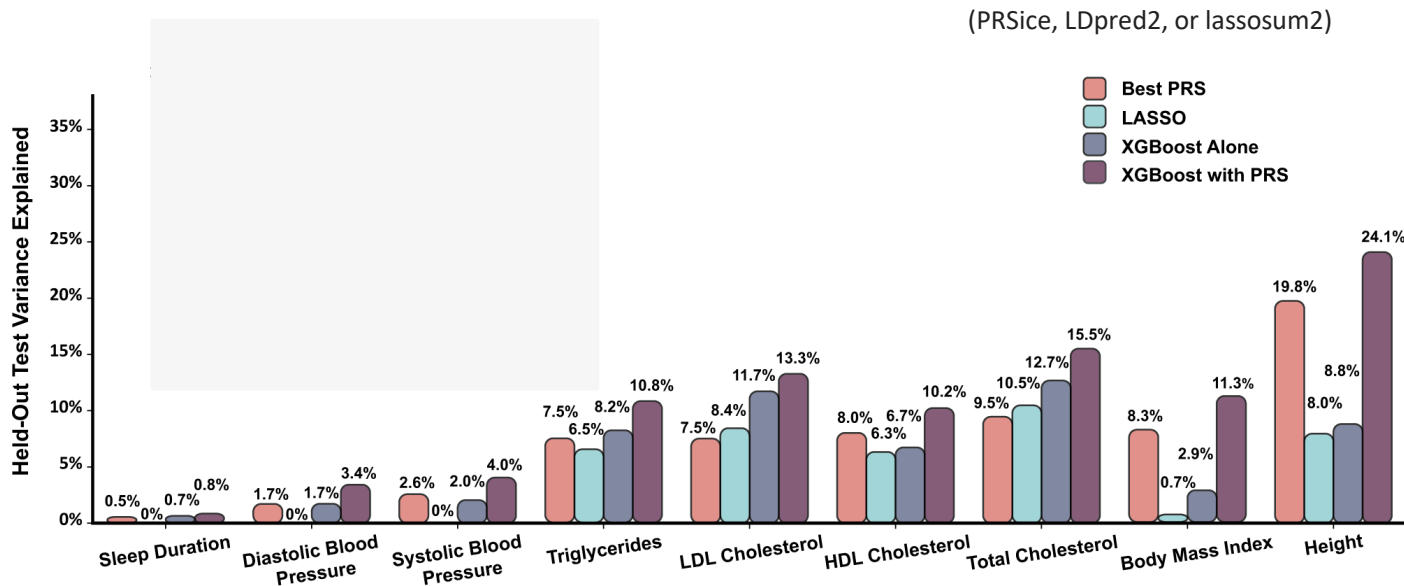
PREDICT *Prostate*, a useful tool in men with low- and intermediate-risk prostate cancer who are hesitant between conservative management and active treatment

Gaëtan Devos & Steven Joniau 

[BMC Medicine](#) **18**, Article number: 213 (2020) | [Cite this article](#)



ML - Optimized PRS



- Combining a PRS and XGBoost model results in a relative increase in the percentage variance explained (Elgart et al., 2022)

ML – Optimized MR

PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS  PEER-REVIEWED


RESEARCH ARTICLE

Deep mendelian randomization: Investigating the causal knowledge of genomic deep learning models

Stephen Malina , Daniel Cizin, David A. Knowles

Version 2 

Published: October 20, 2022 • <https://doi.org/10.1371/journal.pcbi.1009880>

Article	Authors	Metrics	Comments	Media Coverage
				

Considerations for the successful application of ML

- Need for Hyperparameter tuning
- Reproducibility
- Data Dependence -- Overfitting
- Imbalanced datasets
- Interpretability (Black Box)
- Need for individual level data
- Multi-omics : need for same set of samples across different modalities
- Missingness : Imputation vs Exclusion



Selected References

- Accordini, S., Lando, V., Calciano, L., Bombieri, C., Malerba, G., Margagliotti, A., Minelli, C., Potts, J., van der Plaat, D. A., & Olivieri, M. (2024). SNPs in FAM13A and IL2RB genes are associated with FeNO in adult subjects with asthma. *Journal of Breath Research*, 18(1), 16001. <https://doi.org/10.1088/1752-7163/acfbf1>
- Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Applied Sciences* (Switzerland), 13(12), 7082. <https://doi.org/10.3390/app13127082>
- Amiri, R., Razmara, J., Parvizpour, S., & Izadkhah, H. (2023). A novel efficient drug repurposing framework through drug-disease association data integration using convolutional neural networks. *BMC Bioinformatics*, 24(1). <https://doi.org/10.1186/S12859-023-05572-X>
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), 8124. https://doi.org/10.15252/MSB.20178124/SUPPL_FILE/MSB178124-SUP-0002-EVFIGS.PDF
- Avberšek, L. K., & Repovš, G. (2022). Deep learning in neuroimaging data analysis: Applications, challenges, and solutions. *Frontiers in Neuroimaging*, 1, 981642. <https://doi.org/10.3389/FNIMG.2022.981642>

Selected References

- Boulesteix, A. L., De Bin, R., Jiang, X., & Fuchs, M. (2017). IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine*, 2017. <https://doi.org/10.1155/2017/7691937>
- Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Churchhouse, C., Dumont, A., Eriksson, N., Gandal, M., Goldstein, J. I., Grasby, K. L., Grove, J., ... Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention-deficit/hyperactivity disorder. *Nature Genetics*, 51(1), 63. <https://doi.org/10.1038/S41588-018-0269-7>
- Dou, Y., Tan, S., & Xie, D. (2023). Comparison of machine learning and statistical methods in the field of renewable energy power generation forecasting: a mini review. *Frontiers in Energy Research*, 11, 1218603. <https://doi.org/10.3389/FENRG.2023.1218603/BIBTEX>
- Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J. A., Guo, X., Lin, H. J., Raffield, L., Gao, Y., Chen, H., de Vries, P., Lloyd-Jones, D. M., Lange, L. A., Peloso, G. M., Fornage, M., Rotter, J. I., Rich, S. S., Morrison, A. C., Psaty, B. M., ... Sofer, T. (2022). Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology* 2022 5:1, 5(1), 1–12. <https://doi.org/10.1038/s42003-022-03812-z>

Thank you

itunu.isewon@covenantuniversity.edu.ng