ECRONICON OPEN ACCESS

EC PHARMACOLOGY AND TOXICOLOGY

Review Article

Basics of Machine Learning in Drug Discovery: A Bird's Eye Perspective

Rekha Choudhary[1], Pranali Yelchatwar[1], Vinayak Walhekar[1], Ashwini Patil[1], Dileep Kumar[1], Amol Muthal[2], Macha Baswaraju[3], Garige Anil kumar[3], Chandrakant Bagul[4]* and Ravindra Kulkarni[1]*

[1]Department of Pharmaceutical Chemistry, BVDU'S Poona College of Pharmacy, Erandwane Pune, Maharashtra, India.

[2]Department of Pharmacology, BVDU'S Poona College of Pharmacy, Erandwane Pune, Maharashtra, India

[3]Department of Pharmaceutical Chemistry, Jayamukhi Institute of Pharmaceutical Sciences, Narsampet, Warangal, Telangana, India

[4]Department of Pharmaceutical Chemistry, Amrita School of Pharmacy Amrita Vishwa Vidyapeetham, AIMS Health Sciences Campus, Kochi

***Corresponding Author:** Chandrakant Bagul, Department of Pharmaceutical Chemistry, Amrita School of Pharmacy Amrita Vishwa Vidyapeetham, AIMS Health Sciences Campus, Kochi.

Received: May 02, 2022; **Published:** August 31, 2022

Abstract

Artificial Intelligence is an algorithm based computational approach to find solution from existing data. It has wider applications in different areas like medicine, agriculture, etc., and pharmaceutical field is not out of its ambit. Machine learning (ML) that is a subset of artificial intelligence, plays a crucial role in drug discovery and development by employing a massive quantum of structured and semi-structured data so that a ML model can create accurate outputs or provide predictions based on the data under analysis. It is a technical approach in which the machines are trained to process a significant amount of data. Two main categories of ML are supervised and unsupervised learning which are the platforms for the data processing. ML uses historical data which is responsible to generate algorithms for its working to produce the output. This can be accomplished in the ligand-based and structure-based approaches of drug design that helps to anticipate the hit and lead molecules for the drug discovery process.

Keywords: *Artificial Intelligence; Machine Learning; Supervised Learning; Unsupervised Learning; Algorithms*

Introduction

We can see the rapid transformation of data around us in day-to-day life that makes the generated data handling tedious, so its management is important. For example, there are variety of sources available in the surroundings including IOT, network security data, digital data, trade info, cellular phone data, social platform data, data concerning health, SARS-CoV-2 data, etc. Understanding the data pattern and extracting valuable information from such vast data is a challenging task. The rapid increments in the biological data raise copious concerns like adequate data management, its storage, and collecting valuable information from it [1]. It mandates the creation of various methodologies and tools capable of converting these massive databases into behavioral genetics. In the last few decades, machine learning technologies are being widely used in chemo-informatics. It assists the biologists in gaining meaningful conclusions from these data, leading to new drug designs and discoveries [2].

Artificial intelligence (AI) is defined as the simulation of human intelligence in robots that were trained to think and act like humans. Machine learning is a subset of artificial intelligence. It is the most efficient technique to create an AI model that can perform everything

from categorizing data to predicting the outcomes. Machine learning (ML) is used to describe methods or algorithms that allow computers to automate data-driven model programming and build models by recognizing patterns in statistically relevant data. It is the process of directing computers to maximize the data analysis based on example data and past experience. In a modelling task, the streamlined measure might be the effectiveness delivered by a predictive model as well as the degree of a performance or evaluation function in an optimization process [3]. The difference between traditional and machine learning approach is shown below in figure 1. Computational intelligence is used in drug development to examine, learn, and explain how pharmaceuticals were found using AI to identify several remedies in a pre-programmed and integrated fashion. As a result, numerous pharmaceutical companies have shown a stronger desire to contribute technology and resources for obtaining precise drug discovery findings. Finally, this review presents AI strategies in the drug discovery field to address many drug discovery and development applications using ML techniques. Also, the AI sector predicts results in medicinal research and discovery regarding computer intelligence.

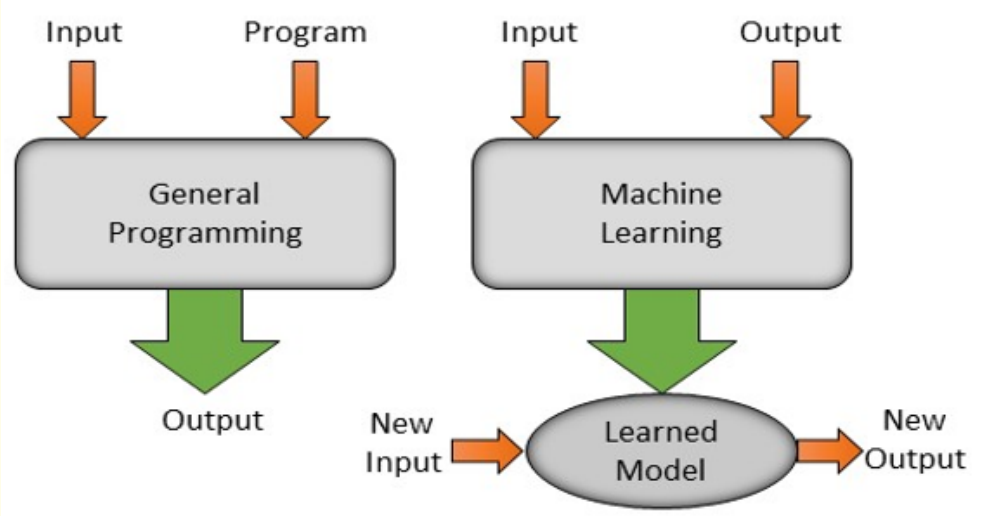

Figure 1: *Difference between traditional and machine learning approach.*

The processes and approaches used in ML are homogenous subset of AI. Supervised and unsupervised learning are the two basic types of machine learning algorithms. Such ML techniques have greatly aided drug development. Various ML algorithms in drug development have benefitted pharmaceutical industries significantly. ML algorithms have been used to construct multiple models for predicting substance's chemical, biological and physical properties in drug development [4]. All stages of the drug discovery process can benefit from machine learning techniques. For example, machine learning algorithms have been used to identify new medication uses, predict drug-protein interactions, uncover drug efficacy, assure safety biomarkers, and improve molecular bioactivity. Random Forest (RF), Naive Bayesian (NB), and Support Vector Machine (SVM) are examples of machine learning techniques that have been frequently employed in drug development [5]. This review article explores the usage of these new approaches in recent years in research. Analyzing the current state of the art in this sector will offer us a sense of where cheminformatics may evolve shortly and the limits and beneficial outcomes it has produced. It will focus on various Machine Learning algorithms used for drug discovery in recent years. The machine learning tree is represented in figure 2.

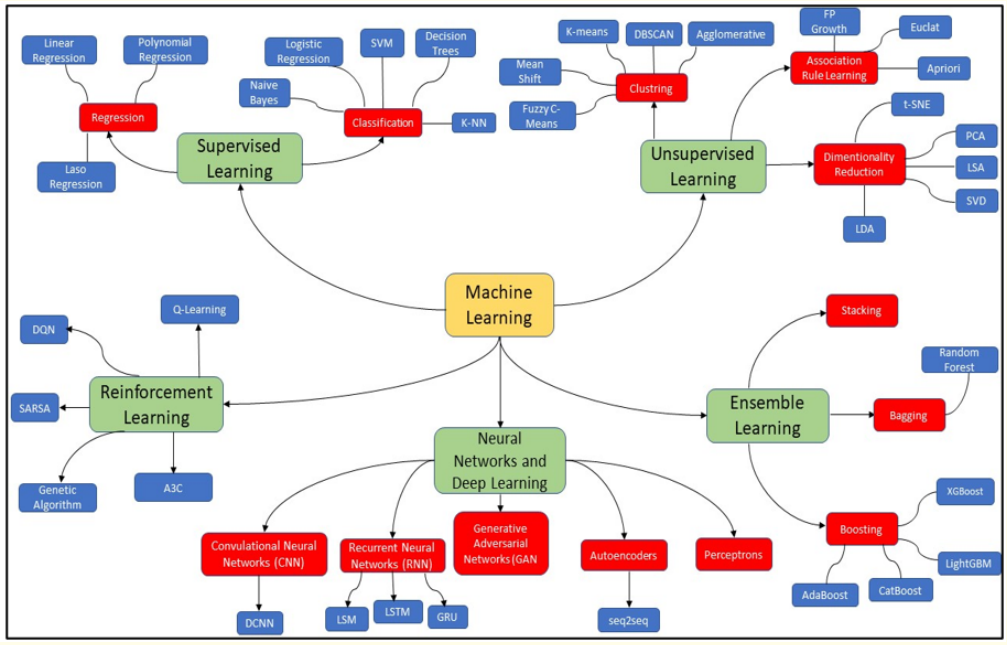

Figure 2: *Machine Learning Tree.*

Data types and algorithms in ML

Types of data

Typically, data availability is an essential factor in developing a machine learning model. In general, the data employed in ML could be categorized into structured, semi-structured or unstructured and their brief descriptions are as follows [6]:

- **Structured:** It has a structured format, is highly organized and is readily available. Structured data is typically stored quantitatively in a very well-organized arrangements or formats including rational databases (RDB). Structured data includes names, addresses, dates, times, among other things.

- **Unstructured:** It has no predetermined format or organization, making it more difficult to encapsulate, operate and investigate, as it predominantly consists of text and multimedia content. It includes email messages, blog entries, sensor data and text files.

- **Semi-structured:** In comparison to structured information, semi-structured data is not stored in a RDB but it does have managerial properties that make it easier to examine. JSON articles, XML, MySQL databases, HTML and other semi-structured data formats are used to represent semi-structured data.

Types of ML algorithms

ML algorithms are divided into three categories depending on the nature of the learning: supervised learning, unsupervised learning and reinforcement learning. In upcoming sections, we will sum up each type of learning techniques in a comprehendible approach [7].

- **Supervised ML algorithm:** Supervised ML algorithms are capable of forecasting future events by applying previous knowledge to new data using annotated data (or labelled data that machines can recognize, understand and memorize the input data using ML algorithms). After examining a predefined training data, the supervised learning generates an inferred function that is used to predict untested new data. After sufficient learning, the algorithm could provide thresholds about any new instances. It can also start comparing the outcomes to the right ones and identify errors in the model to continue to improve [8]. The flow chart of supervised machine learning is depicted below in figure 3.

- **Unsupervised machine learning:** Unsupervised learning utilizes machine learning algorithms to assess and group unlabeled sets of data. This algorithm is accustomed to identify hidden patterns within data and eliminate the need of human involvement. The system does not find out the proper output, but it does examine the data and can make conclusions from it to depict unlabeled data's unseen structures.

- **Reinforcement machine learning:** This type of ML algorithm permits software agents and computer machine to proactively evaluate acceptable behavior in a specific environment and to enhance efficiency.

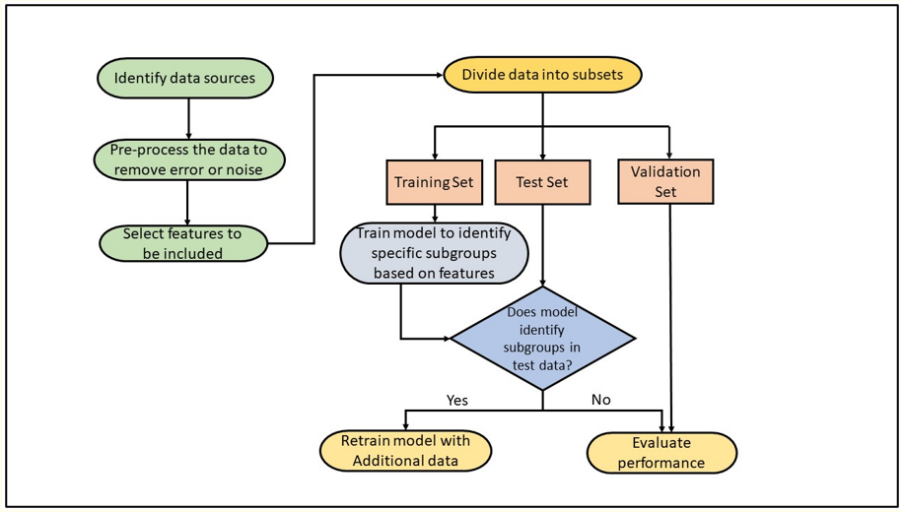

Figure 3: *Flow chart of supervised machine learning.*

Supervised machine learning algorithm

In the model generation the supervised learning is classified into two types:

- **Classification:** It finds specific records in the training data set which subsequently conclude how these subjects should be defined or annotated. Examples of classification techniques in ML include Support Vector Machine (SVM), decision trees, random forest, k-nearest neighbor and linear classifiers.

- **Regression: -** It is employed to discover the interaction between dependent and independent parameters thus useful in making predictions. Logistic regression, linear regression and polynomial regression are the three prominent regression techniques consistently considered in regression model generation.

The classification and regression model are represented in figure 4.

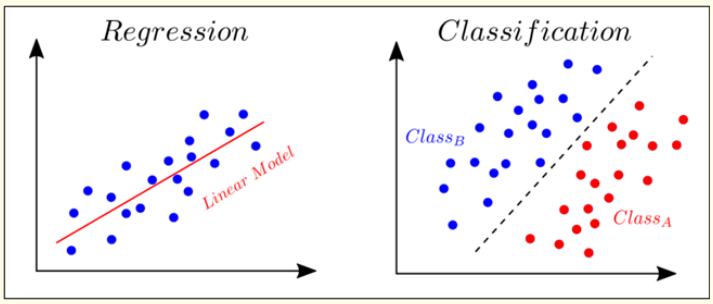

Figure 4: *Classification and regression model.*

Support vector machine

SVMs are supervised ML that are utilized in classification, regression modelling and outlier detection. SVM works on the algorithms fundamental basis of margin estimation theory. SVM strategy requires generating a hyperplane that separates and optimizes the margin between the classifier and the nearest instance of each category that falls within the constraints of the margin of such hyperplane which is also known as the separation margin. Data points on the hyperplanes are known as support vectors [9] and are depicted in figure 5.

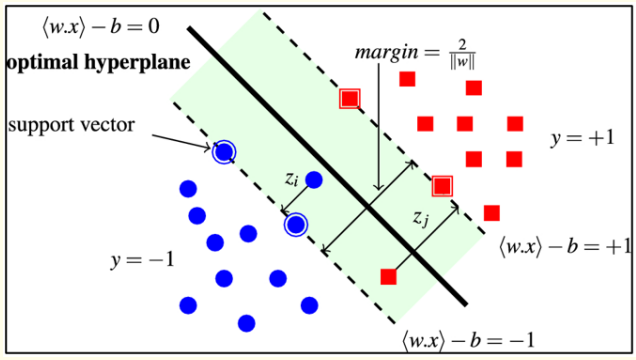

Figure 5: *Support vector machine and hyperplane.*

- It is highly potent and flexible in multi-dimensional space because it can be defined by various kernel functions for the decision boundary.
- The separating hyperplane can be written as: f(x)=<w.x>+b=0
- One of the two applications of SVM were employed in 3D QSAR analysis for evaluating the efficacy of HIV integrase inhibitors and BRAF-V600E, SVMs.
- The algorithm selects the best hyperplane based on the cost function (weight vector and bias) and is mathematically illustrated as

$$minimize\ \frac{1}{2}||w||^2 + X\sum_{i=1}^{n} s_i$$

$$Subject\ to: y_i(w^T x_i + b) \geq 1 - s_i \text{ and } S_i \geq 0, for\ i = 1,2,\ldots.,n$$

Decision tree

It is a well-known multivariate supervised learning approach. It is applicable for both classification as well as regression tasks. As depicted in the figure 6, a decision tree characterizes instances separating the tree from root to leaf nodes. Samples are categorized by inspecting the attributes defined by that node starting from the root node in descending the branches and correlating to the feature value [10]. The goal behind the decision tree is to generate a model that predicts the target variable by learning basic decision rules inferred from the data features.

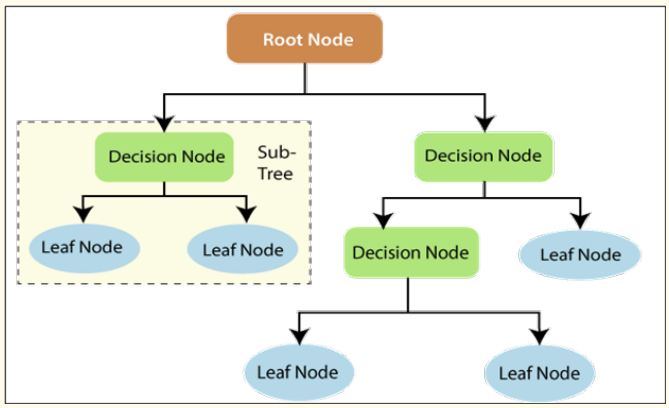

Figure 6: *Decision tree.*

The highly prevalent criterion for partition are "entropy" which stands for the information gained and "Gini" for the Gini impurity, which can be mathematically described as [11,12]:

$$Entropy: H(x) = -\sum_{i=0}^{n} p(x_i) log_2 p(x_i)$$

$$Gini(E) = 1 - \sum_{i=0}^{n} p_i^2$$

Naive bayes classification

It is based on Bayesian statistics and assumes that each pair of features are independent. It performs well and can be used in numerous real-life scenarios, such as a report or text categorization for binary and multi-class groupings [13]. This classifier can be used to efficiently classify noisy occurrences in data and establish a comprehensive predictor. The main advantage is that in comparison to more sophisticated approaches, it requires a small fraction of training data to estimate the necessary functionality. However, due to its strong assertions on feature independence its efficiency may undergo changes. The most widely accepted naive Bayes classifier versions

are Gaussian, Multinomial, Complement, Bernoulli and Categorical [14]. The Bayes formula serves as the foundation for the naive Bayes theorem and is expressed as:

$$P(Y|Z) = \frac{P(Y)P(Z|Y)}{P(Z)}$$

Algorithm of Naive Bayes:

→ *For Each value of* y_n
Estimates $P(Y = y_n)$ *from the data.*
For each value of z_{ij} *of each attribute* Z_i
Estimates $P(Z_i = z_{ij}|Y = y_n)$

→ *Classify a new point via:*
$Y_{new} \leftarrow Arg \max_{y_n} P(Y = y_n) \prod_i P(Z_i = z_{ij}|Y = y_n)$

Where, Z: attributes, Y: class, : probability of even Y given Z has occurred, : probability of even Z given Y has occurred, : probability of event Y, : probability of event Z.

Advantages

It is easier to implement, highly efficient, works well with small training set, grows linearly, ascends with number of variables and data points which are flexible for both binary and classification problems in multiclass and facilitates prediction with probabilistic confidence. It works with both categorical and continuous variables.

Disadvantages

If at least one of the features must be a "continuous variable" (such as time), it is difficult to employ NB directly if "buckets" for "continuous variables" are often incorrect. Scaling will be cumbersome if classes involved are as more than 100K.

Application

NB can be applied effectively as a recommendation system for cancer relapse forecast and post-radiotherapy developments.

K-Nearest Neighbours (KNN)

It is based on multivariate non-parametric approach that is used for classification problems, regression analysis and pattern recognition. KNN is highly prevalent classification type machine learning algorithms [15]. Here figure 7 depicts about KNN algorithm.

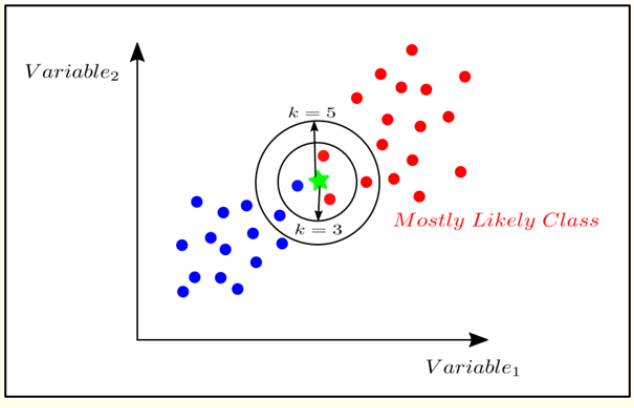

Figure 7: *Graphical representation of K- nearest neighbors.*

Each set of data in a multi-dimensional region is considered as a point. As a result, each training data set is located in this region. So, when data sample is categorized, the KNN searches the multi-dimensional region for the k training data that are in proximity to the input provided t. All the k training data is considered as the input instance's "nearest neighbours". The "closeness" in the defined training data and input is characterized using a distance metric [16]. The most prevalent class among the "neighbours" of the provided input sample can be categorized. This method consists of three major components:

- A series of annotated training data samples
- A separation in distance between data samples
- Nearest neighbors K value.

The distance from the identified item is determined to categorize the input samples, followed by the number of close neighbors. Furthermore, the supplied data sample label is classified correctly of k nearest neighbors.

Distance functions

To determine which points are "closest" to each other, here are a few examples of standard distance functions:

Euclidean Distance	$D(\vec{a},\vec{b}) = \sqrt{\sum_i^n (a_i - b_i)^2}$
Manhattan Distance	$D(\vec{a},\vec{b}) = \sum_i^n \lvert a_i - b_i \rvert$
Hamming Distance	$D(\vec{a},\vec{b}) = \sum_i^n I(a_i, b_i)$ I(p,q) = 0 if identical, 1 if different.

Advantages

The model is inexpensive, simple to integrate and performs well on multiclass problems.

Disadvantages

Computation time is considerable (lazy learning), unspecified records are costly to classify, running a process needs a lot of memory and the KNN can be expensive in determining K if the dataset is massive.

Application

In breast cancer research, KNN algorithms are performed. The K-NN classifier has been broadly used in biomedical signal analysis and rigorous disease diagnosis algorithms implemented to MRI.

Random forest (RF)

Random Forest is a popular ML algorithm that can be used for both regression as well as classification problems. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complex problem and increase the model's performance. Instead of depending on a single decision tree, the random forest analyzes the predictions from each tree and predicts the final output

based on the majority votes of forecasts. The principle of RF is to construct and consolidate complementary classification trees to maximize their variations. This is a pack of trees constructed from a training sample set and independently verified to examine whether the solutions rendered assists a prediction of future data [17,18]. Figure 8 and 9 shows about bagging and boosting aggregation.

- **Bagging:** A sample of data from a training set is picked randomly with replacement which indicates that individual data points can be selected repeatedly. After creating several data samples, these poor models are trained independently based on the problem type-regression or classification.

- **Boosting:** It is an iterative aggregation approach for reducing bias error and achieve good predictive model. The word 'boosting' describes methodologies that transform a weak learner into a strong learner. Since the data samples are weighed, some may engage in the new sets more frequently. Any data points that are inaccurately predicted are identified and their weights are increased in each iteration so that the next learner pays additional focus to get them right [19].

Random Forest Algorithm:

1. *For b = 1 to B:*
   a. *Draw a bootstrap sample D^* of size N from the training data.*
   b. *Using bootstrapped sample data, develop a random-forest tree T_{bs} by successively replicating the basic steps at each terminal node of the tree till the minimum node size n_{min} is achieved.*
      → *Randomly select m variables from the k variables*
      → *Choose the best variable/split-point from the m.*
      →*Divide the node into two daughter nodes.*
2. *Output the ensemble of trees $\{T_{bs}\}_1^B$, to make a prediction at a new point x:*
   - *Regression:* $\hat{F}_{RF}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_{bs}(x)$
   - *Classification: Let $\hat{C}_b(x)$ be the class prediction of the b^{th} random-forest tree then: $\hat{C}_{RF}^B(x)$ = majority vote $\{\hat{C}_b(x)\}_1^B$*

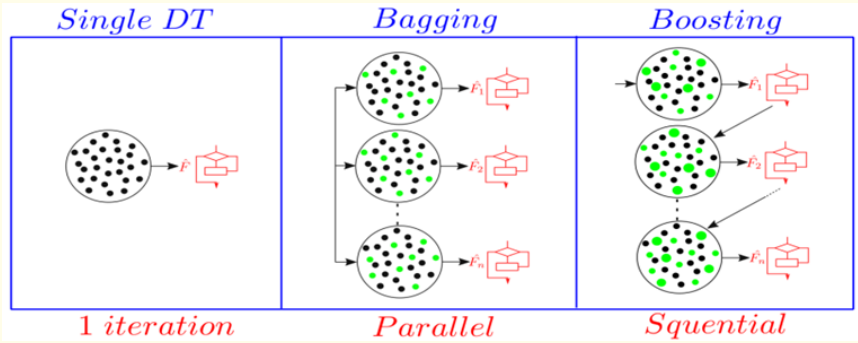

Figure 8: *Random Forest.*

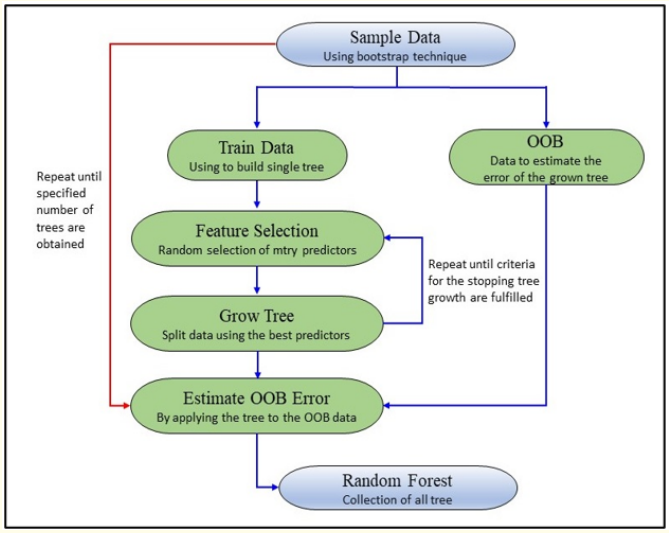

Figure 9: *Flow chart of random forest algorithm.*

Linear regression

This algorithm is used to assess the importance of a dependent variable (also known as explanatory variable) relying on an independent variable (also known as predictor variable). It is basically a technique for determining the relationship among the variables. It is an analytical method in which one or more independent variables are used to predict the dependent variable. Correlation measures the intensity or degree of an association among the two variables whereas linear regression analysis statistically exemplifies the existing relationship. The correlation coefficient "r" is a unitless parameter that lies between −1 to +1. The value of correlation coefficient "−1" shows an inverse relationship and "+1" shows a positive relationship [20,21]. The figure 10 depicts linear regression algorithm.

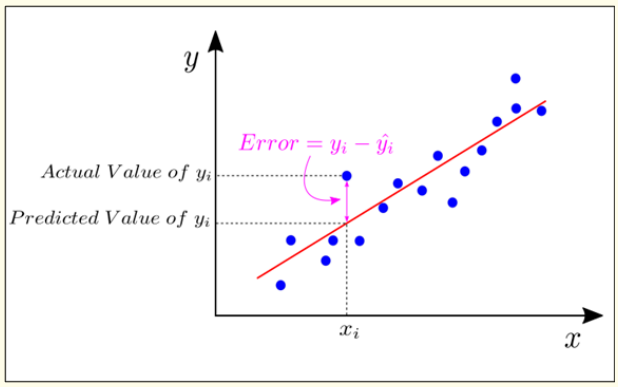

Figure 10: *Linear regression algorithm.*

The linear regression problems are expressed by the following relation, , which represents the model of best curve fitting in association between x (independent or predictor variable) and y (dependent or response variable). The coefficient of regression gives information about how much y varies with x. The regression model can be written as:

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

Where, α is gradient of the line, β is the intercept of the line, and is the error part. Given some estimates $\hat{\alpha}$ and $\hat{\beta}$ for the model coefficients, we predict the value of Y using.

$\hat{y}=\hat{\alpha}\ x+\hat{\beta}$

where indicates a prediction of Y based on X = x

Estimation of the parameters by least squares

Let $\hat{y}=\hat{\alpha}\ x+\hat{\beta}$, x_i be the prediction for Y based on the i^{th} value of X. Then $e_i=y_i-\hat{y}_i$,, represents the i^{th} residual. The residual sum of squares (RSS) is denoted by

RSS = $e_1^2 + e_2^2 + \cdots + e_n^2$

The least-squares method selects to minimize the RSS, and it can be demonstrated as:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x}$$

in which $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Multiple linear regression

The MLR method assists in identifying the relationship between dependent and independent parameters [22]. Only one classifier or attribute is used to predict the response variable in linear regression models. Several classifier or attributes are used to predict the response variable in multiple linear regression models (develop linear relationships between various independent and dependent variables). The following expression can demonstrate the generated regression model [23,24]:

$$y = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + e_{i,j}$$

in which represents the independent or predictor variables, y represents the dependent or response variable and is the error. If there are N independent x variables that are linearly associated with the y variable then the model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N + e_1 + e_2 + \cdots + e_N$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N + e_y$$

In the matrix form: $y=X\beta+e_y$ and MLR is used to estimate regression vector β

$$\beta = (X^T X)^{-1} X^T y$$

The deviation of observed and true value is written as . Developed models are commonly estimated using the ordinary least square method. Where, is the residual error expressed as the subtraction between , the actual value and predicted values, of the response variable, *y* [25].

Application

The MLR is effectively applied in the investigation of the chemical compound's quantitative structure-activity relationship (QSAR). Where, y represents the response variable corresponding to various physicochemical parameters or pharmacological activities, and x represents the predictor variable corresponding to the molecular descriptor.

Logistic regression

It is an effective ML method for binary classification problems (type of target variable is categorical). Logistic regression is significant as linear regression, but it's for classification tasks. The crucial parameter for logistic regression is the logistic function to create a model of a binary output variable. The main difference between logistic and linear regression is that the range of logistic regression is confined to 0 and 1. Furthermore, logistic regression does not mandate a linear relationship between predictor and response variables, unlike linear regression. It is demonstrated by the nonlinear function of the natural log to the odds ratio expressed as [26]:

$$Logistic\ Fuction = \frac{1}{1+e^{-x}}$$

The "Logit function," also known as conditional probability, serves as the foundation of logistic regression. It is regarded as a natural log of the odds. The ranges of conditional probability from 0 to 0.5 in class 0 and 0.5 to 1 are in class 1 [27]:

$$odds = \frac{P}{1-P} \Longrightarrow Logit(P) = ln\left(\frac{P}{1-P}\right)$$

Let us make logit of P equal to to y=mx+c, which yields:

$$\text{Logit(P)} = mx + c \Longrightarrow mx + c = \ln\left(\frac{P}{1-P}\right)$$

$$\text{P} = \frac{e^{mx+c}}{1+e^{mx+c}} \Longrightarrow \text{P(x)} = \frac{1}{1+e^{-(mx+c)}}$$

The schematic diagram for logistic regression is shown in above figure 11.

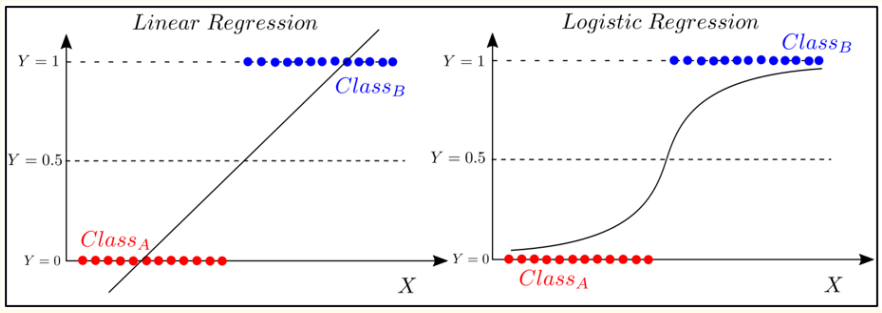

Figure 11: *Schematic diagram for Logistic regression.*

Application

QSAR between structural descriptors and biological activity of carbonic anhydrase inhibitors have been developed using Binary Logistic Regression as non-linear approaches [28].

Performance Indicators

For the classification problem, a variety of performance measures can be introduced [29,30].

Confusion matrix

It is a matrix representation in which columns contain predictor variables and rows have actual variables, also known as a contingency table are given in figure 12.

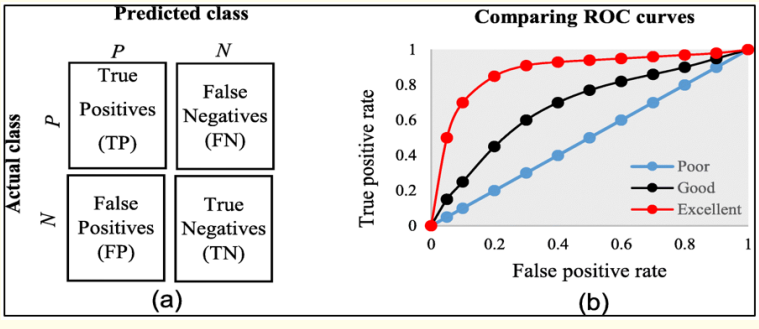

Figure 12: *Classification performance evaluation.*

Area under the curve (AUC)

It is also known as the receiver operating characteristics (ROC) and is mainly used for solving problems related to classification. It attempts to measure the model's capability to differentiate between data. ROC value 1 indicates a strong classifier and 0.5 indicates about presence random guess.

- **Accuracy*:*** The percentage of correctly classified samples is used to measure the accuracy,

  $$\text{Accuracy} = \frac{\text{TrueP}+\text{TrueN}}{\text{TrueP}+\text{TrueN}+\text{FalseP}+\text{FalseN}}$$

- **Sensitivity*:*** It is defined as percentage of positive lists within all positives,

  $$\text{Sensitivity} = \frac{\text{TrueP}}{\text{TrueP}+\text{FalseN}}$$

- **Specificity*:*** the proportion of tests which are negative from all negatives is defined as specificity,

  $$\text{Specificity} = \frac{\text{TrueN}}{\text{TrueN}+\text{FalseP}}$$

- **Precision*:*** It is expressed as the fraction of true positives among all positives,

  $$\text{Precision} = \frac{\text{TrueP}}{\text{TrueP}+\text{FalseP}}$$

The most widely used statistical performance measure for a regression problem is

- **Root Mean Square Error (RMSE):**

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

- **Mean Absolute Error (MAE):**

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

Where $\bar{Y}_i$ represents the estimated value and y_i represents the true value.

Insights to unsupervised machine learning

Unsupervised learning (UL) depicts the process of system learning for the analysis of data by random means. There are no stated goal outputs rather, the UL brings previous biases to bear on what features of the structure of the input should be represented in the output [31]. Indeed, the anatomical and physiological features of synapses in the neocortex are known to be significantly impacted by sensory neuron activity patterns [32]. While learning, almost no information about the composition of scenes is revealed so, human brain is a classic example of unsupervised learning. Clustering methods are the primary basis for UL [33].

Clustering

It is the task of grouping a collection of data points so that data information in the similar classes or groups are more comparable to data points in another group. Categorizing the data instances by clustering method that are similar to each other in a cluster and data that are significantly divergent in multiple clusters [34]. A cluster is portrayed by a unit location known as the cluster's centroid (or cluster center). Centroid is calculated by mean of all data points in the cluster.

$$C_j = \sum X_i$$

The cluster boundary is determined by the cluster's furthest data point.

Types of clustering analysis

The clustering analysis is majorly classified into three types mentioned below as:

1. Exclusive Clustering: K-means
2. Overlapping Clustering: Fuzzy C-means
3. Hierarchical Clustering: Agglomerative clustering, divisive clustering.

Exclusive clustering

K-means

The K-means method is an iterative technique that attempts to split a dataset into K separate non-overlapping subgroups (clusters), each of which contains just one data point. It distributes data points to clusters in such a way that the sum of the squared distances between them and the cluster's centroid (arithmetic mean of all the data points in that cluster) is as little as possible. Within clusters, lesser

the variance, more homogenous the data points are [35]. The plots for K-means clustering are shown in figure 13. The process goes in the following way

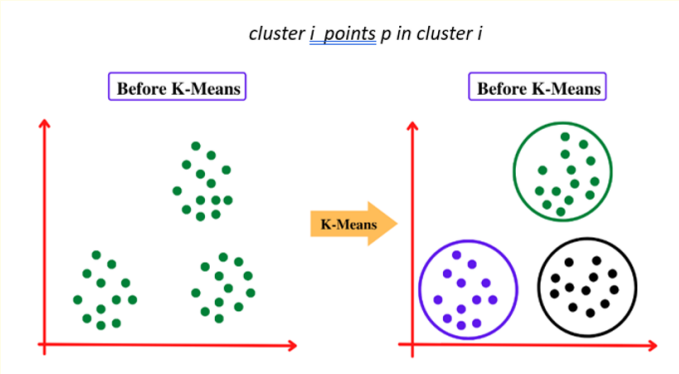

Figure 13: *K- means clustering plots.*

1. Initialize the cluster centres, c_1,..., c_K, in a random order.
2. Given cluster centres, select points in each cluster - Find the nearest c_i for each point *p*. Insert *p* into the cluster *i*.
3. Using the points in a particular cluster, find c_i – and assign c_i equal to the mean of the points in cluster *i.*
4. Repeat Step 2 if c_i has modulated [36].

Properties

1) It renders some sort of conclusion.
2) It is possible to have a "local minimum."
3) It is not always possible to obtain the global minimum of the objective function:

$\Sigma \Sigma \|p\text{-}c_i\|^2$

Fuzzy C-means clustering

On the basis of the distance between the cluster center and the data point, this method assigns membership to each data point corresponding to each cluster center. The closer the data is to the cluster center, more likely it is to belong to that cluster center. Clearly, the total of each data point's membership should equal one. Membership and cluster centers are adjusted after each cycle using the formula [37]:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{i,j}^{m} \left\| x_i^j - C_j \right\|^2$$

where $1 \leq m < \infty$, an extension of k-means

Algorithm

Assume x_i as a vector of values for g_i

1. By chance, initialize membership $U^{(0)} = [u_{ij}]$ for data point cluster cl_j containing data point *gi.*
2. Upon reaching to K[th] step, fuzzy centroid be computed $C^{(k)} = [c_j]$ for $j = 1,..., nc$, in which *nc* represents number of clusters, using [38]

   Here,

   m = fuzzy parameter

   n = number of data points
3. Use $U(k) = [uij]$ to update the fuzzy membership.
4. Suppose $||U^{(k)} - U^{(k-1)}||$, then STOP, alternatively, come back to step 2.
5. Compute the membership threshold.

Allot the g_i to cluster cl_j if *uij* of $U^{(k)}$ > to each data point g_i.

Figure 14 to 16 signifies Fuzzy C-means clustering

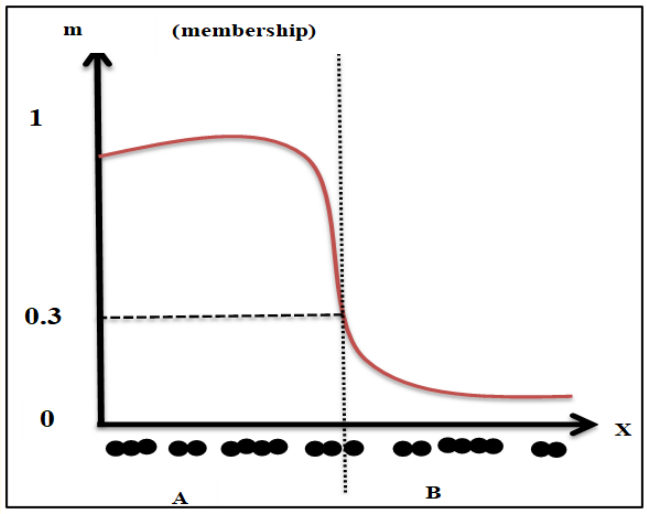

Figure 14: *Fuzzy C-means clustering.*

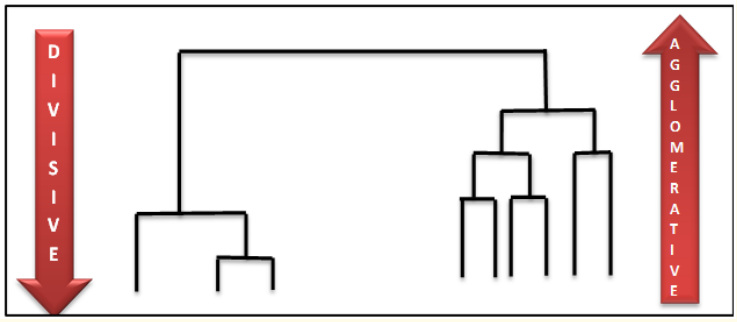

Figure 15: *Hierarchical Clustering.*

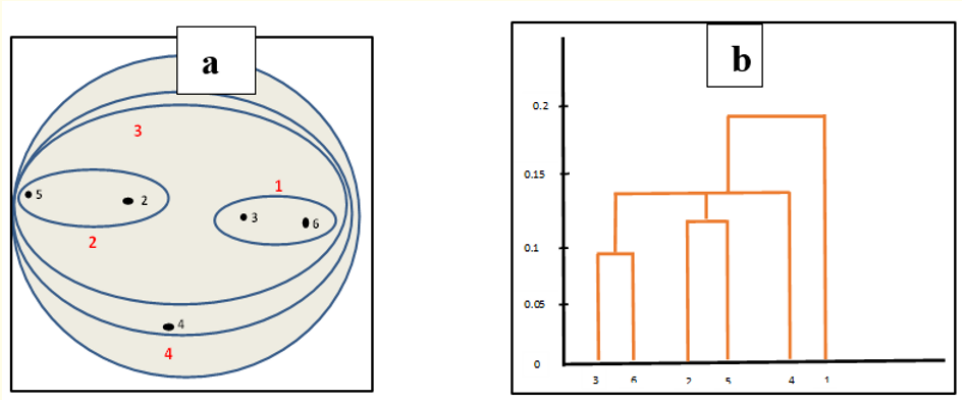

Figure 16: *A) Nested clusters, B) Dendrogram.*

Hierarchical clustering

Hierarchical clustering commonly referred as hierarchical cluster analysis, is a method of grouping similar data points into clusters. Hierarchical clustering results nests of the clusters called dendrogram (tree).

Classification of hierarchical clustering

1. **Agglomerative (bottom up) clustering**: It builds the dendrogram from the bottom level and fuses up with the most equivalent pair of clusters until all data sets are converged into a unit cluster.

2. **Divisive (top down) clustering**: This begins along with entire data points in a single cluster, the root. The root is bifurcated into a number of child clusters and subsequently each child cluster is recursively divided until only singleton clusters of corresponding data points remain, that is, each cluster has only one point [39].

Agglomerative clustering

It is more popularly applied than dividing methods. Initially, each data point generates a cluster which is also referred as a node which then nodes or clusters that are to proximity to each other. At some stage, all nodes will be members of the same cluster.

Divisive clustering

It is also known as top-down approach. It does not require to prespecify the number of clusters [40]. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.

Cluster Criteria functions

- Similarity function
- Stopping criterion
- Cluster Quality.

Similarity function/Distance measure

Distance between data points is calculate by following formulas:

Euclidean distance

The Euclidean distance is defined as the length of the line segment between two points.

$$\boldsymbol{dist}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{(\boldsymbol{x_{i1}} - \boldsymbol{x_{j1}})^2 + (\boldsymbol{x_{i2}} - \boldsymbol{x_{j2}})^2 + \cdots + (\boldsymbol{x_{ir}} - \boldsymbol{x_{jr}})^2}$$

Manhattan distance

The Manhattan distance is defined as the distance between two points measured along axes at right angles.

$$\boldsymbol{dist}(\boldsymbol{x_i}, \boldsymbol{x_j}) = |\boldsymbol{x_{i1}} - \boldsymbol{x_{j1}}| + |\boldsymbol{x_{i2}} - \boldsymbol{x_{j2}}| + \cdots + |\boldsymbol{x_{ir}} - \boldsymbol{x_{jr}}|$$

Weighted euclidean distance

The distance-weighted mean is a measure of central tendency, a special case of weighted mean, where weighting coefficient for each data point is computed as the inverse sum of distances between this data point and the other data points.

$$\boldsymbol{dist}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\boldsymbol{w_1}(\boldsymbol{x_{i1}} - \boldsymbol{x_{j1}})^2 + \boldsymbol{w_2}(\boldsymbol{x_{i2}} - \boldsymbol{x_{j2}})^2 + \ldots + \boldsymbol{w_r}(\boldsymbol{x_{ir}} - \boldsymbol{x_{jr}})^2}$$

Squared distance

The sum of squared differences in their coordinates.

$$\boldsymbol{dist}(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\boldsymbol{x_{i1}} - \boldsymbol{x_{j1}})^2 + (\boldsymbol{x_{i2}} - \boldsymbol{x_{j2}})^2 + \ldots + (\boldsymbol{x_{ir}} - \boldsymbol{x_{jr}})^2$$

Chebychev distance

It is calculated as the maximum of the absolute difference between the elements of the vectors. It is also called the maximum value distance.

$$dist(x_i, x_j) = \max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \ldots, |x_{ir} - x_{jr}|)$$

Stopping criteria

1. No re-assignment of data points to distinct clusters

2. No change in centroids

3. Minimal drop in total squared error (SSE)

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2$$

C_i = j^{th} cluster

m_j = cluster centroid

$C_{j=}$ (the mean vector of all the data points in C_j)

Dist (x, m_j) = distance between data point x and centroid m_j

Plots representing stopping criteria is shown in figure 17.

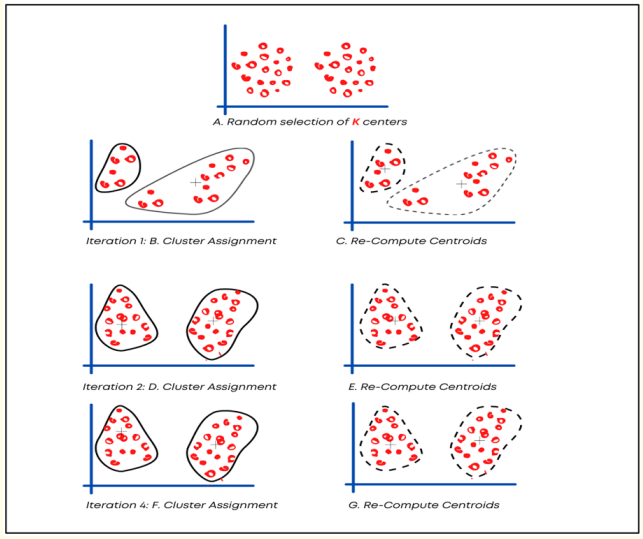

Figure 17: *Plots representing stopping criteria.*

Cluster quality

- **Intra-cluster cohesion (compactness)**

  The cohesion of a cluster is determined by how close the variables are to the cluster centroid in the cluster [41]. The sum of squared errors (SSE) is a popular metric.

- **Inter-cluster separation (isolation)**

  Segregation implies that distinct cluster centroids must be widely apart as shown in figure 18.

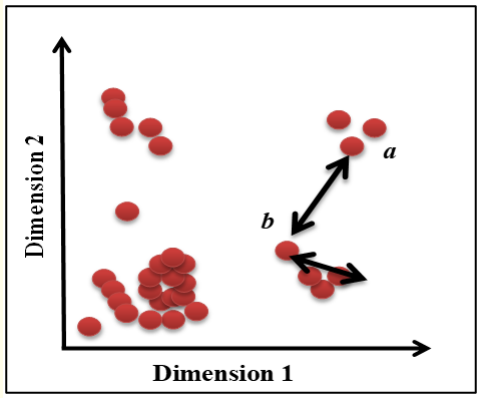

Figure 18: *Inter cluster separation.*

Principal component analysis (PCA)

PCA is a prevalent approach for minimizing the dimensionality which exists in a huge data collection. Reducing the number of components or characteristics depletes the accuracy while assembling the enormous data collection simpler, easier to examine and display. Additionally, it minimizes the complexity of computational model, allowing ML techniques to run quickly.

The concept remains unanswered to how much accuracy is sacrificed to obtain a less complicated and minimal dimension data collection. There is no definite solution for this, but we can attempt to preserve as much diversity as possible when selecting the final set of components.

There are various steps involved in the process of PCA analysis that are [42,43]:

1. Data standardization
2. Composite the covariance matrix of dimensions
3. Gain the eigen vectors and eigen values from the covariance matrix
4. Organize eigen values by arranging in descending order and select the top *k* eigen vectors that correspond to the *k* largest eigen values

5. From the selected k eigen vectors, construct the projection matrix W.
6. To get the new k-dimensional feature Y, transform the original data set X using W.

The PCA works on the basis of its algorithms. The algorithms involved in the functioning are of two types:

1. Variance and Covariance
2. Eigenvalues and Eigen factors.

The PCA can be calculate by the following equation:

$$X = t_1 p_1^T + t_2 p_2^T \ldots\ldots + t_R p_R^T + E$$

$$= TP^T + E$$

Here, X = data matrix, T = scores, P = loadings and E = residuals values

Applications of unsupervised machine learning

- Predictive models for a variety of physical and biological outcomes are being generated using unsupervised machine learning.
- It's also being utilized to create new QSAR molecular representations.
- To detect cellular phenotypes, techniques adopted from domains such as computer vision are being used for the processing of microscopic pictures.
- These methods are being used to examine data from the chemical literature and suggest new organic synthesis pathways.
- Unsupervised models can extract patterns from chemical databases and utilize that information to create new molecules from scratch.

Advantages of unsupervised learning

Unsupervised learning has various advantages that are listed below [44]:

1. No requirement of Preceding knowledge of the image area.
2. Chances of human error is minimal.
3. It affords in unique spectral classes.
4. Fairly easy and consume less time to carry out.

Disadvantages of unsupervised learning

1. The exact information cannot be obtained on data categorizing and the result as data utilized in unsupervised learning is labelled and unknown.

2. The findings are less accurate since the input data is unknown and has not been tagged in advance. This implies that the machine must perform this task on its own.
3. There is no correlation of informative classes with spectral classes.
4. The analyst must devote effort in understanding and labelling the classes that fall within that categorization.
5. Class spectral qualities can also vary over time [45].

Conclusion

Drug research and development is a time-consuming, difficult and expensive process. The goal of drug discovery is to uncover novel molecules with certain chemical characteristics that can be used to cure disorders. With the accessibility of various machine learning techniques, the strategy utilized has become a significant component in computer programming in recent years. Machine Learning techniques can be supervised or unsupervised learning. If you have a little quantity of data that is clearly labelled for training, Supervised Learning is the way to go. For huge data sets, unsupervised learning would provide greater performance and results. This study examines a number of different machine learning algorithms. Today, everyone, intentionally or unintentionally, employs machine learning. This paper gives a quick overview of the most popular machine learning algorithms. Future predictions based on Machine Learning are becoming highly significant in the lead-up to preclinical investigations. In the development of novel pharmaceuticals, this step manages to significantly cut expenses and research time durations. The most recent advancements in the construction of new algorithms in the field of Artificial Intelligence have made it possible to tackle issues in various fields. The pharmaceutical sector has profited enormously from the usage of these models in cheminformatics, and more especially in drug development. As a result, this feature must be heavily changed in order to make conclusive findings. However, in the context of precision medicine and drug development, the opportunities and benefits given by machine learning approaches are enormous. This work contributed to the field by discussing various applications of supervised and unsupervised machine learning algorithms for the identification and development of innovative drug candidates. The employment of machine learning techniques in conjunction with other approaches is growing more widespread by the day.

Acknowledgements

The corresponding author RGK expresses gratitude to the Principal, Poona College of Pharmacy, Pune for his constant encouragement. #Both the authors contributed equally.

Conflict of Interest

Author declares no conflict of interest.

Bibliography

1. Poola I. "The Best of the Machine Learning Algorithms Used in Artificial Intelligence". *International Journal of Advanced Research in Computer and Communication Engineering* 6.10 (2017).
2. Siraj-Ud-Doulah M. "Application of machine learning algorithms in bioinformatics". *Bioinformatics and Proteomics* 3.1 (2019).
3. Schneider P., *et al*. "Rethinking drug design in the artificial intelligence era". *Nature Reviews Drug Discovery* 19.5 (2020): 353-364.
4. Patel L., *et al*. "Machine learning methods in drug discovery". *Molecules* 25.22 (2020): 5277.

5. Carracedo-Reboredo P., *et al*. "A review on machine learning approaches and trends in drug discovery". *Computational and Structural Biotechnology Journal* 19 (2021): 4538-4558.

6. Sarker IH. "Machine learning: Algorithms, real-world applications and research directions". *SN Computer Science* 2.3 (2021): 1-21.

7. Monaco A., *et al*. "A primer on machine learning techniques for genomic applications". *Computational and Structural Biotechnology Journal* 19 (2021): 4345-4359.

8. Uddin S., *et al*. "Comparing different supervised machine learning algorithms for disease prediction". *BMC Medical Informatics and Decision Making* 19.1 (2019): 1-16.

9. Ben-Hur A., *et al*. "Support vector machines and kernels for computational biology". *PLoS Computational Biology* 4.10 (2008): 1000173.

10. Navada A., *et al*. "Overview of use of decision tree algorithms in machine learning". 2011 IEEE control and system graduate research colloquium. IEEE (2011).

11. Patel H., *et al*. "International Journal of Computer Sciences and Engineering Open Access". *International Journal of Science and Engineering* 6.10 (2018).

12. Charbuty B., *et al*. "Classification based on decision tree algorithm for machine learning". *Journal of Applied Science and Technology Trends* 2.01 (2021): 20-28.

13. Salmi N., *et al*. "Naïve Bayes classifier models for predicting the colon cancer". *IOP Conference Series: Materials Science and Engineering*. 546.5 (2019).

14. Rish I. "An empirical study of the naive Bayes classifier". *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* 3.22 (2001).

15. Kumari K., *et al*. "Linear regression analysis study". *Journal of the Practice of Cardiovascular Sciences* 4.1 (2018): 33.

16. Wang L. "Research and implementation of machine learning classifier based on knn". *IOP Conference Series: Materials Science and Engineering* 677.5 (2019).

17. Belgiu M., *et al*. "Random forest in remote sensing: A review of applications and future directions". *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (2016): 24-31.

18. Boulesteix AL., *et al*. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6 (2012): 493-507.

19. Kotsiantis S., *et al*. "Combining bagging and boosting". *International Journal of Computational Intelligence* 1.4 (2004): 324-333.

20. Kulkarni RG., *et al*. "Artificial Intelligence: Drug Discovery and Development Prospective in Medicinal Chemistry". *EC Pharmacology and Toxicology* 9.11 (2021): 87-92.

21. Maulud D., *et al*. "A review on linear regression comprehensive in machine learning". *Journal of Applied Science and Technology Trends 1*.4 (2020): 140-147.

22. Yao D., *et al*. "A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines". *Journal of Computers* 8.1 (2013): 170-177.

23. Uyanık GK., *et al*. "A study on multiple linear regression analysis". *Procedia-Social and Behavioral Sciences* 106 (2013): 234-240.

24. Jason B. "Master Machine Learning Algorithms, Discover How They Work and Implement Them From Scratch", (2017).

25. Schneider A. "Linear regression analysis: part 14 of a series on evaluation of scientific publications". *Deutsches Ärzteblatt International* 107.44 (2010): 776.

26. Peng CYJ., *et al*. "An introduction to logistic regression analysis and reporting". *The Journal of Educational Research* 96.1 (2002): 3-14.

27. Tripepi G., *et al*. "Linear and logistic regression analysis". *Kidney International 73*.7 (2008): 806-810.

28. Sahebjamee H., *et al*. "Quantitative structure-activity relationships study of carbonic anhydrase inhibitors using logistic regression model". *Iranian Journal of Chemistry and Chemical Engineering* 32.2 (2013): 19-29.

29. Siraj UD., *et al*. "Performance evaluation of machine learning algorithms in ecological dataset". (2019).

30. Wu H., *et al*. "Review on Evaluation Criteria of Machine Learning Based on Big Data". *Journal of Physics: Conference Series* 1486.5 (2020).

31. Vamathevan J., *et al*. "Applications of machine learning in drug discovery and development". *Nature Reviews Drug Discovery* 18.6 (2019): 463-477.

32. JIAO., *et al*. "Review of typical machine learning platforms for big data". *Journal of Computer Applications* 37.11 (2017): 11.

33. Gupta R., *et al*. "Artificial intelligence to deep learning: Machine intelligence approach for drug discovery". *Molecular Diversity* 25.3 (2021): 1-46.

34. Réda C., *et al*. "Machine learning applications in drug development". *Computational and Structural Biotechnology Journal* 18 (2020): 241-252.

35. Rodrigues T., *et al*. "Machine learning for target discovery in drug development". *Current Opinion in Chemical Biology* 56 (2020): 16-22.

36. Carpenter KA., *et al*. "Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: a review". *Current Pharmaceutical Design* 24.28 (2018): 3347-3358.

37. Rajula HSR., *et al*. "Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment". *Medicina* 56.9 (2020): 455.

38. Mak KK., *et al*. "Artificial intelligence in drug development: present status and future prospects". *Drug Discovery Today* 24.3 (2019): 773-780.

39. Gaudelet T., *et al*. "Utilizing graph machine learning within drug discovery and development". *Briefings in Bioinformatics* 22.6 (2021): bbab159.

40. Barlow HB. "Unsupervised learning". *Neural Computation* 1.3 (1989): 295-311.

41. Ghahramani Z. "Unsupervised learning". *Summer School on Machine Learning. Springer, Berlin, Heidelberg,* (2003).

42. Hastie T., *et al* "Unsupervised learning". The elements of statistical learning. Springer, New York, NY, (2009). 485-585.

43. Oliver JJ., *et al*. "Unsupervised learning using MML". *ICML* (1996).

44. Berg-Kirkpatrick T., *et al*. "Painless unsupervised learning with features". *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010).

45. Figueiredo MAT., *et al*. "Unsupervised learning of finite mixture models". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.3 (2002): 381-396.

Volume 10 Issue 9 September 2022