Journal of Management and Service Science, 2021,
Vol. 01, Iss. 01, S. No. 003, pp. 1-17

A Comprehensive Survey on Machine Learning

Astha Singh[1], Pawan Singh[2], Anil Kr. Tiwari[3]

Amity School of Engineering and technology, Amity University, Lucknow, India[1,2,3]
singhastha2809@gmail.com[1], pawansingh51279@gmail.com[2]

How to cite this paper: A. Singh, P. Singh and Anil Kr. Tiwari (2021) A Comprehensive Survey on Machine Learning. *Journal of Management and Service Science,* 1(1), 3, pp. 1-17.

https://doi.org/10.54060/JMSS/001.01.003

Received: 25/02/2021
Accepted: 07/03/2021
Published: 08/03/2021

Abstract

The objective of this briefing is to present an overview of the topic, machine learning techniques currently in use or in consideration at statistical agencies worldwide. It is important to know the main reason why real-world scenarios should start exploring the use of machine learning techniques, terminology, approach and about few popular libraries in python, what regression is, by completely throwing light on simple as well as multiple linear and non-linear regression models and their applications, classification techniques, various clustering techniques. The material presented in this paper is the result of a study based on different models and the study of various datasets (analysis and choice of the correct model are important). While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalization to get a system that performs well on yet-unseen data instances. Topics like regression, classification, and clustering, the report covers the insight of various techniques and their applications.

Keywords

Machine Learning, Regression, Classification, Clustering

1. Introduction

1.1. Machine Learning

Machine learning (ML) is utilized in almost every field nowadays like the data scientists forecast if any human body cell which is at a risk to develop cancer is either benign or malignant by using this technology. We also see that this technology plays a vital role in determining health status and wellbeing. Thus, working in support of doctors by helping them in decision making. Decision Trees if made with accuracy and precision from the historical data can help doctors prescribe proper medication to their patients [5]. In banking systems, this technology is used to do bank customer segmentation, approval of loan applications, etc. quite easily. Ever thought of recommendations on the sites like YouTube Amazon or Netflix, this also uses machine learning algorithms to provide certain product or service recommendations that the customer might find interesting to purchase or utilize. An infinite list of examples can be quoted. There is so much that can be very well done by this technology may it be

telecommunication or automobile industry to predict customer churn, all can be done using the available libraries like scikit-learn of python with an ultimate ease [1].

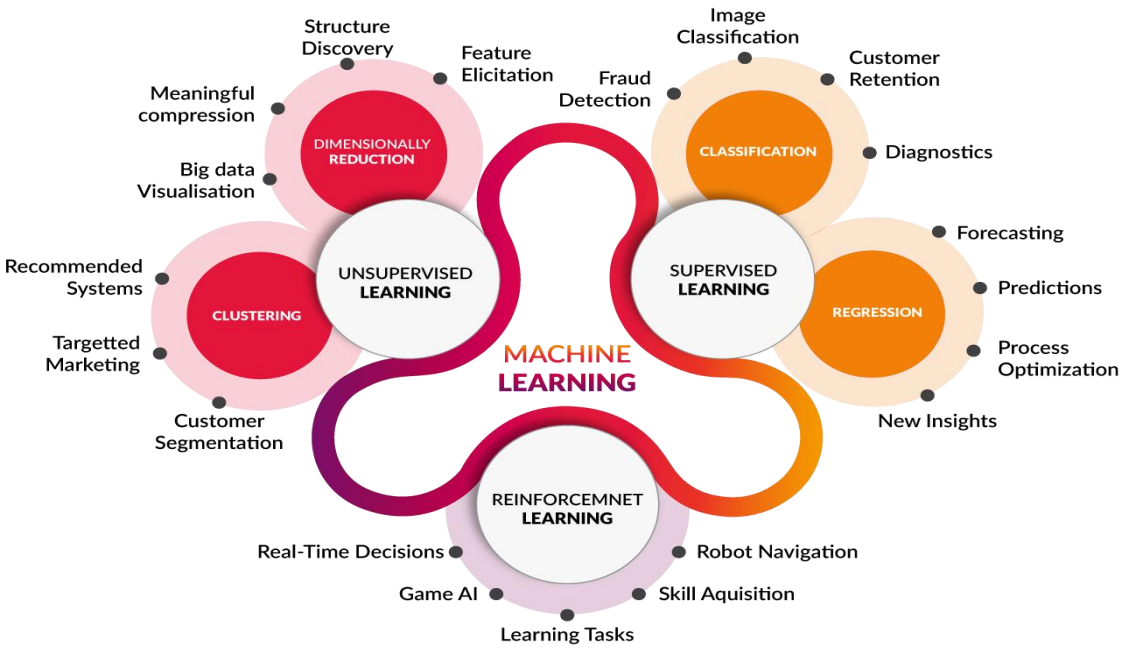


Figure 1. Concept of Machine Learning

1.2. Importance of Machine Learning

ML is defined as a discipline of computer science that gives "computers the learning ability without being explicitly taught," according to a precise definition.

Let's see the meaning of "without being explicitly taught." Presume that we get a databank that consists of the images of cats and dogs, also we have a product or a software that can recognize and differentiating between the two classes [4]. First and the foremost thing would be to interpret these images as a set of features set for the animals. E.g., are the animal's eyes visible? If yes, what is its size? What about its ears? And tail? Legs? Oh! Also does it have wings? What size? If there wasn't the existence of this technology, then every snapshot would really be converted into a feature vector. Then, we had to make some rules and methods in all traditional way to make the computers learn. However, it was a flop. Why?

It required a large number of rules that were very consistent with the current database (i.e. restrictive) and were not generic enough to recognize new databases, instances, or test cases. ML is powerful in this area. This technology helps in building a mod-el by including all the feature sets (to distinguish between objects) i.e. flexibility, and their corresponding type of objects, and it recognizes the pattern of each object on its own by iterating through the dataset available. It recognizes things without getting any programs explicitly. In short, in machine learning the approach is like a child's learning approach. So, machine learning algorithms, are inspired by the process of human learning, iteratively learning from the provided data, and thus, allowing computers to find hidden insights by recognizing patterns.

1.3. Machine Learning Terminology and Approach

To work with any dataset for predictions and gaining information we must know the way we work. Basically, the terminology

we will use is very popular and is universal [1-2].

Figure 2. Terminology used in Machine Learning

Now, after one knows the terminology we use, he must be aware of the approach to follow [1]. The approach is simple, you need to know your dataset and do the following analysis:

- Understand the problem/dataset
- Extract the features from your dataset
- Identify the problem type
- Do pre-processing of your dataset:
    - Imputer: used to impute NaN i.e., missing values of the dataset.
    - Label Encoder: handles categorical data.
    - Dummy Variable/ OneHotEncoder: the categorical data is defined as variables with a finite set of label values.
    - Standard Scaler: to fix the outlier or different scales (if any) in the dataset.
    - Splitting of dataset into train and test data: split arrays or matrices into random train and test subsets.
- Choose the right model (there are several models in machine learning according to your information extraction requirements).
- Train and test the model
- Strive for its accuracy

And that's all one need. Your prediction model is ready, you can go ahead give your model inputs and it will give you predictions/information as per your requirement.

1.4. Few Popular Techniques

Regression/Estimation technique is used to predict label with continuous values. E. g. Estimating house price v/s its characteristics or to estimation of CO_2 emission v/s car's engine. Classification technique is used to find a class to which an object belongs to depending on the provided features, for example, if a human cell is benign or malignant, etc. Clustering means grouping similar types/cases, for example, can find similar patients/customers/areas by knowing the commons in each. Association technique is used for finding items/events that often co-occur i. e. occurring together, for example, applying association rule for market basket analysis on grocery items that are usually bought together by a particular customer. Anomaly detection is used to detect abnormalities by finding abnormal/unusual cases, for example, used for credit card fraud detection. Sequence mining

is used to predict the next event. Recommendation systems, this connect or relate people's preferences with those whose choices match, thus, grouping things, recommending new items, such as books or movies to the customers [1-3].

1.5. Some Important Python Libraries

There are number of libraries used for the numerical computations such as NumPy, Pandas, SciPy. NumPy is a math library allows working with images, video, sound waves in array of n-dimensions. More efficient because of its easy usage making complex mathematical operations easy. Pandas is a high-level Python library providing high-level data structures manipulation and tools for analysis. It offers grouping, filtering, combining information with time-based functionality. SciPy is a collection of numerical algorithms, including signal processing, optimization, statistics, etc. [6-7].

For data visualization the separate libraries are provided in python such as Matplotlib, Seaborn. Matplotlib is a renowned plotting library that provides different dimensional plots like 2-D and 3-D. Seaborn library is also used for interactive plots. SciKit Learn is a specific library used for machine learning. It is provided for the purpose of:

- Free software machine learning library.
- Classification, Regression and Clustering Algorithms.
- Works with NumPy and SciPy.
- Great documentation.
- Easy to implement

2. Regression

2.1. Introduction to Regression

The dataset shown here is related to the CO_2 emissions from different cars.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Figure. 3. Dataset of Regression

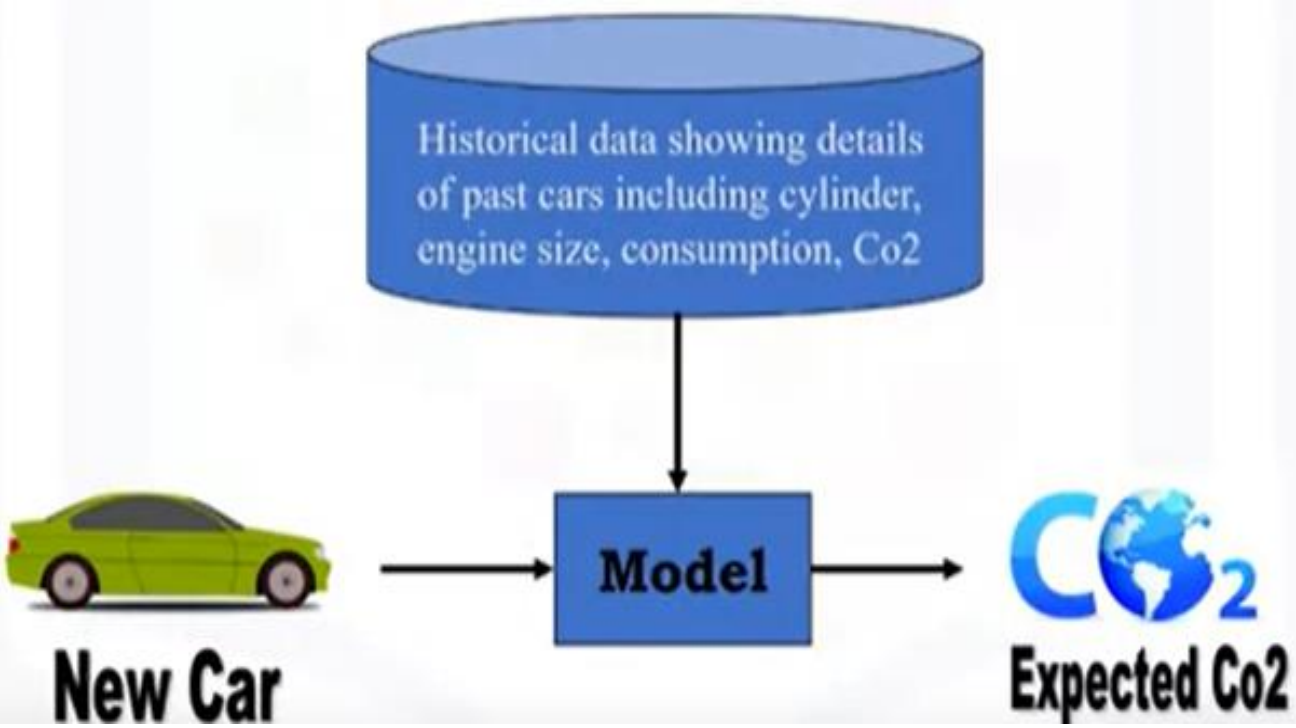


Figure. 4. Regression Model

Can we predict the CO_2 emissions for record index = 9? Yes, we can do that using regression which is used to predict continuous value. The two variables used are addressed as a dependent variable(Y) and one or more independent variables(X). Y is the target variable. Here X can be taken as engine size, cylinders, fuel Consumption Comb and Y is CO2emissions. Y is continuous value while X can be continuous or discrete or even categorical [8-10]. Thus, a linear regression is thereby an approximate linear mod-el describing the dependency between two or more variables.

Types of Regression Models:

Simple Regression: here X is a single variable.
Simple Linear Regression
Simple Non-linear Regression
Multiple Regression: here X can be more than one variable.
Multiple Linear Regression
Multiple Non-linear Regression

Applications of regression: Sales forecasting, Satisfaction Analysis, Price Estimation, Employment Income and many more

2.2. Simple VS Multiple Linear Regression

Simple Linear Regression

X→EngineSize,Y→CO_2Emissions

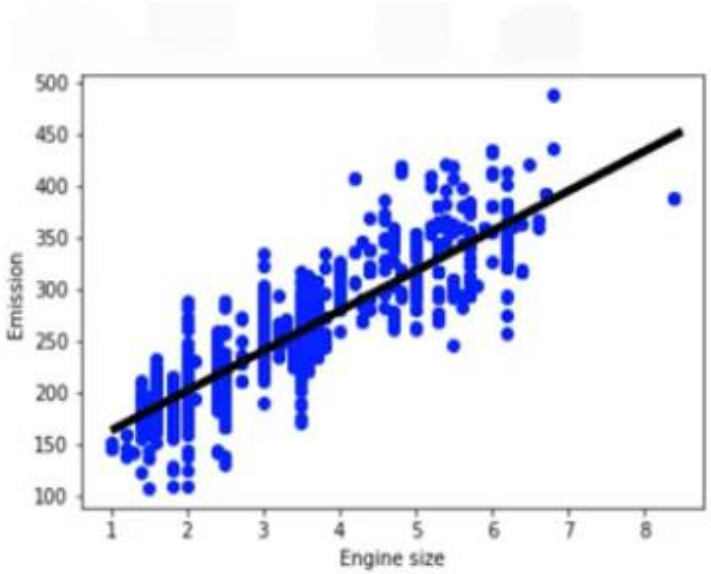


Figure. 5. Line of Regression

Multiple Linear Regression

X→Engine Size, Cylinders, FuelCnsumptionComb; Y→CO2Emissions

For simple linear regression by building a Regression Model and making the above plot using Scikit-Learn library. Now let us assume a best fitting line, thus, let's predict the emission for a given car engine, eg. If engine size=2.4 then emission=214 from the graph. This model follows the equation: (bo is the y-intercept and b1, b2, is the slope)

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Constant Coefficients

The distance from the data point from the line of regression is called the residual error, the mean of all the distances shows in-accurately the line fits with the whole dataset. This error is shown by the equation of MSE i.e. mean square error.

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

We do not require to learn these formulae as everything is already available in the python libraries. Pros of linear regression is fast, no parameter tuning and easy to implement

2.3. Model Evaluation

Model evaluation approach can be either train or test on the same dataset or Train/Test split. Also, there are some metric that can be utilized for model evaluation [12]. We can write the error of the model as the average difference between the actual and the predicted values for all the rows.

$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

What do you mean by training accuracy and out-of-sample accuracy?

- Training Accuracy: is the percentage of correct predictions that the model makes when using the test dataset for evaluation. A high training accuracy isn't always appreciable because this may mean that the model is over-trained and thus, result in over-fitting of the data. Over-fit? Yes, over-fitting means the model is overly trained to the provided dataset, which may capture noise i.e. outliers and thereby produce a non-generalized model.
- Out-of-Sample Accuracy: is the percentage of the correct predictions that the model makes on the data on which it has not been trained on. It is important for a model to have high out-of-sample accuracy. By using train/test split we can get achieve this

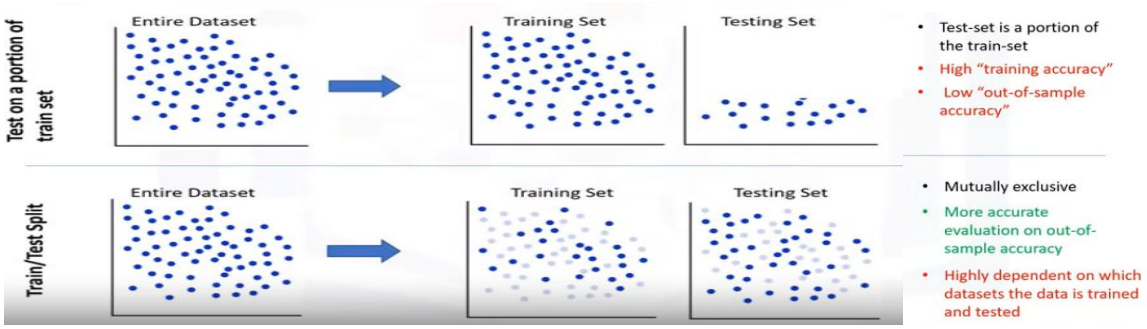


Figure. 6. Plots for Model Evaluation

Next technique to evaluate the model, K-fold-cross-validation. If we have K equal 4 folds, this means the dataset is split into 4 parts, in this each 25% of the data is split and used as test data while the left over 75% is set as training data. Then the accuracy of the model is calculated by taking average of all 4 evaluations.

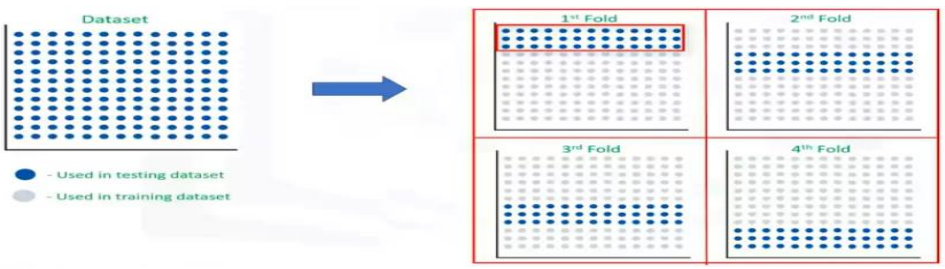
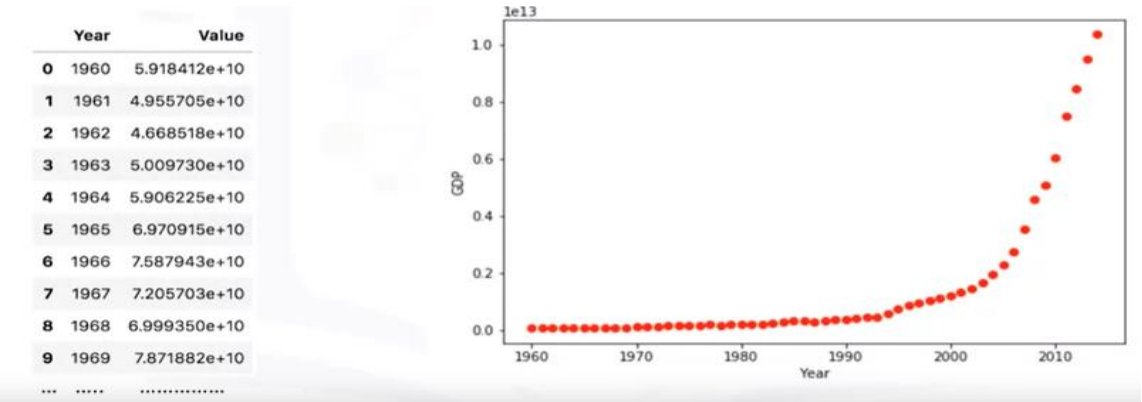


Figure. 7. K-fold Cross-validation

2.4. Non-Linear Regression

This is a dataset showing GDP values through years. We see that the plot obtained isn't a linear one but a curve.



	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	.....

Figure. 8. Non-Linear Regression Curve

It seems either logistic or an exponential function. We may fit quadratic or cubic regression lines here i.e. a polynomial function. Thus, an nth degree polynomial curve may fit this data precisely.

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

It is important to note that increasing the degree to a large extent may overfit the model which is not our requirement. We require recording patterns and not noise.

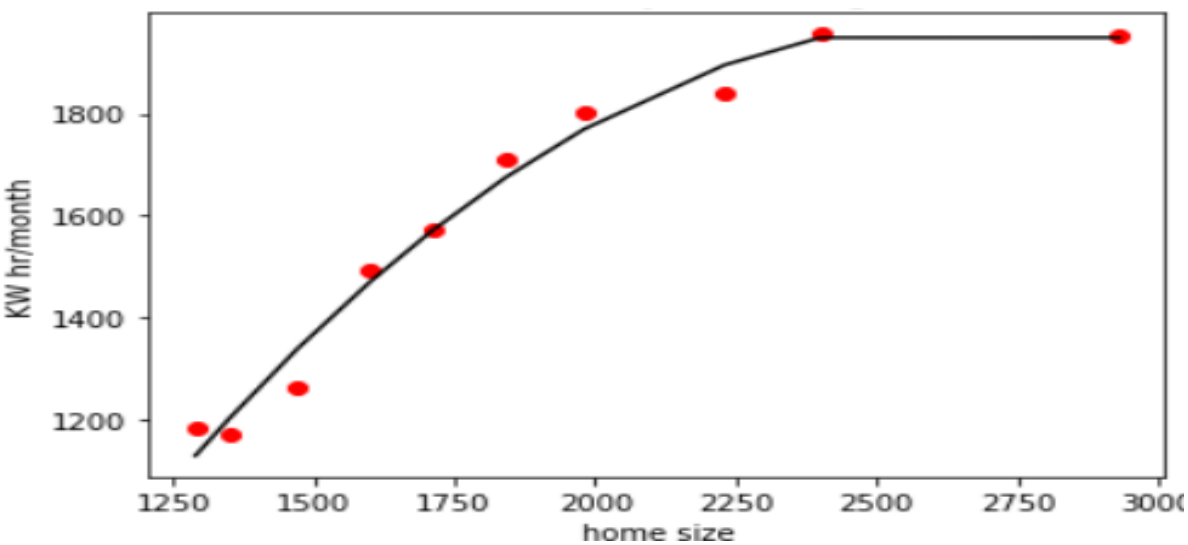


Figure 9. Polynomial linear regression curve, Degree = 3

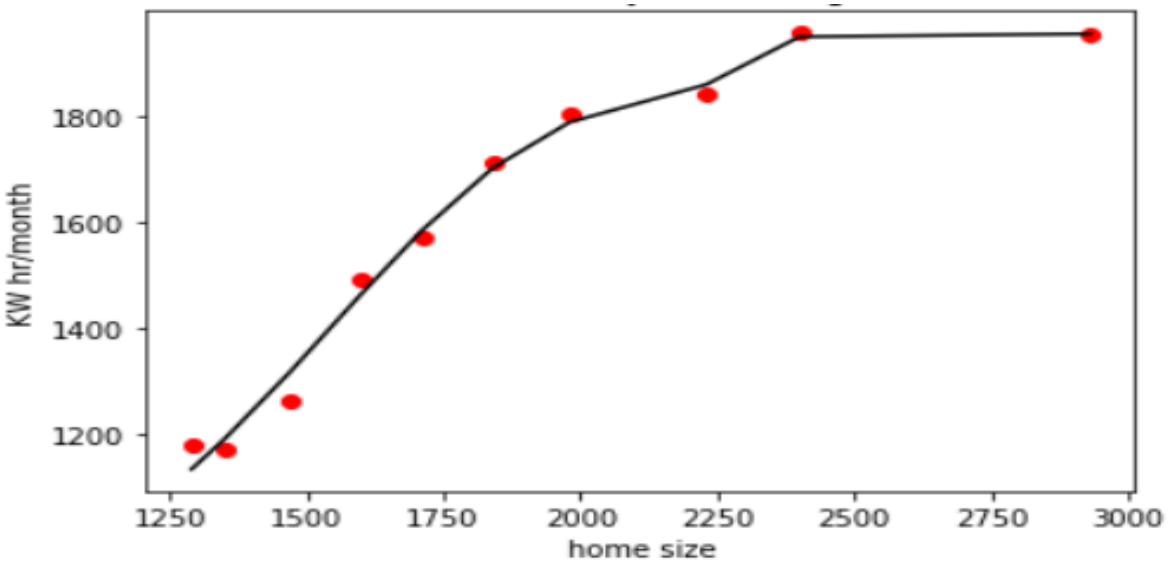


Figure 10. Overfitting, Degree = 10

3. Classification

3.1. Introduction to Classification

Classification falls under supervised machine learning approach. Classification means categorizing some unknown items into sets of categories or classes. The target variable in this method is a category which can be Label Encoded into discrete categorical values. The dataset below shows a bank customer data who had taken a loan [13].

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

Categorical Variable

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
37	2	16	10	130	9.3	10.23	3.21	

Figure 11. Dataset for classification

The banks concern is which of the upcoming customers applying for loan can turn to be a risk. Bankers can review the historical data and identify the risky customers and decline their applications using a classifier. We may have more than two classes. Here in the example there are two classes: default (Yes=1 and No=0). Customer segmentation is the most renowned example of classification. Other applications are email filtering, handwriting recognition, object classifiers, etc. Many

algorithms can be used to make a classifier including naïve bayes, k-nearest neighbor, decision tree, neural networks, logistic regression, support vector machine, etc.

3.2. K-Nearest Neighbors

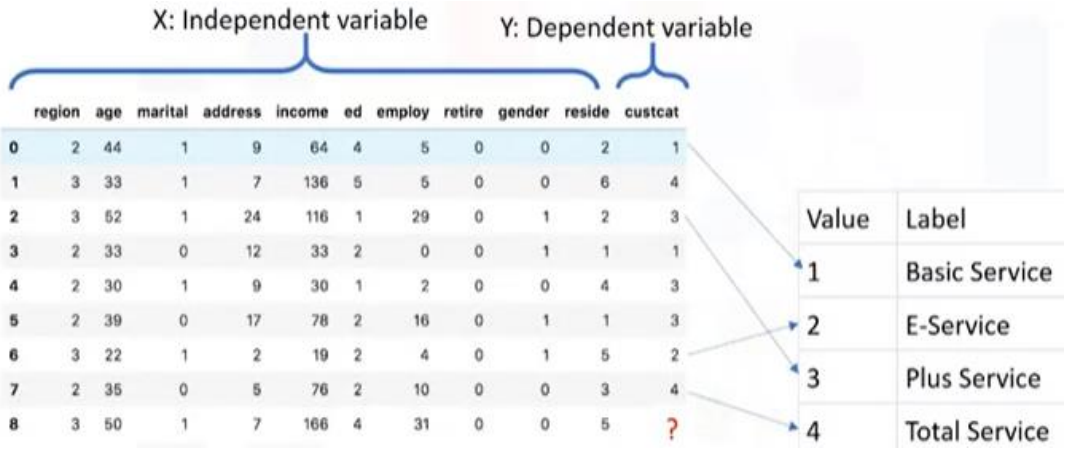


Figure12. Dataset features

The above given dataset corresponds to a company who has to provide customer services to its new customers based on the demographic data present. There are 4 groups. We need a classifier to classify the new customers. Let's build KNN model using two features Age and Income and plotting a graph accordingly. We see we have a new customer and we need to classify him. We use the technique of finding the 5 nearest neighbors to the new customer and 3/5 are "Plus Service" so it is quite feasible to put this customer in the "Plus Service" class. Thus, the value of K in K-Nearest Neighbors here becomes 5. A lower value of K(say K=1) if chosen by the user may cause over-fitting of the model. Also choosing a very high value of K(say K=20) then the model becomes overly generalized [8].

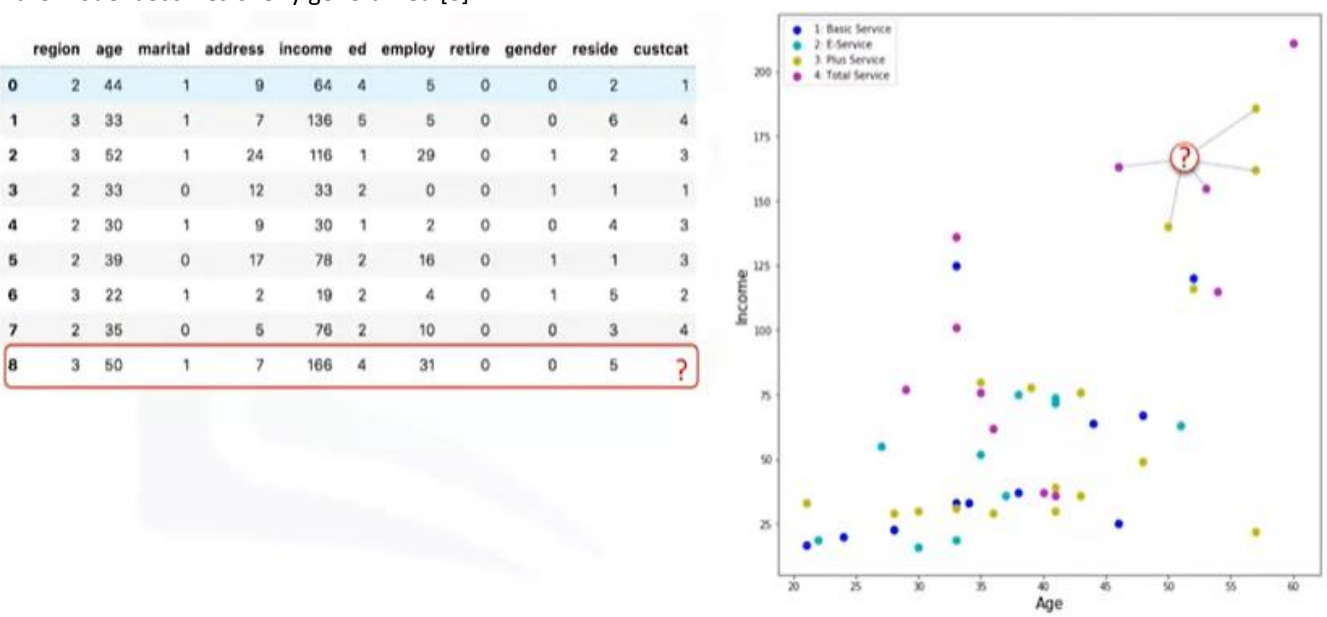


Figure 13. Plot for K-nearest neighbor algorithm

Thus, this algorithm classifies cases based on the classes to which the nearest cases belong to with majority votes. Thus, distance between two plots plays a major role. Farther the points the more dissimilar they are. This distance is called Euclidean distance. $\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$

3.3. Decision Tree

Suppose we have the following dataset where we need to decide using a decision tree that which drug A or B was effective according to the previously available data so that we can prescribe the correct drug in future. We can create the following decision tree [3, 14].

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

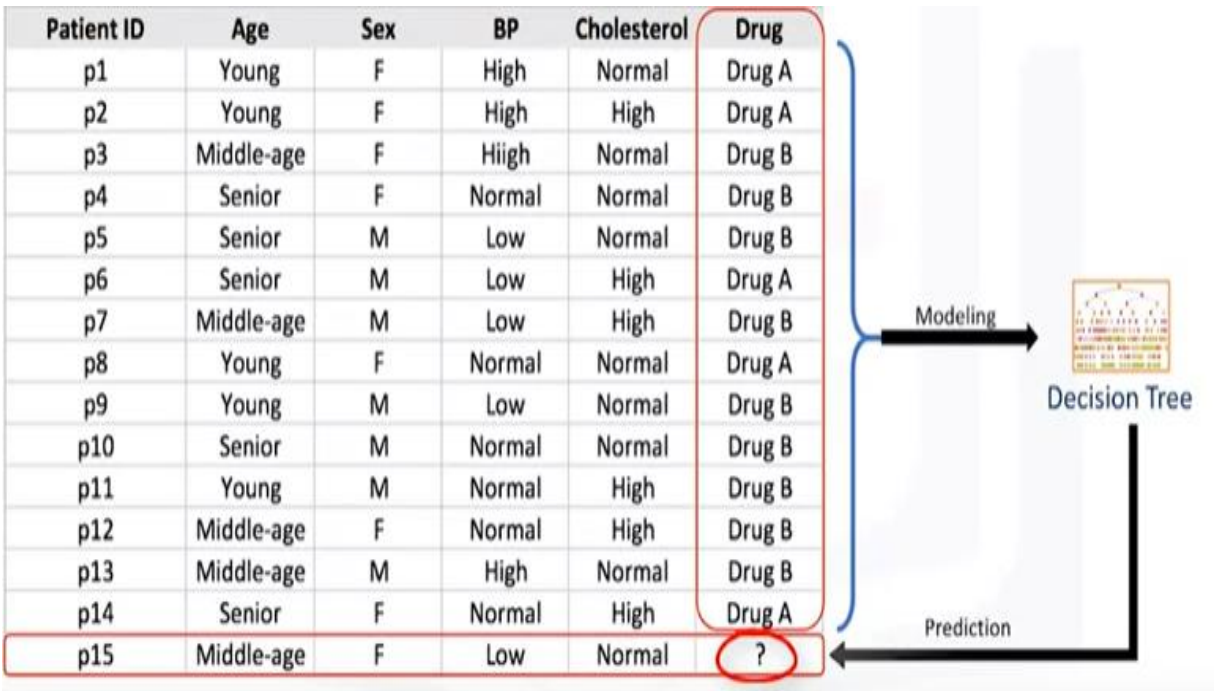


Figure 14. Dataset for decision tree

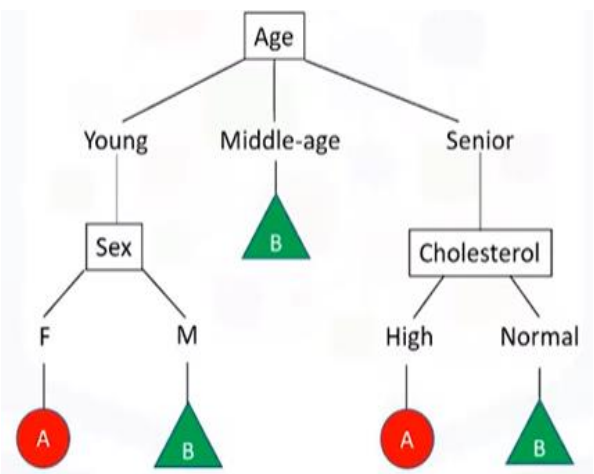


Figure 15. Decision Tree

Now we can easily prescribe a drug to patient p15, drug B will be the most suitable choice. We can note that a decision tree is about testing an attribute and thereby branching the cases on the result of the test done. Thus, an internal node means a test, a branch means a result and a leaf means a class. Decision Tree Learning Algorithm:

1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go to Step 1 and repeat it for rest of the attributes.

3.4. Logistic Regression

This algorithm is a classification algorithm for categorical values. Logistic regression works with binary as well as multiple classes in target variable. In this method ***independent variable must be continuous***, if categorical it must be dummy coded. Let us take the dataset of a telecom industry and analyze the customers who will leave next month.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

Figure 16. Dataset for Logistic Regression

The four major scenarios where logistic regression is a best fit are:

- If data if binary: yes/no, 0/1, True/False
- If we require probabilistic results: logistic regression returns value [0,1] for a given data.
- If we know that the data is linearly separable: need of linear decision boundary.
- If we need to understand the impact of a feature.

To perform logistic regression, we use the sigmoid function instead of the coefficients that are there in the linear regression model.

Sigmoid function output is always [0,1]

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

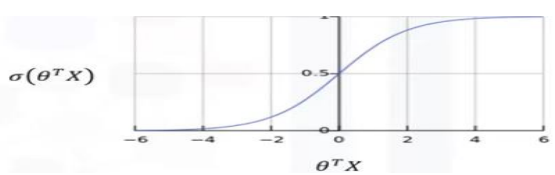


Figure 17. Sigmoid function curve

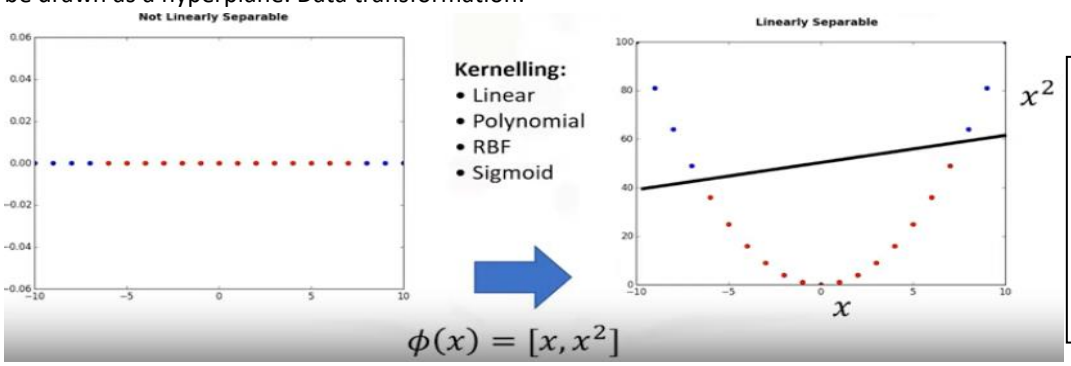
3.5. Support Vector Machine

This method is also used for classification by finding a separator. Let's see the following dataset for human cell classification as benign or malignant for the new sample of human cells [8].

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Figure 18. Dataset for Support Vector Machine

Firstly, mapping the data to a high dimensional feature space like 3-D is done. Then, a separator is found out that could be drawn as a hyperplane. Data transformation:



Kernelling means mapping data into higher dimensional space such that there is a way to get linearly separable dataset from the inseparable one.

Figure 19. Hyperplane

The goal now is to choose a hyperplane with the biggest margin possible after kernelling. Once we have achieved this hyperplane we can classify new points telling whether it lies above or below the line.

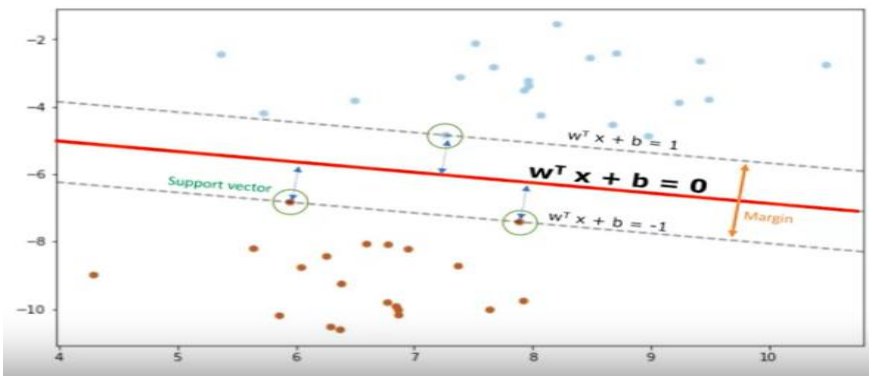


Figure 20. Identify your hyperplane

Advantages of SVM: Accurate, memory efficient. Disadvantages of SVM: it is prone to over-fitting, no probability estimations, ineffiecient computations for large datasets. SVM Applications: Image Recognition, Text Catorizing, detecting spam, sentiment analysis, etc.

4 Clustering

4.1. Introduction to Clustering

Having the given dataset for customer segmentation to find potential customers. Clustering can be used which is an unsupervised learning approach. Clustering refers to finding clusters in the dataset.

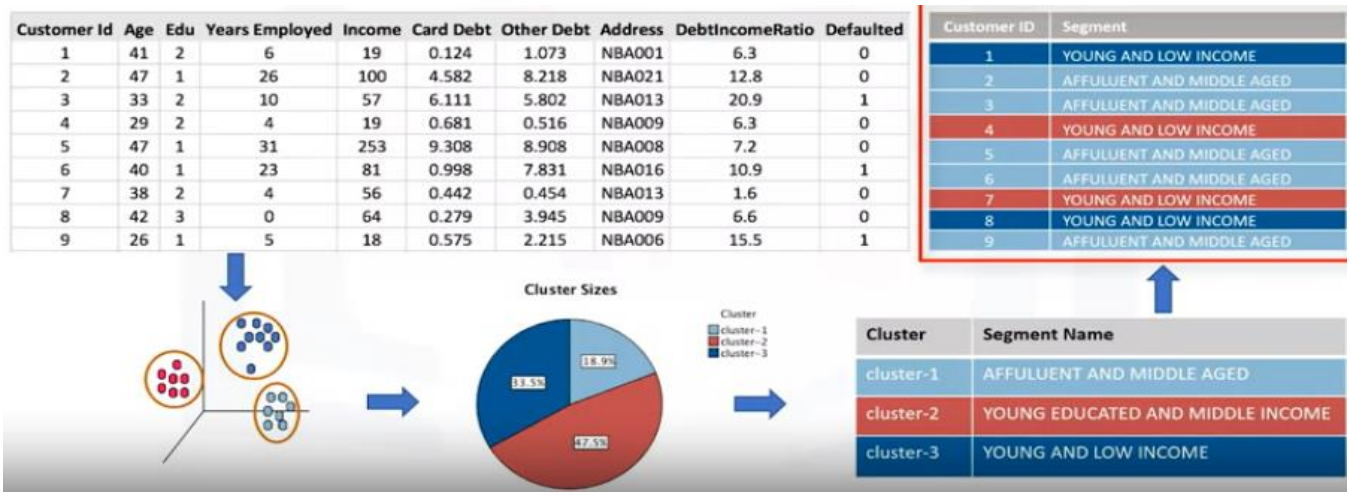


Figure 21. Dataset for clustering

A cluster can be seen as a group of similar data points. Classification is supervised while clustering is unsupervised. Thus, clustering can be used to find association among objects. Applications of clustering: Customer Segmentation, Recommendation Systems, Fraud detection, Customer insurance risk, Auto-categorizing news articles, Characterize patient behavior, etc. Where can we use clustering? Exploratory data analysis, Summary generation, Outlier detection, Pre-processing steps, Duplicate extraction. Clustering Algorithms: Partitioned based clustering(efficient): K-means, Hierarchical clustering (produces trees of clusters): Agglomerative, DBSCAN (for arbitrary shaped clusters)

4.2. Introduction to K-means

This algorithm aims at dividing the data into k non-overlapping datasets/subsets without any cluster-internal structure. Objects in same clusters are very similar while others are very dissimilar. In this method, we use dissimilarity metrics. K-means tries to minimize the intra cluster (within same cluster) distances while maximize the inter cluster (between different clusters) distances. We use Mankowski or Euclidean distance to find dissimilarities [4-5].

First step is to determine the number of clusters (k) to be formed. It randomly picks up any points say k=3 therefore 3 data points called centroids. We need to find the distances between these centroids and the other data points. We now create distance matrix. Nearest data points to the respective centroids are grouped as clusters. Here forming 3 clusters. The error (avg. of distances from centroid to their data points) is too much because the centroids were chosen randomly. We move centroids to minimize error. This is an iterative process of computing new centroids and again checking for dissimilarities and distances. We repeat this process until there are no more movements in centroids.

4.3. Introduction to Hierarchical Clustering

These algorithms build a hierarchy of clusters, here each node represents a cluster which consist of the clusters of daughter nodes. Two approaches exist: divisive (top down) and agglomerative (bottom up). Agglomerative is a more popular approach. In this each data point is a cluster itself and according to the closest distance between the clusters, the clusters are grouped together [8]. We repeat the process until we built one tree of clusters often visually recognized using dendrogram (all clusters clustered as a single cluster).

This type of clustering doesn't require a pre-specified number of clusters. Thus, in agglomerative clustering, we merge the clusters with the nearest distances between clusters. We may use Euclidean distance to calculate distances between two clusters. This approach may take long to compute [5].

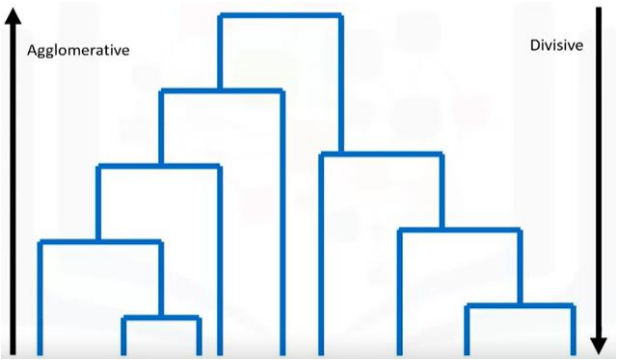


Figure 22. Hierarchical Clustering

4.4. DBSCAN

When the cases appear where we have arbitrary shaped clusters or clusters within clusters traditional clustering approach fails and here, we have to use clustering technique called DBSCAN or Density Based Spatial Clustering of Applications with Noise. DBSCAN clusters regions of high density and separates outliers which cannot be done by k-means algorithm [4].

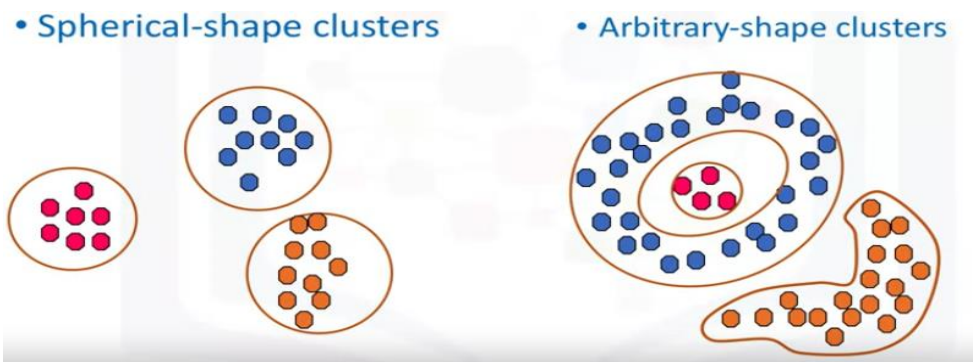


Figure 23. Clusters in DBSCAN

DBSCAN works on two parameters: Radius of neighborhood (R) and Minimum number of neighbors (M). How DBSCAN works? Let's define R=2 units and M=6. Each point in a given dataset can be either a core (this is a point with at least M data points within the radius of the neighborhood), border (this is a point with less than M data points within the radius of the

neighborhood or it is reachable from some core point) or an outlier point(neither a core nor a border point). This way we label each point as core, border or outlier. The next step is to close all the core points within the neighborhood as a cluster. Thus, a cluster is formed by closing at least one core point and all reachable points from those core points as well as the borders, thus, filtering out all the outliers. It is robust to the existing outliers. Unlike K-means clustering, we do not need to specify the number of clusters in the beginning [6-8].

Conclusion

There are several researches still going on for this field to make a significant difference in getting the insights of data to improve and benefit the industries at large. It is worth pointing out that machine learning algorithms always involves the use of historical data in order to understand the relationship between two or more variables. Some observers refer to this as the issue of the lag between the past and the present and the future. Nonetheless, machine learning algorithms are popular for real world problems. Although many users might find the mathematics involved quite difficult, the techniques are itself relatively easy to use, especially when a model or template has previously been developed. However, users who do not understand the underlying mathematics should obtain some assistance in the interpretation of the results. Actually the important part is just that one needs to know the dataset and apply the correct model according to the total data present By training the dataset over and over on a large data accuracy can be increased and thus the regression model is completed by using certain libraries and concepts It is a tender module which delivers a display using which operators could cooperate in command to perform approximately, like business the itinerant, snap a picture, guide the email, or view a record. Individually movement is prearranged a window in which to draw its manipulator boundary. The window naturally plugs the shelter. Whenever an original activity starts, it is pushed onto the vertebral heap and takes user focus. The back stack abides to the basic “last in, first out” stack mechanism, so, when the user is done with the current activity and presses the Back mutton, it is popped from the stack (and destroyed) additionally the preceding action resumes. When an activity is stopped because an innovative activity starts, it is notified of this change in state through the activity’s lifecycle callback methods.

Acknowledgements

I would like to thank my university for giving me this golden opportunity to research on such an interesting topic. I would also like to thank Dr. Anil Kumar, Assistant Pro Vice Chancellor & Director, ASET and Dr. Deepak Arora, Professor & Head, Dept of CSE & IT, ASET for encouraging students to indulge in valuable research activities to enhance their technical skills. Next, I want to thank my faculty guide Dr. Pawan Singh for guiding me throughout the course of this project. I would also like to thank the internet, books and the institute for the knowledge I acquired with their help.

References

[1] S. A. Macskassy and F. Provost, “Classification in networked data: A toolkit and a univariate case study,” The Journal of Machine Learning Research, vol. 8, pp. 935–983, 2007.

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning (still) requires rethinking generalization,” Commun. ACM, vol. 64, no. 3, pp. 107–115, 2021.

[3] S. Subudhi, R. N. Patro, and P. K. Biswal, “Superpixel clustering based segmentation algorithm for hyperspectral image classification,” in 2019 International Conference on Information Technology (ICIT), 2019.

[4] R. Gandhi, “Introduction to machine learning algorithms: Linear regression,” Towards Data Science, 27-May-2018. [Online]. Available: https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a. [Accessed:11-Jan-2021].

[5] M. Summerfield, Programming in python 3: A complete introduction to the python language, 2nd ed. Boston, MA: Addison-Wesley Educational, 2010.

[6] W. McKinney, Python for Data Analysis, 2e. Sebastopol, CA: O'Reilly Media, 2017.

[7] S. Nabwire, H.-K. Suh, M. S. Kim, I. Baek, and B.-K. Cho, "Review: Application of artificial intelligence in phenomics," Sensors (Basel), vol. 21, no. 13, 2021.

[8] A. Ganguly, IBM Watson solutions for machine learning: Achieving successful results across computer vision, natural language processing and AI projects using Watson cognitive tools. New Delhi, India: BPB Publications, 2021.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In Proceedings of the International Conference on Machine Learning (ICML'99), pp. 200–209, 1999.

[10] N. Zulkarnain and M. Anshari, "Big data: Concept, applications, & challenges," in 2016 International Conference on Information Management and Technology (ICIMTech), 2016.

[11] I. Arel, D. C. Rose and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," in IEEE Computational Intelligence Magazine, vol. 5, no. 4, pp. 13-18, Nov. 2010, doi: 10.1109/MCI.2010.938364.

[12] Y. Low, J. E. Gonzalez, A. Kyrola, D. Bickson, C. E. Guestrin, and J. Hellerstein, "GraphLab: A new framework for parallel machine learning," arXiv [cs.LG], 2014.

[13] B. Singh, P. Sihag, S. M. Pandhiani, S. Debnath, and S. Gautam, "Estimation of permeability of soil using easy measured soil parameters: assessing the artificial intelligence-based models," ISH j. hydraul. eng., pp. 1–11, 2019.

[14] A. Sharma, P. Agrawal, V. Madaan, and S. Goyal, "Prediction on diabetes patient's hospital readmission rates," in Proceedings of the Third International Conference on Advanced Informatics for Computing Research - ICAICR '19, 2019.