Comparative Study and Correlative Analysis of Machine Learning and Data Mining Algorithms for better accuracy

Shirisha Kampati, Dr. Kakara Santhi Sree

Abstract—Machine Learning the popular expression in the new years that we are hearing is an important subfield of Artificial Intelligence. ML algorithms enable PCs to learn all alone without the need for programming expressly and upholds for development with experience. ML, the emerging advancements today that discovers applications taking all things together with everyday life exercises. ML is utilized in different fields, for example, Bioinformatics, Intrusion Detection, Information Retrieval, Game playing, Advertising, Malware Detection so on. This paper gives a general thought on different Machine Learning Algorithms. First, we center on different ML algorithms, highlight the promising learning strategies in this field. Next, we investigate the close associations of Machine Learning techniques with Data Mining approaches for better accuracy.

Index Terms—Artificial Intelligence, Data Mining, Machine Learning

1 Introduction

THE term Machine Learning was instituted by Arthur Samuel in 1959,an American pioneer in the field of PC gaming and AI,expressed that "It gives computers the ability to learn without being explicitly programmed."

ML is a sub field of Data Science that emphasizes on designing algorithms that learns from previous data and predict future outcomes. Machine Learning includes acquiring knowledge from past information and utilizing that knowledge to make future forecasts, this can be achieved without programming explicitly. It will probably empower PCs to learn all alone. ML is the science that enables frameworks to learn and enhance their efficiency without being specifically programmed [10]. The way toward learning starts with perceptions or information, which might be models, direct insight, or guidance, to look for patterns in information and settle on better choices later on dependent on the examples that we provide. The essential objective is to permit the PCs adapt automatically without human intervention and change the activities as required[11]-[12] The process of extracting knowledge from large volumes of data stored in various data sources like databases, data warehouses, and other information repositories is called as Data Mining. It is an important step in Knowledge Discovery from Data (KDD).

Data Mining and Machine Learning use similar key algorithms to discover patterns in the data. However their process and utility will differ. In contrast to Data Mining, the computer in Machine Learning can automatically learn model parameters from data [2], [5].

2 LEARNING APPROACHES IN ML

The main types of learning problems in ML are Supervised learning,Unsupervised Learning, Reinforcement Learning and Semi-supervised Learning as shown in Fig.1 [11],[1].

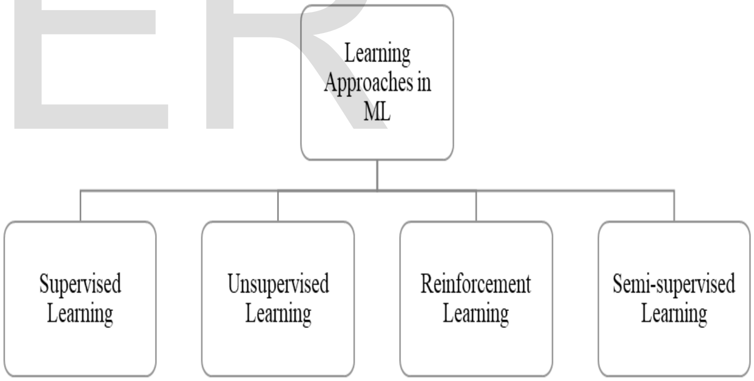


Fig. 1.Different Learning approaches in ML

2.1 Supervised Learning

This learning approach refers to a class of systems and algorithms that determine a predictive model using data points with known outcomes. The model is learned by training through an appropriate learning algorithm [3].

There are two main Supervised Learning problems are-

Classification: Supervised learning problem that includes anticipating a class label.

Regression: Supervised learning problem that includes foreseeing a numerical label.

- *Shirisha Kampati is currently pursuing PhD in Computer Science from JNTU, Hyderabad, Telangana ,India, PH-9866785488. E-mail: shirsha.krishna@gmail.com*
- *Dr. Kakara Santhi Sree is currently working as Professor ,Department of Computer Science and Engineering , School of Information Technology(SIT), Jawaharlal Nehru Technological University(JNTUH), Kukatpally, Hyderabad, Telangana, India.*

2.2 Unsupervised Learning

Unlike supervised learning, UL extracts similar features among the data and try to cluster the data based on the similarity. The data does not contain prior class labels. Local information is used for learning, it find out the output on its own by identifying the patterns that exists in the given raw data. There are numerous kinds of Unsupervised Learning problems, the two fundamental types are Clustering and Density Estimation.

Clustering: It includes discovering clusters or groups based on similarity in data.

Density Estimation: Density Estimation includes summarizing the distribution of data.

2.3 Reinforcement Learning

Reinforcement Learning (RL) is a class of ML problems that learns what needs to be done in order to find the finest solution through repeated experiences. More specifically, RL is the process of learning by having close communication with the environment and observing the outcomes of various actions. It allows machines to automatically determine the ideal behaviour in a specific context, to maximize its performance. For that, a simple return of the results is necessary to learn how the machines must act. Reinforcement learning is a type of cognitive learning model in which the algorithm provides feedback based on data analysis. It differs from other learning approaches as the sample set of data does not train the machine, it learns by trial and error.

An example of reinforced learning is the recommendations based on the prior searching or viewing content on the online platforms like shopping, content viewing and browsing through videos etc. However,this machine understands based on the completion of the prior activities and suggest accordingly.For example, searching different items in e-commerce application,not purchased,in this case machine understand that the item would not be a good one to recommend and will try another approach next time[3],[13].

2.4 Semi Supervised Learning

A hybrid learning problem called Semi Supervised Learning is Supervised ML approach, which encompasses the training data with very few labelled samples and a huge sample size that haven't been labelled [13]. Semi-supervised learning models will make effective use of all data, not only labelled data.

3 MACHINE LEARNING ALGORITHMS

The approach towards finding algorithms that have improved courtesy of experience derived from data. It is the planning, research, and development of algorithms that enable machines to learn without the need for human intervention [4]-[5].

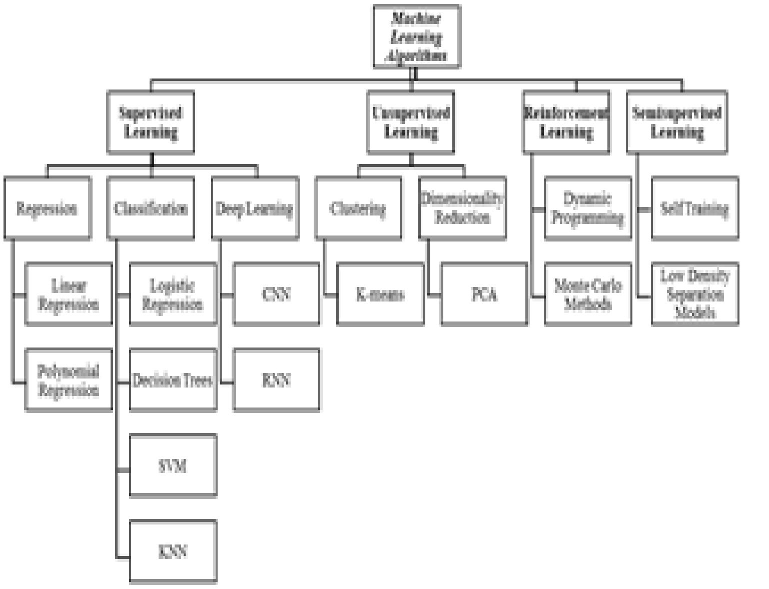


Fig. 2. Machine Learning Algorithms

Various Machine Learning Algorithms available as shown in Fig.2 are [14]:

3.1 Linear Regression

Regression analysis focuses on one dependent variable and a series of other changing variables–making it particularly useful for prediction and forecasting [6]-[7].

Linear Regression algorithms shows the relationship that exists between 2 variables and how the change in one variable effect the other variable. It shows the impact of changing independent variable on dependent variable. By fitting the best line, we define the relationship between the independent and dependent variables. This best fit line is referred to as a regression line, and it is defined as a linear equation is $Y= p *X + q$. In this, dependent variable as Y, independent variable as X, The coefficients p and q are determined by minimising the sum of squared distance variations between data points and regression line.

Types of Linear Regression:

Simple Linear Regression: It is described by one independent variable.

Multiple Linear Regression: It is described by multiple (>1) independent variables. We may fit a polynomial or curvilinear regression when evaluating the best fit rows. These are also known as polynomial or curvilinear relapse.

3.2 Logistic Regression

Logistic Regression algorithm is for classification tasks, it is not a regression algorithm. This algorithm predicts a categorical dependent variable's outcome based on predictor variables by using a logistic function on linear combination of features. Logistic regression algorithms estimate the likelihood of falling through a certain degree of categorical dependent variable based on predictor variables [6].

Types of Logistic Regression:

Binary Logistic Regression –the target variable has 2 potential outcomes i.e. either 1 or 0. Example –Predicting student result on a given exam as pass or fail, predicting whether lump/tumor on a human body is cancerous or not.

1. Multi-nominal Logistic Regression – In this the target variable has three or many more possible outcomes with no order. Example- Knowing the food preferences through online food delivery platforms like Uber Eats, EatClub, Swiggy, Zomato etc.
2. Ordinal Logistic Regression – The target variable has 3 or more probable outcomes with ordering among them. Example- Product feedbacks, reviews on a 1 to 5 scale.

3.3 Decision Tree

A Decision Tree is a flowchart-like tree structure, where each internal node (non-leaf node) signifies a test on an attribute, each branch represents a result of the test, and each leaf node (or terminal node) holds a class label. Given a tuple, T, the class label is not available, the tuple's attribute values are validated against each node of the decision tree. A path is followed from the root node through non-leaf nodes to a leaf node that contains the class prediction for that specific tuple [12], [3], [5], [9].

3.4 Support Vector Machine

Support Vector Machine is a classification or regression strategy. Given two types of set of points in N dimensional place, SVM generates a (N — 1) dimensional hyper plane to distinct those points into two groups.

SVM classifier is mathematically more multifaceted or complicated than a distance-based classifier. A SVM classifier, on the other hand, has better generalisation competences than a distance-based classifier and it is faster on a large sample set because it only operates on support vectors [12], [3], [5].

3.5 Naïve Bayes

Naive Bayes classifiers are a kind of simple probabilistic classifiers based on applying Bayes' theorem, the features are considered to have a high degree of independence from one another. This classification algorithm is based on Bayes theorem. The assumption of Naive Bayesian algorithms is class conditional independence which means the values of the attributes are conditionally independent of one another given class label of the sample. It suites best for large datasets [6].

Bayes theorem provides a way of calculating posterior probability P(X|Y) from P(X), P(Y) and P(Y|X) and is given by:

$P(X|Y) = P(Y|X) * P(X) / P(Y)$

Where P(X|Y) is the posterior probability of X given Y, P(X) is the prior probability of X, P(Y|X) is the probability of data Y given hypothesis X and P(Y) is the probability of data Y. [11], [12], [9].

3.6 k-Nearest Neighbor

k-Nearest Neighbor (kNN) is a supervised non parametric learning algorithm used for both classification and regression problems. The key idea behind kNN is, the class or value of a data point is determined with the help of data points around it. Nearest Neighbor classifiers depend on learning by likeness (analogy), that is, by contrasting a given test tuple with training tuples that are similar to it. The algorithm stores every single accessible case and classifies new cases by a majority vote of its k neighbors. The case being relegated to the class is most common among its k-Nearest Neighbors estimated by a distance measure. These distance measures can be Euclidean, Minkowski, Manhattan and Hamming distance. The First three distance measures are used for continuous variables and the fourth measure is (Hamming) for categorical variable [12], [5], [6].

3.7 K-Means

K-Means is an unsupervised algorithm for cluster analysis. It is a non-deterministic (exhibits different behaviours on different runs) and iterative approach that groups similar data in to clusters. The data points inside a cluster are homogeneous and heterogeneous to peer groups. It finds the centroids of k groups and assigns a data point to that cluster having least distance between its centroid and the data point. The k-means algorithm works as follows. To begin with, it arbitrarily chooses 'k' of the items, every one of which at first addresses a group mean or centre. For each of the left over objects, the object is allotted to the cluster to which it is the most alike comparative, based on the object's distance from the group mean. Then new mean is computed for each cluster. This procedure is repeated until the criterion function congregates [11].

4 EXPERIMENTAL PROCEDURE

4.1 Experimental Datasets

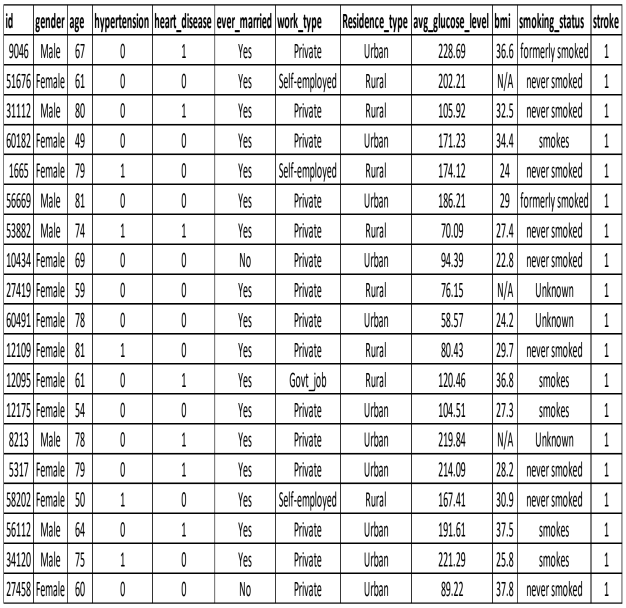
For comparison between various ML Algorithms the following 2 datasets are used.

1. Stroke Prediction Dataset (11 Clinical features for predicting stroke events) from Kaggle.com.Sample data is shown in TABLE 1.

Based on the input parameters, this dataset is used to determine when a patient is likely to face a stroke. Each row in the data provides relevant information about the patient. The data set consists of 5110 observations with 12 attributes.

TABLE 1

STOCK PREDICTION SAMPLE DATA

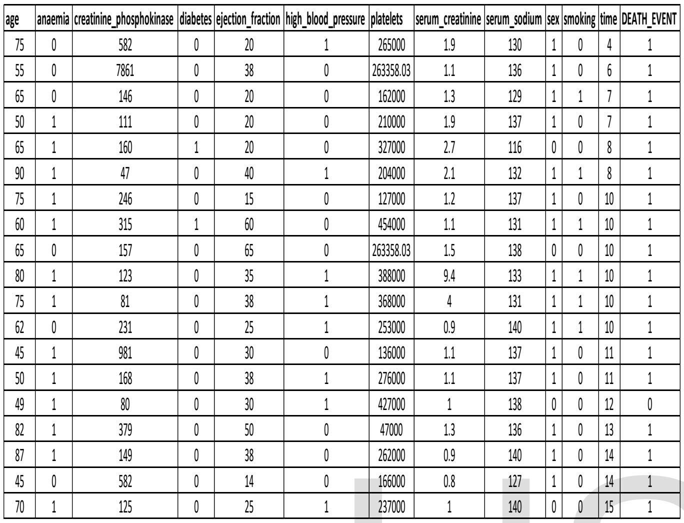


id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1

2. Heart Failure Clinical records Data set from UCI Machine Learning Repository.Sample data is shown in TABLE 2.

This dataset includes the 299 heart failure patients medical records are obtained during their follow-up period, with each patient profile containing 13 attributes.

TABLE 2
HEART FAILURE SAMPLE DATA



age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1
62	0	231	0	25	1	253000	0.9	140	1	1	10	1
45	1	981	0	30	0	136000	1.1	137	1	0	11	1
50	1	168	0	38	1	276000	1.1	137	1	0	11	1
49	1	80	0	30	1	427000	1	138	0	0	12	0
82	1	379	0	50	0	47000	1.3	136	1	0	13	1
87	1	149	0	38	0	262000	0.9	140	1	0	14	1
45	0	582	0	14	0	166000	0.8	127	1	0	14	1
70	1	125	0	25	1	237000	1	140	0	0	15	1

4.2 Experimental Results

We have applied various ML algorithms on the above 2 datasets in Jupiter Notebook and the Accuracies of different algorithms is shown in TABLE 3.

TABLE 3
COMPARISON OF VARIOUS ALGORITHM ACCURACY ON 2 DATASETS

Algorithm	Stroke Prediction dataset	Heart Failure dataset
SVM	77%	80%
Logistic Regression	78%	78%
Decision Tree	92%	73%
Random Forest	92%	85%
XGBoost	91%	85%
K -Means	95%	50%

From TABLE 3, it can be stated that Random Forest is giving the more accuracy when compared to other algorithms on both the datasets. And the algorithms Random Forest, XGBoost, Decision Tree imply that the patient who is having a Stroke will likely to have more chances for Heart failure.

4.3 Comparative Analysis

In this section we have presented few graphs that depicts about how the stroke and deceased prediction analysis over few attributes (like age, bmi) from the data sets.

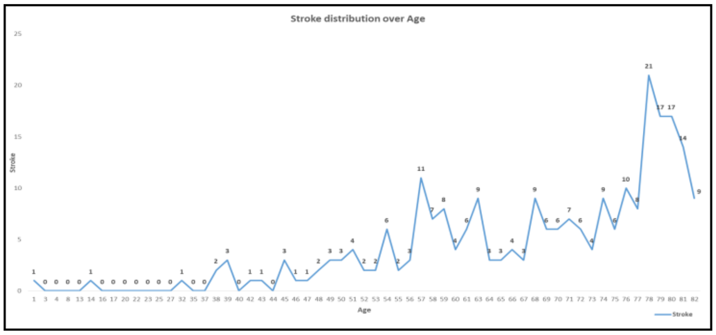


Fig. 3. Stroke distribution over Age attribute

The Fig. 3 shows the distribution of Stroke prediction over the Age attribute. The graph shows that there is higher chances of stroke having age more than 50 years in the Stroke prediction data set.

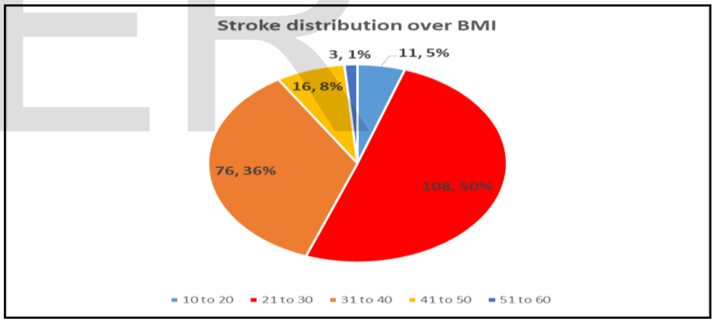


Fig. 4. Stroke distribution over BMI attribute

The Fig. 4 shows the distribution of Stroke prediction over the BMI attribute. The graph shows that lower BMI is having higher risk of stroke.

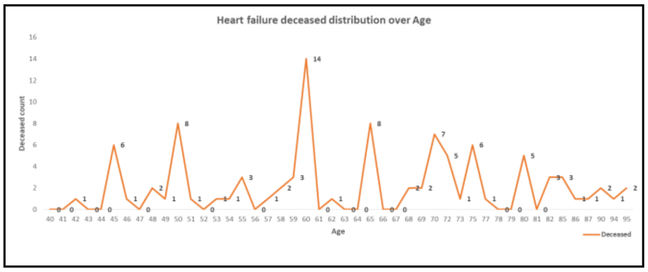


Fig. 5. Heart Failure deceased distribution over Age attribute

The Fig. 5 shows the distribution of Heart Failure over the Age attribute. The graph shows that there is higher chances of having Heart Failure deceased at the age of 55-61 years of Age for the Heart Failure Data set.

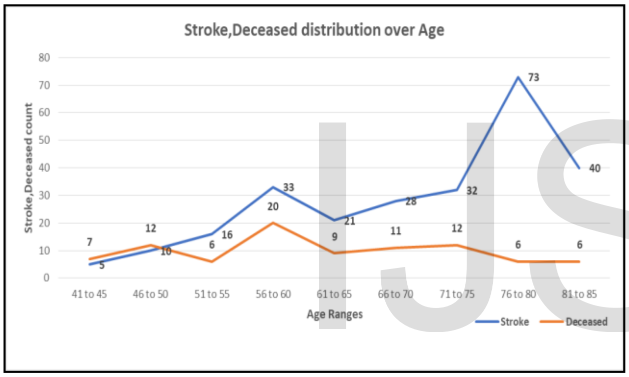


Fig. 6. Stroke, Deceased Distribution over Age

The Fig. 6 shows the distribution of Stroke, Heart failure Deceased distribution over the Age attribute from both the datasets.

It is clear from the graph in Fig. 6 that there are higher chances of getting stoke over 50 years of age and also higher fatality rate, if we dig further into the graph, the stroke rate is higher in the age range of 75-81 but the deceased rate is less in this age range. It's clear evident that age ranges from 55 to 61 is critical as there is higher chance of getting stroke as well as leading to deceased. In the age range of 65 to 85, there is higher chances of getting stroke but has very less fatality rate.

5 CONCLUSION

The fundamental commitment of this paper is to comprehend different Machine Learning Algorithms and Data Mining Algorithms. Based on the findings and discussions, it is clear that Random Forest gives better accuracy when compared with other algorithms on the 2 datasets. Age attribute is the crucial attribute among the other attribute which effects the stroke rate, deceased rate. Finding the relevant attribute which effects the accuracy of the models is very crucial. Machine Learning and Data Mining are two fundamental areas of Data Science today. It is possible that we will see more overlap as the two are used in combination to improve the usability and predictive capabilities of vast quantities of data.

REFERENCES

[1] Arun Kumar Rana, Ayodeji Olalekan Salau, Swati Gupta, Sandeep Arora," A Survey of Machine Learning Methods for IoT and their Future Applications", Amity Journal of Computational Sciences (AJCS),Vol.2,pp.1-5,2018.

[2] Anna L. Buczak and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, Vol. 18, pp.1153-1176,2016.

[3] B V Chowdary, Dr.Y Radhika, "A Survey on Applications of Data Mining Techniques", International Journal of Applied Engineering Research, Vol.13, pp.5384-5392, 2018.

[4] Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol.5, pp.1301-1309, 2017.

[5] Ayon Dey, "Machine Learning Algorithms: A Review', International Journal of Computer Science and Information Technologies", Vol. 7, pp.1174-1179, 2016.

[6] J.Deepika, T.Senthil, C.Rajan, A.Surendar, "Machine learning algorithms: a background artefact", International Journal of Engineering & Technology, Vol.7, pp.143-149, 2018.

[7] Keshav Singh Rawat "Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics", IOSR Journal of Computer Engineering (IOSR-JCE),Vol.19,pp.56-61, 2017

[8] Manoj Kumar Gupta, Pravin Chandra "A comprehensive survey of data mining", International Journal of Information Technology, Springer Link, Vol.12, pp.1243-1257, 2020.

[9] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, MK Series,2012.

[10] Kapil Sethi Ankit Gupta Gaurav Gupta Varun Jaiswal, "Comparative Analysis of Machine Learning Algorithms on Different Datasets", International Conference on Innovations in Computing (ICIC 2017), pp.87-91, 2017.

[11] Devanshi Dhall, Ravinder Kaur and Mamta Juneja, "Machine Learning: A Review of the Algorithms and Its Applications", Proceedings of ICRIC 2019, Vol.597, pp.47-63, 2020.

[12] Seema Sharma, Jitendra Agrawal, Shikha Agarwal, Sanjeev Sharma, "Machine Learning Techniques for Data Mining: A Survey", IEEE International Conference on Computational Intelligence and Computing Research, 2013.

[13] Herleen Kour(&) and Naveen Gondhi, "Machine Learning Techniques: A Survey", ICIDCA 2019, LNDECT,Vol.46, pp.266–275,2020

[14] Dr. O. Obulesu M. Mahendra M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey", Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018), pp.605-611, 2018.