Machine Learning Techniques for Better Data Driven Decisions Revisited

Tarika Verma, Nasib S. Gill

Abstract: ***The main goal of machine learning is to accurately predict the decisions to the problems without human expert intervention. These decisions depend upon patterns found and facts learnt during training tenure. However, prior incorporation of human knowledge is necessary for better prediction of the test data. The main aim is to make machines self-reliant for decision making. Providing machine with this vision makes it useful in every modern field. This makes the stepping stone to make computers behave as the humans do. Enhancing its speed and accuracy are the next step in this field. This paper presents a stock of techniques used to train the machines to respond to patterns present in the data sets so that useful information may be extracted for its potential use.***

Keywords: Machine Learning Techniques, Supervised ML, Unsupervised ML, Reinforcement ML, Naïve Bayes, SVM, Decision Tree, Regression, Clustering, Association Rule, Apriori

I. INTRODUCTION

Machine learning (**ML**) is the scientific and the statistical study in which computers are used to draw inference regarding a task without being given the explicit instructions by the programmer [1]. This inference is drawn based on some patterns found in the available sample data and which can further be used in making predictions or taking decisions in future [2].

ML algorithms works just like mathematical and statistical models in which the computer is provided with the sample training data. This is to make the computer able to respond to problems and so that itis able to conclude and to make predictions dynamically without being explicitly programmed [3]. Thus, unconventionally computer is made to solve the problems using its own decision-making system. Therefore, ML is also taken as a subset of Artificial Neural Network in which a machine learns to make correct predictions based on past experiences.

Various applications of MLinclude, Text Filtering, Fraud Detection, Natural Language Processing (NLP), Validating transactions, Business problem solving (Predictive Analytics), Data Mining, Machine Vision, Optical Character Recognition and so on.

Revised Manuscript Received on March 16, 2020.

***** Correspondence Author

Tarika Verma, Research Scholar, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak (Haryana), India. Email: tarika.verma@yahoo.co.in

Nasib Singh Gill, Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak (Haryana), India. Email: nasibsgill@gmail.com

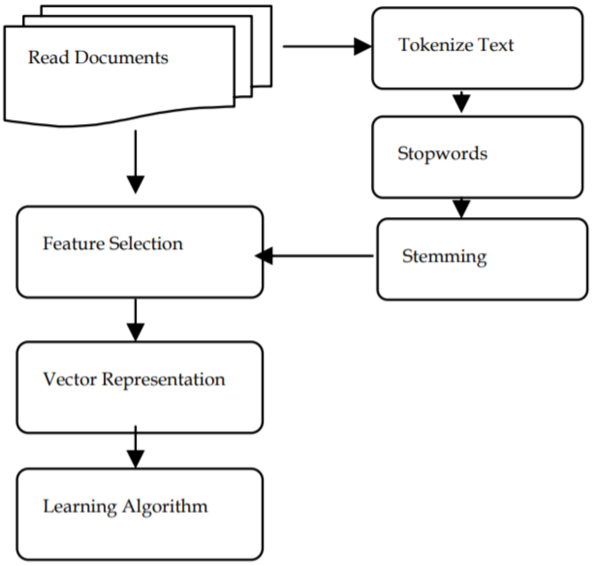


Fig 1: Document Classification Process [6]

Today ML is being used to handle large data sets at a minimal expenditure that is being generated on daily basis from the social life and networking world. The requirement of auto-knowledge retrieval from huge text sets of data to help in the human predictions and analysis is fully obvious [4].

For a better text classification, first the document is represented using a data structure and then a classifier is used to predicate the label of the class. Then the classifier may be evaluated based on accuracy when the test data is evaluated [5]. Documents pre-processing where dimensionality reduction (DR) is done firstly for well-planned data manipulation and depiction to remove irrelevant and redundant features. This increases performance speed and accuracy of classification algorithms [6].

II. MACHINE LEARNING TECHNIQUES

Machine learning (ML) includes training of computer systems to recognize patterns and behavior based on historical data to make decisions and to solve computational problems without using any explicit aid. ML is considered as an inseparable aspect of artificial intelligence [7]. ML uses training data to teach machine how to respond to the real-life problems and imitate the behavior of human or the living organisms. Machine learning is about predicting the future based on the past Implementation of machine leaning is broadly classified into two types: supervised and unsupervised learning. [12]

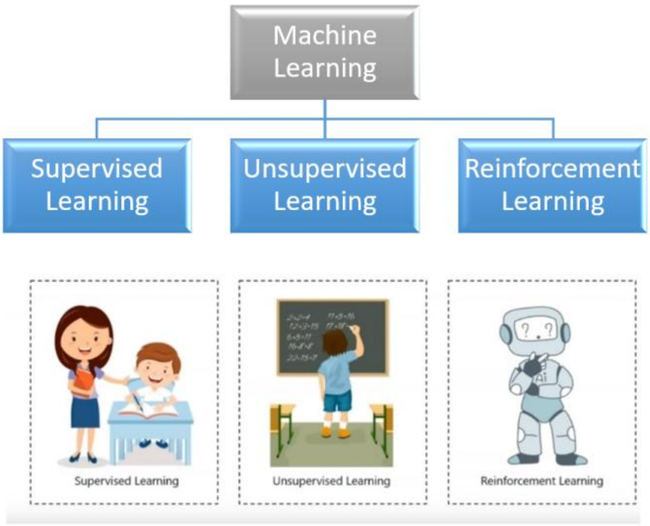


Fig. 2: Types of ML [10]

III.SUPERVISED ML

In this technique the machine learns under guidance. As evident from its name, its training data includes both input and output in the form of labelled training data so that results can be compared to desired output. It includes Classification and Regression techniques. [8]

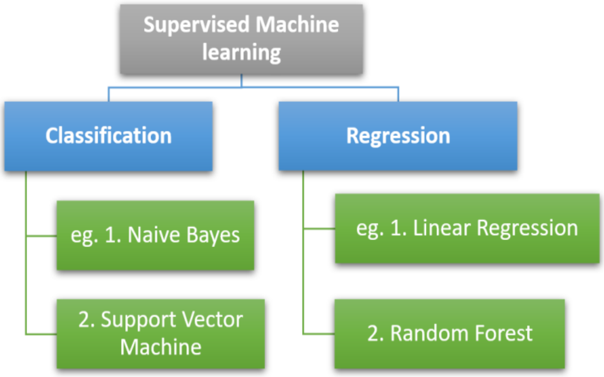


Fig. 3: Types of Supervised ML [8]

In Classification, the dataset is divided in labelled groups or discrete classes or categories. This labeling of classes is based on some condition or feature whose end result is a particular class or the label. For eg. Spam or Non-Spam classification of an Email.

In Regression, some continuous quantity or variable is predicted. For eg.Predicted price vs Actual Price of some product over a period of time. This is usually done by visualizing the relation between the dependent and independent variables. In this we have to predict the dependent variable (Y) value on the basis of the dependent variable (X). It is generally used while predicting a continuous quantity. This dependent variable is always continuous in Regression Model. The independent variable can be discrete or continuous. [9]

IV.BACKGROUND OF SUPERVISED ML ALGORITHMS

This section provides the background of supervised machine learning techniques.

A. Naïve Bayes Classifier

This method assumes that the features value of the entity is independent of each other statistically [10]. Let a feature vector ȳ = ($\hat{y}1$,$\hat{2}$,…..,$\hat{y}n$) and all the possible classes to which it belongs be D = (d_1,d_2,…..,d_m). The goal of NB classifier is to find out the probabilities p_1, p_2, ..., p_m for ȳ where p_a is the probability that ȳ belongs to category d_a. To find the belonging class of the feature vector ȳ, the value of "maximum ($p1$, $p2$, ..., pj)" is determined. Thus, the classification problem is

$$P(da\,|\hat{y}1,\hat{y}2,\ldots\ldots,\hat{y}n) = \frac{P(y1,y2,\ldots\ldots,yn|da\,)P(da\,)}{P(y1,\ \ y2,\ \ \ldots\ldots,\ \ yn)}$$

Where

P(da |ŷ1,ŷ2,…..,ŷn) = Probability of belonging to class d_a,given the feature vector ȳ = ($\hat{y}1$,$\hat{y}2$,…..,$\hat{y}n$)

P(d_a) = Probability of a random sample that is from class d_a.

P(y_1,y_2,…..,y_n|d_a) = Probability that given class d_a contains the feature vector ȳ = ($\hat{y}1$,$\hat{y}2$,…..,$\hat{y}n$)

And P(y1, y2, ….., yn)= Probability of occurrence of the feature vector. [12]

Naïve Bayes works incorrectly when the feature vectors in the training set becomes dependent on each other.

B. Support Vector Machine (SVM)

A SVM is a ML algorithm based on the "structural risk minimization principle" from "statistical learning theory" and itsgoal is to lessen error and computational complexity. The SVM aim is to discover an optimal dissociating training dataset classes usinga hyper-plane which acts as a boundary between data points. The optimum dissociation is attained by the hyper-plane that has the longest distance from the closest training dataset. During training dataset, assigned value of class is "1" or "0"depending its position overhead the hyper-plane, or below. New record is assigned a class as per its characteristics.[13]

The optimal hyper-plane is got by "minimizing" the objective function as below:

$$Min\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j\,(x_i,x_j)$$

Subject to

$$Min\sum_{i=1}^{n}\alpha_i\,y_j\ = 0$$

and

$$0 \le\ \alpha_i\ \le P$$

Here α_i= Lagrange multipliers and P= penalty.

The decision function h(x) for new data classification is defined as:

$$\mathrm{h}\ (\mathrm{x}) = sgn\left(\sum_{i=1}^{n} y_i a_i\ K(x_i, x_j) + b\right)$$

Here$K(x_i, x_j)$=Kernel functions (which includes linear, radial, polynomial, sigmoid functions) [13].

C. Decision Tree

Decision Tree is a "divide-and-conquer" approach, carried out from a given self-sufficient instances set, which give rise to a learning problem. In it, the root node represents a problem-statement with one or more solutions. Each solution further guides us to final decision via some intermediate problem representing nodes based on it. [14]

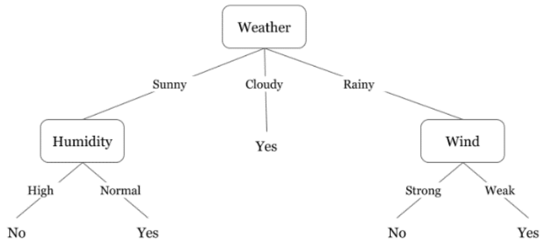


Fig. 4: Decision Tree for Visiting Park [15]

Here all the nodes belonging to a decision tree involve testing of one or more properties, or use some function of these properties. Leaf nodes of the decision tree give a classification that applies to all the instances that come in the path to reach that particular leaf node. Thus when an unseen instance is tested, it is routed from the root to the node on the next tree level according to its properties tested in consecutive nodes, and on reaching a leaf that instance is classified as per the classification of the leaf. [14]

D. Linear Regression

In it, the relation between the "dependent variable (Y)" and "independent variable (X)" is a straight line with best fit. Both variables vary linearly with respect to each other. [16] It can be represented by:

$$Y = b_0 + b_1 X + error$$

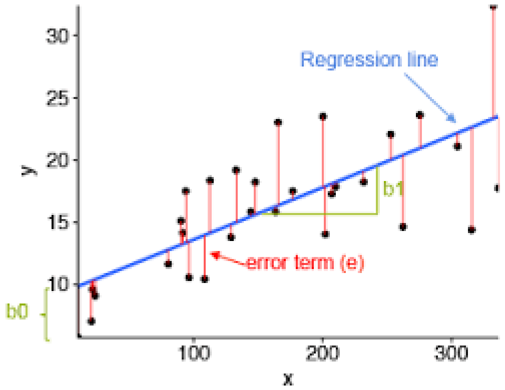


Fig. 5: Linear Regression [17]

For eg predicting a stock value (dependent variable) depends upon time (independent variable) as follows.

E. UNSUPERVISED ML

In it the machine learns is not under guidance and the machine gets unlabeled data. Machine itself identifies the patterns hidden in the data to make output predictions. In this we only have the "input data" and no corresponding "output variable". The goal is to imitate the basic structure of the data so as to gain insights [18].For e.g. We may classify people according to their age, knowledge, gender etc. when we visit a new place.

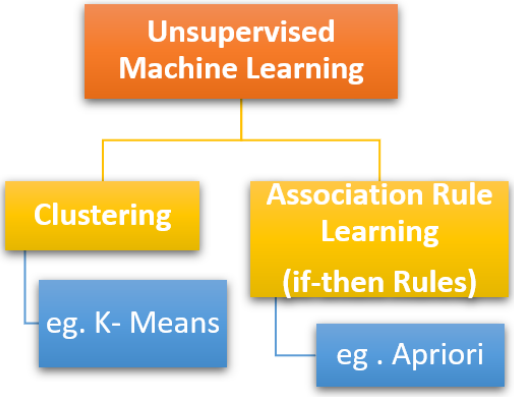

Unsupervised Machine Learning
Clustering
Association Rule Learning (if-then Rules)
eg. K- Means
eg . Apriori

Fig. 6: Types of Unsupervised ML [18]

In Clustering, similar data items are grouped together in order to identify the clusters or groups of data. [18]

In **Association Rule Learning**, emails are classified on basis of "rule-based classifiers" by using a set of "IF-THEN" rules. [19] Unlike other classifiers, here the emails' feature vectors are not required.

An example rule used is
IF "word FREE appears in subject" OR "word !!!! appears in subject"
THEN "the email is spam.

This is not generally the casualty but the co-occurrence pattern that comes to the force.

F. BACKGROUNDS OF VARIOUS UNSUPERVISED ML ALGORITHMS

A. K- Means Algorithm

This algorithm divides "M data points" (which are in "N dimensions") into "K clusters" in order to minimize the within-cluster sum of squares. We try to achieve the "local optima solutions" such that no inter-cluster data point manoeuvre reduces the sum of squares within-cluster. [20] Basically, this algorithm creates k clusters and pairs similar type of objects in a unique cluster. Thus, k clusters are formed in such a way that the constituents of a certain cluster are similar as compared to the non-cluster constituents of a certain data set. [18]

Proceedings

Initially, “k initial cluster centres” are selected and then iteratively refined as:

1. Each data instance “di” is allocated to cluster centre which is in its closest proximity.
2. Each cluster-centre“Cj” is then revised and this becomes equivalent to the mean of its elemental instances.

These steps are iterated until no further change is there in the apportionment of instances to clusters. Simply we may say that iterations are continued till cluster memberships are stabilized. This is called convergence. [20]

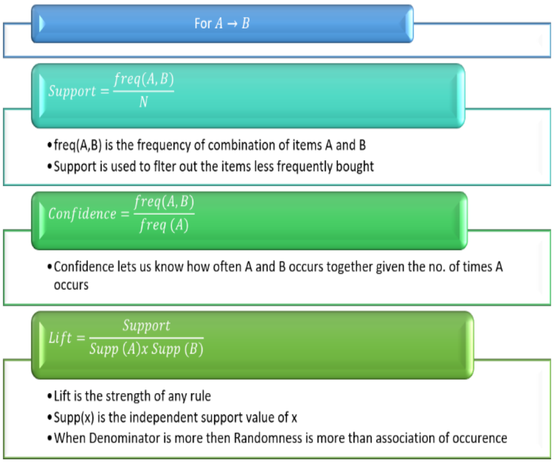
B. Hierarchical Clustering

It this algorithm clusters are created by makinggroups of similar objects. It is of two types: Agglomerative Algorithms (The bottom-up approach for making clusters i.e. small clusters are converted to bigger ones on combining some common characteristics) and Divisive Algorithms (The top-down approach for making clusters in which bigger cluster is divided in smaller clusters on basis of any difference) [18]

The final output is a “hierarchical tree” of the clusters or groups which is also called a “dendrogram”.

C. Apriori

It uses frequent item sets (having support > threshold) to create association rules. It is based on the fact that a “subset of a frequent itemset”must also be a “frequent item set”[21].



For A → B

$Support = \frac{freq(A,B)}{N}$

- freq(A,B) is the frequency of combination of items A and B
- Support is used to flter out the items less frequently bought

$Confidence = \frac{freq(A,B)}{freq(A)}$

- Confidence lets us know how often A and B occurs together given the no. of times A occurs

$Lift = \frac{Support}{Supp(A) x Supp(B)}$

- Lift is the strength of any rule
- Supp(x) is the independent support value of x
- When Denominator is more then Randomness is more than association of occurence

Fig. 7: Matrices to measure association [21]

D. REINFORCEMENT ML

In it, the agent learns to react to the environment. It allows the software agents or the machine to auto-learn the apt behavior within a context with max. performance. [11] It’s about interaction of two elements i.e. Learning Agent and Environment.

Environment leverages Exploration (when agent acts on trial and error basis) and Exploitation (when agent acts on the basis of knowledge gained from environment) mechanisms. This “Reinforcement Signal” is reward by environment to the agent for correct action and “Penalty” for the incorrect action.The agent selects the next action as per the rewards obtained with its improved environment knowledge. [22]

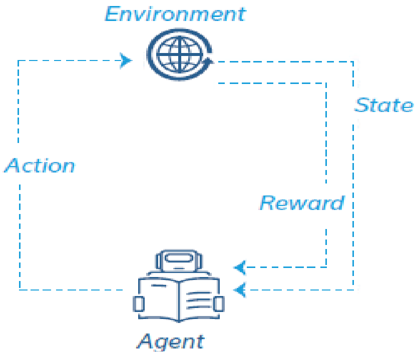


Fig. 8: Reinforcement ML [11]

Steps taken by learning agent for Reinforcement Learning [22]

1. Observation of environment
2. Select Action using some Policy
3. Take Action
4. Get Reward or Penalty
5. Update Policy (Learning Step)
6. Iterate above steps to get optimal Policy

E. COMPARITIVE ANALYSIS OF ML TECHNIQUES

Here we will consider the recently published papers and their analysis.

Table- I: Literature Analysis

Year	Title	Techniques	Analysis
2014	Email mining: tasks, common techniques, and tools [10]	• Naïve Bayes • SVM • Rule Based Classifier • KNN	Spam detection, Email categorization, Contact analysis, Email network property analysis, Email visualization and Other tasks are discussed
2018	Spam Filtering: A Comparison Between Different Machine Learning Classifiers [12]	• SVM • Naive Bayes • J48	Classifiers compared and SVM got max. accuracy and False Positive Rate.
2018	A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling [13]	• MultiBoost • AdaBoost • SVM • Logistic Regression	SVM and MultiBoost ensemble is used to model landslide susceptibility. Non linear relationship in landslide and its causing factors is studied.
2017	Data Mining Techniques for the Knowledge Discovery[14]	• Association • Classification • Clustering • Decision Tree • NN	Data mining techniques are explained in comprehendible manner.

Retrieval Number: D6766049420/2020©BEIESP
DOI: 10.35940/ijeat.D6766.049420

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

2019	Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges	• Hierarchical Learning • Data Clustering • Outliers Detection • Dimensionality Reduction • Neural Networks	It explains recent advancements, applications in unsupervised learning and Deep learning, in the context of networking. It also covers pitfalls and future challenges.
2017	Enhancing K-Means and Naive Bayes for Data Mining [20]	• K Means • Naïve Bayes	Algorithms ensembled way is proposed to procure aflexible mining model.
2019	A Weighted Frequent Itemset Mining Algorithm for Intelligent Decision in Smart System [21]	• Frequent Itemset Mining Algorithm (FIM)	Algorithm is proposed to limits the searching area of weighted frequent itemsets and time efficiency improved.

We have thus studied that one can focus on frequent item sets for an increased time efficiency. One can also ensemble the algorithms also so as to cover the non-linear data items and train the machine for both linear and non- linear situation.

V. CONCLUSION

Data driven problems are to be solved in ML without human intervention by analysing the trends in the given data set. In this paper, we have presented how the machine learning is useful in the current scenario by deeply explaining various types and algorithms and analysing recent developments in this field. For further studies one can go for Deep Learning combined with ML for a better accuracy and fast response of the ML algorithms.

REFERENCES

1. "Paraphrasing Arthur Samuel (1959), the question is: How can computers learn to solve problems without being explicitly programmed?" in Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170. doi:10.1007/978-94-009-0279-4_9
2. Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2
3. Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?". Computing Science and Statistics. 29 (1): 3–9.
4. Merrill lynch, Nov.,2000. e-Business Analytics: Depth Report. 2000.
5. Sebastiani, F., (2002) “Machine learning in automated text categorization” ACM Computing Surveys (CSUR) 34, pp.1 – 47.
6. Aurangzeb Khan, et al., (2010) "A Review of Machine Learning Algorithms for Text-Documents Classification," Journal of Advances in Information Technology, Vol. 1, No. 1, pp. 4-20, February.doi:10.4304/jait.1.1.4-20
7. T. Oladipupo, “Types of Machine Learning Algorithms,” New Adv. Mach. Learn., 2012.
8. [Online], Jason Brownlee (2016), “Supervised and Unsupervised Machine Learning Algorithms”, https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
9. [Online], Jason Brownlee (2016), “Linear Regression for Machine Learning”, https://machinelearningmastery.com/linear-regression-for-machine-learning/
10. Tang, G., Pei, J., &Luk, W.-S. (2013). “Email mining: tasks, common techniques, and tools.” Knowledge and Information Systems, 41(1), 1–31. doi:10.1007/s10115-013-0658-2
11. [Online] (2019), https://www.netscribes.com/reinforcement-learning-industry-impact/
12. M Shajideen, Nasreen & V, Bindu. (2018). Spam Filtering: A Comparison Between Different Machine Learning Classifiers. 1919-1922. 10.1109/ICECA.2018.8474778.
13. Pham, B. T., Jaafari, A., Prakash, I., & Bui, D. T. (2018). “A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling”. Bulletin of Engineering Geology and the Environment. doi:10.1007/s10064-018-1281-y
14. Tarika Verma, Chhavi Rana. (2017). Data Mining Techniques for the Knowledge Discovery. International Journal of Engineering and Technology. 9. 351-354. 10.21817/ijet/2017/v9i3/170903S054.
15. [Online], Decision Tree, https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/
16. [Online], Jason Brownlee (2016), “Linear Regression for Machine Learning”, https://machinelearningmastery.com/linear-regression-for-machine-learning/
17. Kassambara (2018), “Linear Regression Essentials in R”, http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/
18. Usama, M., et al. (2019). “Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges”. IEEE Access, 1–1. doi:10.1109/access.2019.2916648
19. Sarno, R., et al.. (2015). “Hybrid Association Rule Learning and Process Mining for Fraud Detection”. IAENG International Journal of Computer Science, 42(2).
20. Tarika Verma et al. (2017). Enhancing K-Means and Naive Bayes for Data Mining. International Journal of Engineering and Technology. 348-350. Doi 10.21817/ijet/2017/v9i3/170903S053.
21. Kowsalya, A., et al. (2019). “A Weighted Frequent Itemset Mining Algorithm for Intelligent Decision in Smart System”., IEEE Access. DOI 10.1109/ACCESS.2018.2839751,
22. Hessel, M., et al. (2019, July). Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 3796-3803).

AUTHORS PROFILE

Ms. Tarika Verma, has passed B.Tech. in 2015 and M.Tech. in 2017 in Computer Science and Engineering from University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak, India. She had topped in M.Tech (CSE) at UIET, M.D. University in 2017. She has also worked as Assistant Professor at AIJHM College, Rohtak. She is currently pursuing Ph.D. in Computer Science at M. D. University, Rohtak. She has published several research papers in Journals and Conference Proceedings. Her research interests include IoT, Machine Learning, Big Data Analytics and Data Mining.

Dr. Nasib Singh Gill, is at present senior most Professor of Dept. of CS & Applications, M. D. University, Rohtak, India and is working in the Dept. since 1990. He earned his Doctorate in CS in the year 1996 and carried out his Post-Doctoral research at Brunel University, West London during 2001-2002. He is a recipient of Commonwealth Fellowship Award of British Government for the Year 2001. Besides, he also has earned his MBA degree. He has published more than 245 research papers in reputed National & International Journals, Conference Proceedings, Bulletins, Edited Books, and Newspapers. He has authored seven books. He is a Senior Member of IACSIT as well as a fellow of several professional bodies including IETE and CSI. He has been serving as Editorial Board Member, Guest Editor, Reviewer of International/National Journals and a Member of Technical Committee of several International/National Conferences. He has guided so far 9 Ph.D. scholars as well as guiding about 7 more scholars presently in the areas – IoT, Machine Learning, Information and Network Security, Computer Networks, Measurement of Component-based Systems, Complexity of Software Systems, Decision Trees, Component-based Testing, Data mining & Data warehousing, and NLP.