Machine Learning Concept, Algorithms and Applications: A Survey

Mr. Chintu Kumar

Research Scholar, SSSUTMS, Sehore[1]

ABSTRACT

From the past decade, Machine Learning (ML) has progressed from the endeavour of few computer aficionados exploiting the prospect of computers learning to play games, and Mathematics (Statistics) that seldom considered computational methods, to an autonomous research discipline that has not solitary provided the obligatory base for statistical-computational principles. Machine Learning is a discernment which consensuses the machine to attain from examples and experience, and that furthermore deprived of being obviously programmed. This also offers many other commercial resolutions and has led to a distinct research interest in data mining to recognize hidden symmetries or asymmetries in social data that growing by second. This paper mainly emphases on explaining the perception and evolution of Machine Learning, some of the prevalent Machine Learning algorithms and try to associate three most popular algorithms based on some basic notions. We also discuss the application of machine learning in different sectors.

KEYWORDS: *Machine Learning, Data Mining, Statistics, Bayesian algorithm*

I. INTRODUCTION

Machine Learning is a perception which consents the machine to acquire from examples and experience, and that moreover deprived of being overtly programmed. Thus, instead of you scripting the code, what you do is you feed data to the generic technique, and the technique/ machine builds the logic based on the given data.[8] It permits the computers system or the machines to construct data-driven decisions rather than being explicitly programmed for carrying out a certain task. These programs or techniques are designed in a way that they learn and improve over time when are exposed to new data. Machine Learning Technique is trained using a training data set to create a model. When new input data is introduced to the ML technique, it makes a prediction on the basis of the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning technique is deployed. If the accuracy is not acceptable, the Machine Learning technique is trained again and again with an augmented training data set.[1]

This is just a very high-level example as there are many factors and other steps involved shown in fig.1. While designing a machine (a software system), the programmer always has a specific purpose in mind. For instance, consider J. K. Rowling's Harry Potter Series and Robert Galbraith's Cormoran Strike Series. To confirm the claim that it was indeed Rowling who had written those books under the name Galbraith, two experts were engaged by The London Sunday Times and using Forensic Machine Learning they were able to prove that the claim was true. They develop a machine learning algorithm and "trained" it with Rowling's as well as other writers writing examples to seek and learn the underlying patterns and then "test" the books by Galbraith. The algorithm concluded that Rowling's and Galbraith's writing matched the most in several aspects. So instead of designing an algorithm to address the problem directly, using Machine Learning, a researcher seek an approach through which the machine, i.e., the algorithm will come up with its own solution based on the example or training data set provided to it initially.[2]

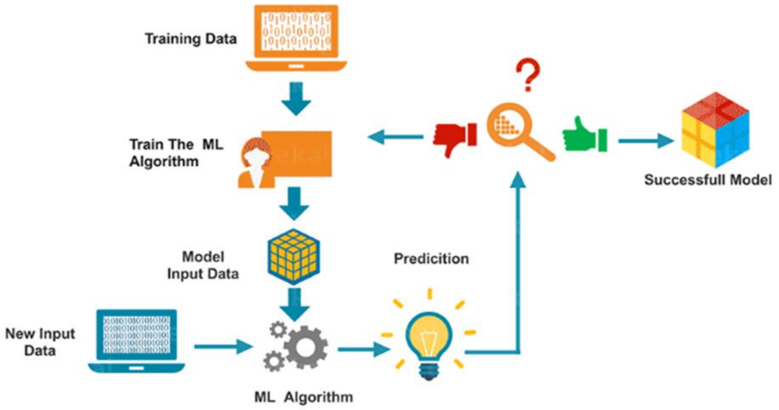

Fig.1: Working steps of Machine learning technique [1]

1.1 Classification of Machine Learning

There are three important types of Machine Learning Techniques such as supervised learning, unsupervised learning and reinforcement learning which we are discussing in detail:

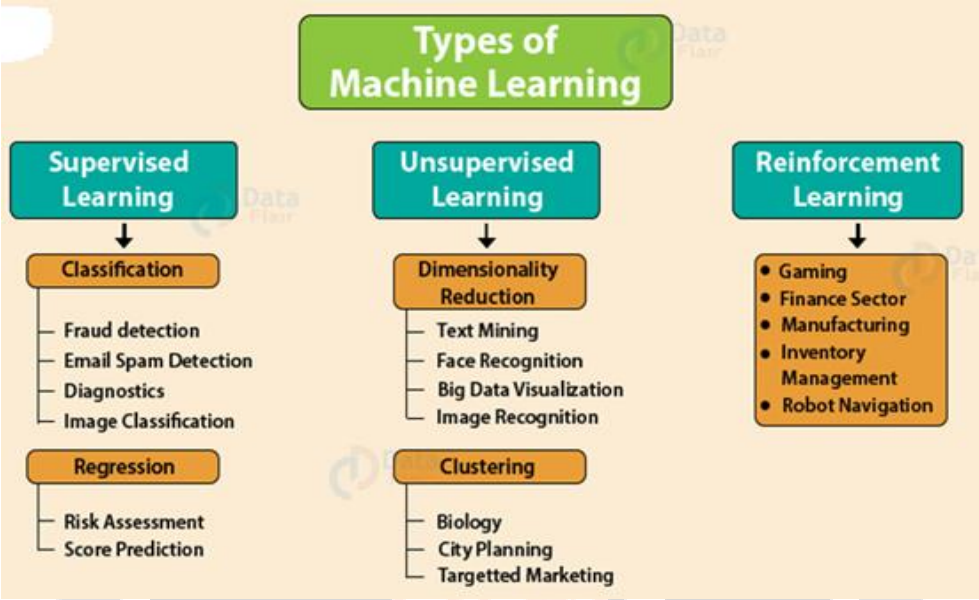

Fig.2: Classification of Machine Learning Techniques

A. Supervised Learning

Supervised Learning is the most popular paradigm for performing machine learning operations. It is widely used for data where there is a precise mapping between input-output data. The dataset, in this case, is labeled, meaning that the algorithm identifies the features explicitly and carries out predictions or classification accordingly. [2] As the training period progresses, the algorithm is able to identify the relationships between the two variables such that we can predict a new outcome. Resulting Supervised learning algorithms are task-oriented. As we provide it with more and more examples, it is able to learn more properly so that it can undertake the task and yield us the output more accurately. Some of the algorithms that come under supervised learning are as follows: Linear regression, random forest, support vector machine, artificial intelligence [3], etc.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.[3]

- Classification: Supervised learning problem that involves predicting a class label.
- Regression: Supervised learning problem that involves predicting a numerical label.

Both classification and regression problems may have one or more input variables and input variables may be any data type, such as numerical or categorical.

B. Unsupervised Learning

Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program. The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it.[4]

In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings. The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures.[4] This offers more post-deployment development than supervised learning algorithms. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes. Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters. The example of unsupervised learning is k-mean clustering, principle component analysis, SVD, FP-growth etc.

There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data.[4]

- Clustering: Unsupervised learning problem that involves finding groups in data.
- Density Estimation: Unsupervised learning problem that involves summarizing the distribution of data.

C. Reinforcement Learning

Reinforcement learning directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favorable outputs are encouraged or 'reinforced', and non-favorable outputs are discouraged or 'punished'. Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favorable or not.[5]

In case of the program finding the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favorable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result.[5]

In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward.[5] This simple feedback reward is known as a reinforcement signal.

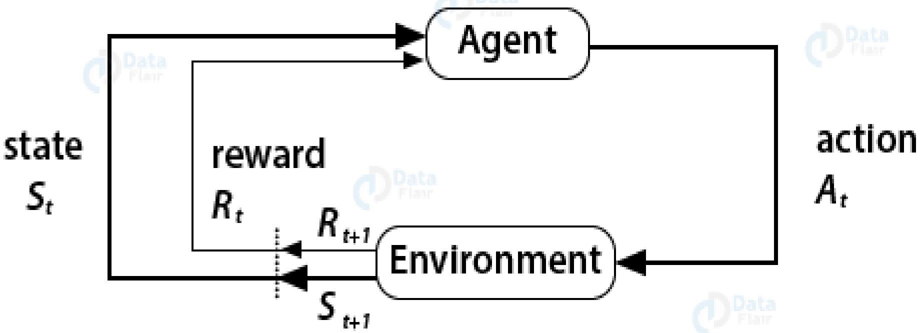

Fig. 3: Example of reinforcement learning

The agent in the environment is required to take actions that are based on the current state. This type of learning is different from Supervised Learning in the sense that the training data in the former has output mapping provided such that the model is capable of learning the correct answer. Whereas, in the case of reinforcement learning, there is no answer key provided to the agent when they have to perform a particular task. When there is no training dataset, it learns from its own experience.

D. Semi-Supervised Learning

In this type of learning, the given data are a mixture of classified and unclassified data. This combination of labeled and unlabeled data is used to generate an appropriate model for the classification of data. In most of the situations, labeled data is scarce and unlabeled data is in abundance (as discussed previously in unsupervised learning description).[6] The target of semi-supervised classification is to learn a model that will predict classes of future test data better than that from the model generated by using the labeled data alone. The way we learn is similar to the process of semi-supervised learning. A child is supplied with:

- Unlabeled data provided by the environment. The surroundings of a child are full of unlabeled data in the beginning.

Labeled data from the supervisor. For example, a father teaches his children about the names (labels) of objects by pointing toward them and uttering their names.

II. COMPARING PERFORMANCES OF GENERAL ML ALGORITHMS

Though various researchers have contributed to ML and numerous algorithms and techniques have been introduced as mentioned earlier, if it is closely studied most of the practical ML approach includes three main supervised algorithm or their variant. These three are namely, Naive Bayes, Support Vector Machine and Decision Tree. Majority of researchers have utilised the concept of these three, be it directly or with a boosting algorithm to enhance the efficiency further. These three algorithms are discussed briefly in the following section.

A. NAIVE BAYES CLASSIFIER

It is a supervised classification method developed using Bayes' Theorem of conditional probability with a 'Naive' assumption that every pair of features is mutually independent. That is, in simpler words, presence of a feature is not affected by presence of another by any means. Irrespective of this over-simplified assumption, NB classifiers performed quite well in many practical situations, like in text classification and spam detection. Only a small amount of training data is need to estimate certain parameters. Besides, NB classifiers have considerably outperformed even highly advanced classification techniques.

B. SUPPORT VECTOR MACHINE

SVM, another supervised classification algorithm proposed by Vapnik in 1960s have recently attracted an major attention of researchers. The simple geometrical explanation of this approach involves determining an optimal separating plane or hyperplane that separates the two classes or clusters of data points justly and is equidistant from both of them. SVM was defined at first for linear distribution of data points. Later, the kernel function was introduced to tackle non-linear datas as well.

C. DECISION TREE

A classification tree, popularly known as decision tree is one of the most successful supervised learning algorithms. It constructs a graph or tree that employs branching technique to demonstrate every probable result of a decision. In a decision tree representation, every internal node tests a feature, each branch corresponds to outcome of the parent node and every leaf finally assigns the class label. To classify an instance, a top-down approach is applied starting at the root of the tree. For a certain feature or node, the branch concurring to the value of the data point for that attribute is considered till a leaf is reached or a label is decided.

Now, the performances of these three were roughly compared using a set of tweets with labels positive, negative and neutral. The raw tweets were taken from Sentiment140 data set. Then those are pre-processed and labeled using a python program. Each of these classifiers were exposed to same data. Same algorithm of feature selection, dimensionality reduction and k-fold validation were employed in each case. The algorithms were compared based on the training time, prediction time and accuracy of the prediction. The experimental result is given below.

Table - 1: Comparison Between Gaussian NB, SVM and Decision Tree

Algorithm	**Training Time (In sec.)**	**Prediction Time (In sec.)**	**Accuracy**
Naïve Bayes (Gaussian)	2.708	0.328	0.692
SVM	6.485	2.054	0.6565
Decision Tree	454.609	0.063	0.69

But efficiency of an algorithm somewhat depends on the data set and the domain it is applied to. Under certain conditions, a ML algorithm may outperform the other.

III. APPLICATIONS OF MACHINE LEARNING

Machine Learning (ML) is a buzzword in the technology world right now and for good reason, it represents a major step forward in how computers can learn. The need for Machine Learning Engineers is high in demand and this surge is due to evolving technology and generation of huge amounts of data aka Big Data. The ML is used various sectors now a days in which some of the real-world applications is discussing below:

A. Traffic Alerts (Maps)

Now, Google Maps is probably the app we use whenever we go out and require assistance in directions and traffic. The other day I was traveling to another city and took the expressway and Maps suggested: “Despite the Heavy Traffic, you are on the fastest route”. But, how does it know that?

It’s a combination of People currently using the service, Historic Data of that route collected over time and few tricks acquired from other companies. Everyone using maps is providing their location, average speed, the route in which they are traveling which in turn helps Google collect massive Data about the traffic, which makes them predict the upcoming traffic and adjust your route according to it.

B. Social Media (Facebook)

One of the most common applications of Machine Learning is Automatic Friend Tagging Suggestions in Facebook or any other social media platform. Facebook uses face detection and Image recognition to automatically find the face of the person which matches its Database and hence suggests us to tag that person based on DeepFace. Facebook’s Deep Learning project DeepFace is responsible for the recognition of faces and identifying which person is in the picture. It also provides Alt Tags (Alternative Tags) to images already uploaded on Facebook.

C. Oil and Gas

This is perhaps the industry that needs the application of machine learning the most. Right from analyzing underground minerals and finding new energy sources to streaming oil distribution, ML applications for this industry are vast and are still expanding.

D. Transportation and Commuting (Uber)

If you have used an app to book a cab, you are already using Machine Learning to an extent. It provides a personalized application which is unique to you. Automatically detects your location and provides options to either go home or office or any other frequent place based on your History and Patterns. It uses Machine Learning algorithm layered on top of Historic Trip Data to make a more accurate ETA prediction. With the implementation of Machine Learning, they saw a 26% accuracy in Delivery and Pickup.[9]

E. Products Recommendations

Suppose you check an item on Amazon, but you do not buy it then and there. But the next day, you're watching videos on YouTube and suddenly you see an ad for the same item. You switch to Facebook, there also you see the same ad. So how does this happen? Well, this happens because Google tracks your search history, and recommends ads based on your search history. This is one of the coolest applications of Machine Learning. In fact, 35% of Amazon's revenue is generated by Product Recommendations.[9]

F. Healthcare

With the advent of wearable sensors and devices that use data to access health of a patient in real time, ML is becoming a fast-growing trend in healthcare. Sensors in wearable provide real-time patient information, such as overall health condition, heartbeat, blood pressure and other vital parameters. Doctors and medical experts can use this information to analyze the health condition of an individual, draw a pattern from the patient history, and predict the occurrence of any ailments in the future. The technology also empowers medical experts to analyze data to identify trends that facilitate better diagnoses and treatment.[8]

G. Speech Recognition

All current speech recognition systems available in the market use machine learning approaches to train the system for better accuracy. In practice, most of such systems implement learning in two distinct phases: pre-shipping speaker independent training and post-shipping speaker-dependent training. [7]

H. Computer Vision

Majority of recent vision systems, e.g., facial recognition software's, systems capable of automatic classification microscopic images of cells, employ machine learning approaches for better accuracy. For example, the US Post Office uses a computer vision system with a handwriting analyzer thus trained to sort letters with handwritten addresses automatically with an accuracy level as high as 85%.

I. Government

Government agencies like utilities and public safety have a specific need for ML, as they have multiple data sources, which can be mined for identifying useful patterns and insights. For example, sensor data can be analyzed to identify ways to minimize costs and increase efficiency. Furthermore, ML can also be used to minimize identity thefts and detect fraud.[15]

J. Bio-Surveillance

Several government initiatives to track probable outbreaks of diseases uses ML algorithms. Consider the RODS project in western Pennsylvania. This project collects admissions reports to emergency rooms in the hospitals there, and an ML software system is trained using the profiles of admitted patients in order to detect aberrant symptoms, their patterns and areal distribution. Research is ongoing to incorporate some additional data in the system, like over-the counter medicines' purchase history to provide more training data. Complexity of this kind of complex and dynamic data sets can be handled efficiently using automated learning methods only. [7]

K. Robot or Automation Control

ML methods are largely used in robot and automated systems. For example, consider the use of ML to obtain control tactics for stable flight and aerobatics of helicopter. The self-driving cars developed by Google uses ML to train from collected terrain data. [7]

L. Empirical Science Experiments

A large group data-intensive science disciplines use ML methods in several of it researches. For example, ML is being implemented in genetics, to identify unusual celestial objects in astronomy, and in Neuroscience and psychological analysis. The other small scale yet important application of ML involves spam filtering, fraud detection, topic identification and predictive analytics (e.g., weather forecast, stock market prediction, market survey etc.).

M. Financial Services

Companies in the financial sector are able to identify key insights in financial data as well as prevent any occurrences of financial fraud, with the help of machine learning technology. The technology is also used to identify opportunities for investments and trade. Usage of cyber surveillance helps in identifying those individuals or institutions which are prone to financial risk, and take necessary actions in time to prevent fraud.

.

IV. RELATED WORK

Kumar, and Kumar (2018) proposed framework predicts the achievement of a motion picture in light of its gainfulness by utilizing chronicled information from different sources. Utilizing informal community examination and content mining methods, the framework naturally separates a few gatherings of highlights, including "who" are on the best composition (actor and director) what a film is about, "when" a motion picture will be released, and in addition "semi variety" highlights that match "who" with "what", and "when" with "what". They proposed likewise made extraordinary commitments to the expectation. Moreover, to planning a choice emotionally supportive network with reasonable utilities, our investigation of key factors for motion picture productivity may likewise have suggestions for hypothetical research on group execution and the achievement of imaginative work.[11] *Latif and Afzal (2016)* used IMDB for our experimentation. They created dataset and then transformed it and applied machine learning approaches to build efficient models that can predict the movies popularity. Performing data mining on IMDB is a hard task because of so many attributes related to a movie and all in different dimensions with lots of noisy data and missing fields. After performing classification, they have found out that their best results are achieved through simple logistic and logistic regression at around 84 %. The attributes that contributed the most to information are metascore and number of votes for each movie, Oscar awards won by the movies and the number of screens the movie is going to be screened. [12] *Chaudhari et al. (2016)* developed a tool, which can predict the success of movie being a hit or flop. As this factor is important for everyone involved in the movie, for example: If a movie is flop, it exacerbates the image of actor or director. The tool will use searching algorithms and then use of bespoke system to predict the percentage of success of movie which is yet to be released. Their analysis of the data collected from various resources like IMDb, Kaggle. They gather a series of interesting facts and relationships using a variety of data mining techniques such as Bayes Classification Algorithm, Decision Tree etc. Subsequently, a classifier is learned and used to classify new movies with respect to their predicted box-office collection. Experimental results showed that the proposed approach improved the classification accuracy as compared to a fully independent setting.[13] *Meenakshi et al., (2018)* developed a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. They gathered a series of interesting facts and relationships using a variety of data mining techniques. In particular, they concentrated on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed. They additionally reported on the techniques used, giving their implementation and utility. Additionally, they found some attention-grabbing facts, such as the budget of a movie isn't any indication of how well-rated it'll be, there's a downward trend within the quality of films over time, and also the director and actors/actresses involved in the movie.[14] *Quader et al. (2017)* proposed a decision support system for movie investment sector using machine learning techniques. This research helps investors associated with this business for avoiding investment risks. The system predicted an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. Using Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-released features. They showed Neural Network gives an accuracy of 84.1% for pre-released features and 89.27% for all features while SVM has 83.44% and 88.87% accuracy for pre-released features and all features respectively when one away prediction is considered. Moreover, they figured out that budget, IMDb votes and no. of screens are the most important features which play a vital role while predicting a movie’s box-office success.[15] *Taegu Kim et al. (2017)* forecasting models developed for considering competition and word-of-mouth (WOM) effects in addition to screening-related information. Nationality, genre, ratings, and distributors of motion pictures running concurrently with the target motion picture are used to describe the competition, whereas the numbers of informative, positive, and negative mentions posted on social network services (SNS) are used to gauge the atmosphere spread by WOM. Among these candidate variables, only significant variables are selected by genetic algorithm (GA), based on which machine learning algorithms are trained to build forecasting models. The forecasts are combined to improve forecasting performance. Experimental results on the Korean film market show that the forecasting accuracy in early screening periods can be significantly improved by considering competition. In addition, WOM has a stronger influence on total box office forecasting. Considering both competition and WOM improves forecasting performance to a larger extent than when only one of them is considered.[16] *Minhoe Hur et al.(2016)* new box-office forecasting models are presented to enhance the forecasting accuracy by utilizing review sentiments and employing non-linear machine learning algorithms. Viewer sentiments from review texts are used as input variables in addition to conventional predictors, whereas three machine learning-based algorithms, i.e., classification and regression tree (CART), artificial neural network (ANN), and support vector regression (SVR), are employed to capture non-linear relationship between the box-office and its predictors. In order to provide variable importance for machine learning-based forecasting algorithms, an independent subspace method (ISM) is applied. Forecasting results from six different forecasting periods show that the presented methods can make accurate and robust forecasts.[17] *Antara Upadhyay et al. (2018)* proposed the use of a review system for predicting the success rate of a movie. Moviegoers’ opinions of a movie before and after the release of the movie will be determined using sentiment analysis. A custom dictionary will be developed comprising words commonly used in movie reviews, which will be mapped to their corresponding weight-age in order to score reviews on a scale of one to five, and accordingly classify the success rate of movies. Due to this methodology, user can easily decide whether to book the ticket in advance or not. The long-term gain from this approach is that any kind of movie like Hollywood, Bollywood, etc can be reviewed on the website. In future our website can be used for reviewing sports events and music concerts and also for reviewing product sales, etc. [18]

V. CONCLUSION

Machine learning is widely used technology nowadays which is very helpful for the statistical analysis of data. It also gains the information from the historical data. The machine learning algorithm is classified into three categories namely supervised, unsupervised and reinforcement learning and each classification techniques has their merits and demerits. In this paper we present the comparative analysis of mostly used algorithm such as decision tree, support vector machine, and naïve bayes classifer and it is found that naïve bayes gives more accurate results than the other two. This paper also presented the literature survey about one of the application movie prediction of machine learning techniques and application of machine learning technique in various sectors. In future, need to develop an ensemble classifier which can predict the data more accurately and effectively.

REFERENCES

[1] https://www.edureka.co/blog/what-is-machine-learning/

[2] Christopher M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", 2006. Page-3

[3] Russel, "Artificial Intelligence: A Modern Approach", January 1, 2015

[4] Hastie et al. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition (Springer Series in Statistics) , 2016, pp. 28. [online]. Available: https://www.amazon.com/Elements-Statistical-Learning-Prediction-Statistics.

[5] Richard S. Sutton et al., "Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning",2018, second edition, pp.-2. [Online]. Available : https://www.amazon.com/Reinforcement-Learning-Introduction-Adaptive-Computation.

[6] Mohammed et al., "Machine Learning Algorithms and Applications", 2017, CRC Press Taylor & Francis, London, ISBN-13:978-1-4987-0538-7.

[7] Das and Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017, pp. 1301-1309.

[8] [Online]: Available. https://www.outsource2india.com/software/articles/machine-learning-applications-how-it-works-who-uses-it.asp

[9] [Online]: Available. https://www.edureka.co/blog/machine-learning-applications/

[10] Vaseem Naiyer, Jitendra Sheetlani, Harsh Pratap Singh, "Software Quality Prediction Using Machine Learning Application", Smart Intelligent Computing and Applications, Springer 2020.Pp. 319-327

[11] Hemant Kumar, Santosh Kumar, "Predicting Movie Success or Failure using Linear Regression & SVM over Map-Reduce in Hadoop", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 6, June 2018, pp. 6426-6433.

[12] Latif and Afzal, "Prediction of Movies popularity Using Machine Learning Techniques", International Journal of Computer Science and Network Security, VOL.16 No.8, August 2016, pp 127-131.

[13] Chaudhari et al., "A Data Mining Approach to Language Success Prediction of A Feature Film", International Journal of Engineering Sciences & Management Research, 2016, pp. 1-9.

[14] Meenakshi et al., "A Data mining Technique for Analyzing and Predicting the success of Movie", Journal of Physics: Conf. Series 1000 (2018) 012100 doi :10.1088/1742-6596/1000/1/012100. Pp 1-9.

[15] Quader et al., "A Machine Learning Approach to Predict Movie Box-Office Success", 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017.

[16] Taegu Kim et al., "Box Office Forecasting considering Competitive Environment and Word-of-Mouth in Social Networks: A Case Study of Korean Film Market", Hindawi Computational Intelligence and Neuroscience, Volume 2017, Article ID 4315419, 16 pages.

[17] Minhoe Hur et al., "Box-office Forecasting based on Sentiments of Movie Reviews and Independent Subspace Method", Information Sciences August 16, 2016. Pp 1-31.

[18] Antara Upadhyay, "Movie Success Prediction Using Data Mining", International Journal of Engineering Development and Research, 2018. Volume 6, Issue 4, pp 198-203.