4th International Conference on Multidisciplinary Research
Osmania University Centre for International Program, Osmania University Campus, Hyderabad (India) (ICMR-2019) Conference World
2nd February 2019, www.conferenceworld.in ISBN:978-93-87793-67-5

APPROACH TO BUILD A MODEL USING DATA SCIENCE/MACHINE LEARNING

K Swathi[1], Dr. Baddam Indira [2]

[1]*Research Scholar, Department of Informatics, UCE, OU, Hyderabad*
[2]*Assistant Professor, Department of MCA, CBIT, Hyderabad*

ABSTRACT

The amount of data is increasing gradually in the real world. As the data is increasing, the machines can make use of the data to make effective decisions using certain algorithms called machine learning algorithms. The machines can learn based on the existing data and their corresponding outcomes and can tune themselves such that they can predict the future out of the machine learning models built. This paper aims to provide a means of understanding the roles and responsibilities of each and every individual who plays a part in building machine learning model, steps to build a model and types of various machine learning algorithms.

Keywords: Data scientist, exploratory data analysis (EDA), Feature engineering, Machine Learning, reinforcement learning, Semi-supervised learning, supervised learning, Unsupervised learning

I. INTRODUCTION

In the current modern era, the organizations collect large amounts of data, both for individual studies or continuous operations. The amount of data is growing gradually and becoming a great part of the everyday lives. The widely used term of the massive data is BIG DATA [10].

Bit→Byte→KB→ MB→GB→TB→PB→BIG DATA

As the data is growing, there needs to be some process which makes the lives easier to analyze process and make meaningful information out of it which helps in the growth of business. Keeping this in mind, tremendous powerful software tools are designed which helps the business make conclusions out of the vast data [12]. Data science is the field of study which is a combination of domain knowledge, computer programming and mathematical and statistical implementation on the digital data to extract meaningful information that helps in making business decisions. Machine learning is a category of algorithm that allows applications to become more accurate in predicting outcomes without being explicitly programmed.

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed"-Arthur Samuel (1959) [4].

II. ROLES AND RESPONSIBILITIES OF INDIVIDUALS IN BUILDING MACHINE LEARNING MODEL

There are various roles playing a vital role in making the business decisions fast and accurate.

2.1 DATA SCIENTIST

Data Science is the process of data cleansing, data preparation and data analysis and gets insights from existing raw data to make meaningful. This is the responsibility of Data Scientist. The Skills of the Data Scientist are demonstrated in Fig.1

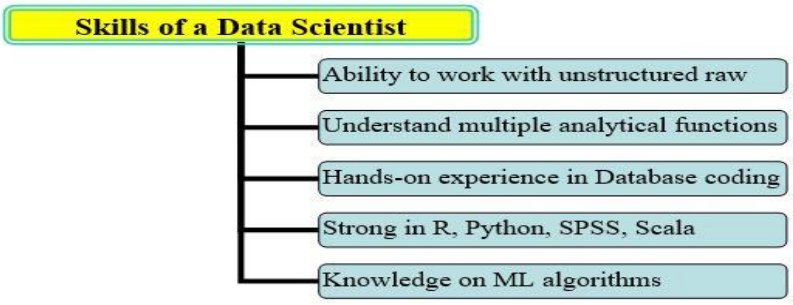

Figure 1. Skills of Data Scientist

2.2 DATA ANALYST

Data analysis is the process of performing basic operations of basic descriptive statistics, data visualization and data communication. This is the responsibility of Data Analyst. The Skills of the Data Analyst are demonstrated in Fig.2

Figure 2. Skills of Data Analyst

2.3 MACHINE LEARNING EXPERT

Machine learning is the process of using the algorithms on existing data to discover the hidden patterns in predicting the future trends. This is the responsibility of Machine Learning Experts. The machine learning algorithms are used to build the model. The Skills of the Machine Learning Expert are demonstrated in Fig.3

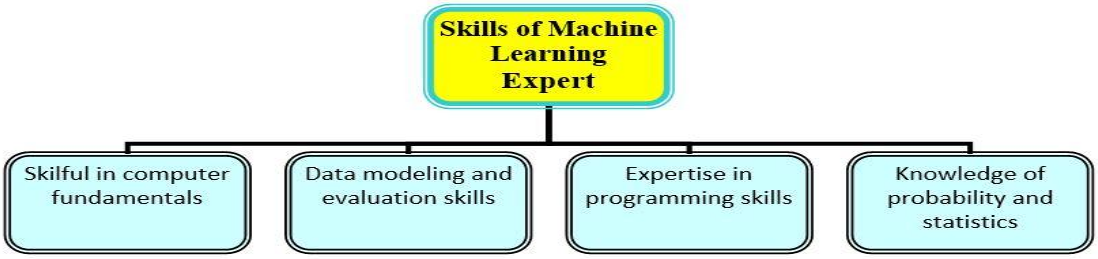

Figure 3. Skills of Machine Learning Expert

III. APPROACH IN BUILDING AN ACCURATE MACHINE LEARNING MODEL

- Define/Understand the Problem: The first and foremost thing to start with is to identify and understand the problem for which the solution is required.

- Data Collection: The process of moving the raw data to the environment where the analysis on the data is planned.

- Data Pre-processing: It is the process of transforming raw data to a form that the analysis can be performed without any difficulty to get accurate results. The actions performed are like eliminating the gaps, organizing the data, filtering the data, removing the duplicates transforming the data, normalization.

- Feature Engineering: The principal part of the feature engineering is to have the domain knowledge and creative ideas. The main task is to identify the new features in the data set that helps effectively in prediction and classification. Identifying the new feature or improving the existing feature to best fit in the model. Discovery of meaningful features contribute to quality in the final model.

- Exploratory Data Analysis (EDA): It involves actions and methods performed on data that help to summarize the characteristics, describe the facts, detect the patterns within the data, spot the outliers using plots, define and test the hypotheses. Various statistical operations like Univariate, Bivariate and multivariate analysis are performed to achieve the basic analysis on data.

- Algorithm Selection: Among various supervised, unsupervised, semi-supervised, reinforcement machine learning algorithms available, select the appropriate algorithm which best fits to the current model.

- Model Training and Evaluation: Split the data set into training set and test set. The test set should be selected by satisfying the conditions like the size of the test set is large enough and yield statistically meaningful results, it should represent the whole data set. The training data is selected and fed into the machine learning algorithm and design a model. Evaluate the model with the test set. If the performance is good on testing data also then the model designed is accurate, if not the model is again evaluated by readjusting the data and features considered [11].

After performing all the above steps, the final selected model will be deployed in applications. The flow of building Machine learning model is demonstrated diagrammatically in Fig.4

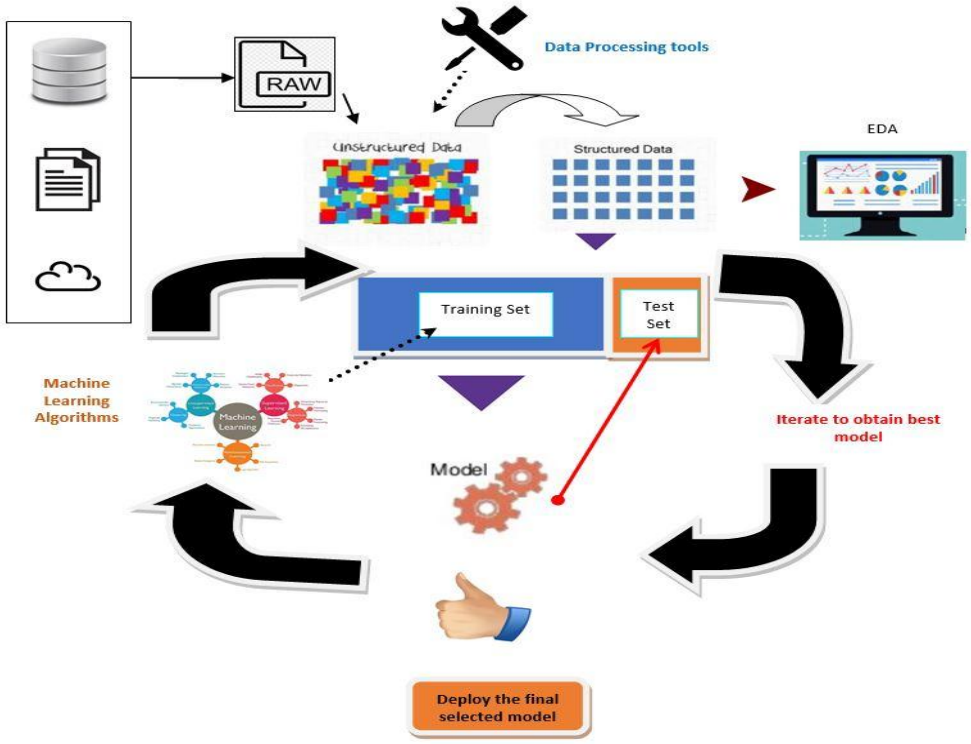

Figure 4. Building Model using Machine Learning

IV. TYPES OF MACHINE LEARNING ALGORITHMS

Machine Learning is a subset of artificial intelligence which helps the machines in learning from the existing data and makes decisions based on the previous experience. The Machine learning makes use of certain algorithms which are designed in such a way that the machines learn and improve as the new data is fed.

The algorithms are of 4 types.

- Supervised Learning Algorithms
- Unsupervised Learning Algorithms
- Semi-supervised Learning
- Reinforcement Learning

4.1 SUPERVISED LEARNING ALGORITHMS

Supervised learning algorithms are implemented to find out the mapping function f(X) from input variables(X) to output variable (Y). This algorithms works only if all the input variables are labeled. The goal is to obtain the below function.

Y = f(X) where X→ Input Variable (Labeled)

Y→ Output Variable

With the help of the existing data the model is trained and once the model gets trained, it can make decisions on the given new data. Supervised Machine Learning is diagrammatically represented in Fig.5

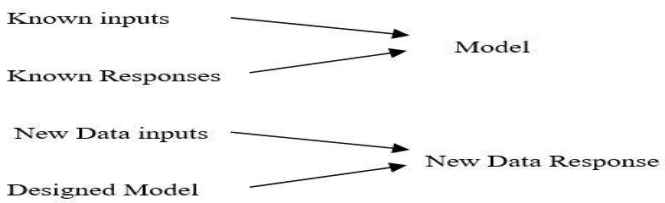

Figure 5. Supervised Machine Learning

Supervised learning algorithms are categorized as below:

- Classification Algorithms: Predict Discrete/Nominal Values
- Regression Algorithms: Predict Continuous values

Some of the most popularly used supervised learning algorithms are as follows [13]:

- Decision Trees
- Support vector machines for classification problems
- Random forest for classification and regression problems
- Linear regression for regression problems
- Ordinary Least Squares Regression
- Naive Bayes Classification
- Logistic Regression

4.2 UNSUPERVISED LEARNING ALGORITHMS

Unsupervised learning algorithms are implemented where there is only raw input data (X) and no corresponding output variables. From the raw data based on the interpretations, algorithms are implemented to find out the common patterns and relationships to create the clusters. Labels are not assigned to the clusters but the data set can be divided into groups based on the patterns and associations.

The model learns through observation from the existing data and finds structures in the data. Once the model is built on the existing dataset, it automatically finds patterns and relationships in the new dataset. Supervised Machine Learning is diagrammatically represented in Fig.6

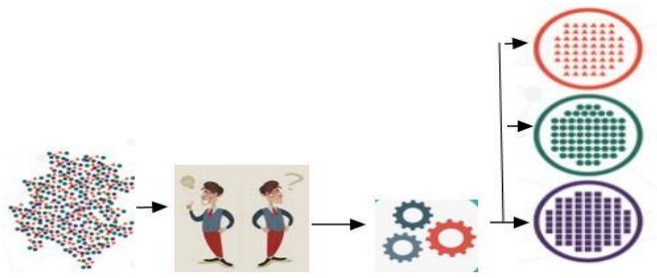

Figure 6. Unsupervised Machine Learning

Supervised learning algorithms are categorized as below:

- Clustering: Clustering problem is to discover the inherent groupings in the data.
- Association: An association is to discover rules that describe large portions of the data.

Some popular examples of unsupervised learning algorithms are [13]:

- K-means for clustering problems
- Apriori algorithm for association rule learning problems

- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis

4.3 SEMI-SUPERVISED LEARNING ALGORITHMS

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems. Here the raw data contains both labeled and unlabeled data set, majority is unlabeled.

These problems sit in between both supervised and unsupervised learning.

Semi-Supervised learning algorithms are categorized as below:

- Classification Algorithms: Predict Discrete/Nominal Values
- Clustering: Clustering problem is to discover the inherent groupings in the data.

4.4 REINFORCEMENT LEARNING ALGORITHMS

It is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. The trained model is used to predict the new data presented to it.

Reinforcement learning goes through the following steps:

- The agent observes the Input State.
- Using Decision making function the agent performs an action.
- After the action is performed, the environment sends a reward to the agent.
- Finally, the state and the action information about the reward is stored.

Reinforcement learning algorithms are categorized as below:

- Classification Algorithms: Predict Discrete/Nominal Values
- Control Algorithms: Control based on reward

Some of the most popularly used supervised learning algorithms are as follows [14]:

- Q-Learning
- Temporal Difference (TD)
- Deep Adversarial Networks

V. CONCLUSION

We conclude that each and every stage of building machine learning model plays a crucial role. A finely cleansed and wrangled data can give better results. There are various data processing tools which can perform

this task effectively. The one who performs these tasks should be very careful because any flaw at this stage needs an iteration of the complete cycle. The effective splitting of the training set and test set have great impact on the model. As the model is the key part in making the future predictions, the best model design can help in making accurate decisions by the business. Before choosing any algorithm, one should also consider its flaws and try to find if any other algorithm can build a better model.

REFERENCES

[1] Kirby McMaster, Brian Rague, Stuart L. Wolthuis, Samuel Sambasivam, "A Comparison or Key Concepts in Data Analysis and Data Science" Information Systems Education Journal, ISSN 1545-679X

[2] Taiwo Oladipupo Ayodele, "Types of Machine Learning Algorithms"

[3] Judith Hurwitz, Daniel Kirsch, "Machine Learning" IBM Limited Education

[4] Ahmad F.Al Musawi, "Introduction to Machine Learning" Machine Learning, UTQ, CSD, 3rd Satge

[5] Mohssen Mohammed, Muhammad Badruddin Khan, Eihab Bashier Mohammed Bashier, "Machine Learning: Algorithms and Applications" International Standard Book © 2017 by Taylor & Francis Group, LLC.

[6] Sebastian Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning"

[7] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, Martin Zinkevich, "Data Infrastructure for Machine Learning".

[8] Sumit Das, Aritra Dey, Akash Pal, Nabamita Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect" International Journal of Computer Applications (0975-8887) Volume 115-No. 9, April 2015.

[9] Iqbal Muhammad, Zhu Yan, "Supervised Machine Learning Approaches: A Survey" ICTACT Journal on Soft Computing, April 2015, Volume 05, Issue 03, ISSN: 2229-6956.

[10] Kirby McMaster, Brian Rague, Stuart L. Wolthuis, Samuel Sambasivam, "A Comparison of Key Concepts in Data Analytics and Data Science" Information Systems Education Journal (ISEDJ) ISSN: 1545-679X

[11] Rahul Chourasiya, Vaibhav Patel, Anurag Shrivastava, "Classification of Cyber Attack Using Machine Learning Technique at Microsoft Azure Cloud" International Research Journal of Engineering & Applied Sciences, IRJEAS, ISSN (O): 2322-0821, ISSN (P): 2394-9910, Volume 6 Issue 1, Jan 2018-Mar 2018.

[12] Mike Innes, David Barber, Tim Besard, "On Machine Learning and Programming Languages" SysML 2018, February 2018, Stanford CA, USA

[13] Taiwo Oladipupo Ayodele, "Types of Machine Learning Algorithms" New Advances in Machine Learning

[14] Ayon Dey, "Machine Learning Algorithms: A Review" International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1174-1179, ISSN: 0975-9646