Advanced Role of Machine Learning in Data Science Analysis of Current Trends and Future Implications

[1]Dr.Guru Kesava Das.Gopisetty [2] P.Jagadeesh Babu [3]M. Bharathi Rani [4] Boyina Kalyani
[1.] *Professor & HOD, Dept. of CSE, KITS Akshar Institute of Technology, Yanamadala, Guntur India*
[2.]*Assistant professor, Dept. of CSE, Eluru College of Engineering and Technology, Eluru*
[3.]*Assistant professor, Dept. of CSE, Malineni Lakshmaiah Woman's Engineering College, Guntur*
[4.] *Assistant professor, Dept. of IT, KKR&KSR Institute of Technology & Sciences, Guntur*

Abstract: The machine learning empowers data science to reduce human efforts and most valuable asset for business needs through pattern recognition, prediction, analysis and efforts. One notable aspect is the meticulous examination of extensive data sets, enabling the generation of valuable forecasts that may inform improved decision-making and prompt intelligent actions in real-time without the need for human involvement. With the development of economic globalization the rapid development of industries in various fields, big data technology has attracted more and more attention. Network data is constantly being generated at an unprecedented rate and it is necessary to intelligently process huge data, and then to make full use of the value in the data, you need to use machine learning method. The growing role of data science (DS) and machine learning (ML) in high-energy physics (HEP) is well established and pertinent given the complex detectors, large data, sets and sophisticated analyses. It was recommended that more research is conducted on the impact of ML on society, stronger regulations and laws to protect the privacy and rights of individuals when it comes to ML should be developed, transparency and accountability in ML decision-making processes should be increased, and public education and awareness about ML should be enhanced. In this paper a detailed overview of different structures of Data Science and address the impact of machine learning on steps such as Data Collection, Data Preparation, Training the model.

Index Terms: Machine Learning, Data Science, Data Collection, HEP, Data Set, Data Preparation, Domain Knowledge.

1. INTRODUCTION

Data Science is many field of study with Computer/ IT, Mathematics/Statistics and Business need Domain Knowledge [1]. The three domains separately result in a variety of careers as Software, Research and Machine learning with these areas Data Scientist can maximize their performance interpreting data and providing innovative solution and new improvements in prediction [2]. Machine learning is the field of intersecting computer science, mathematics and statistics. It is used to identify patterns, recognize behaviors, and make decisions from data with minimal human intervention [3]. It is a method of data analysis that automates data collection, data preparation, feature engineering, training the model, and eventually model evaluation and prediction [3]. Machine learning and data scientists implement very complex models such as neural networks or support vector machines and an ensemble of simple models for random forests and decision trees [4]. Machine Learning (ML) Machine learning is a branch of artificial intelligence that focuses to creating algorithms data to improve their performance on a given data set [5]. Traditional machine learning focuses on using pre-set statistical methods to find the value of data in data analysis [6]. The goal of machine learning in large data environment is to search out specific rules that hidden behind dynamic, changeful, multi-source heterogeneous data and finally maximize the data value [7]. We consider data science to refer to scientific new process algorithms and systems used to extract meaning and insights from data and machine learning to refer to techniques used by data scientists to learn from data [8]. The most widely used method in dimensionality reduction is principal component analysis (PCA) PCA is a simple method that finds the directions of greatest variance in the dataset and represents each data point by its coordinates each of these directions [9]. The findings of this research is provide a comprehensive understanding of the current trends and future implications of ML on society which

will be useful for policymakers researchers and practitioners in making informed decisions about the development and use of this technology [10].

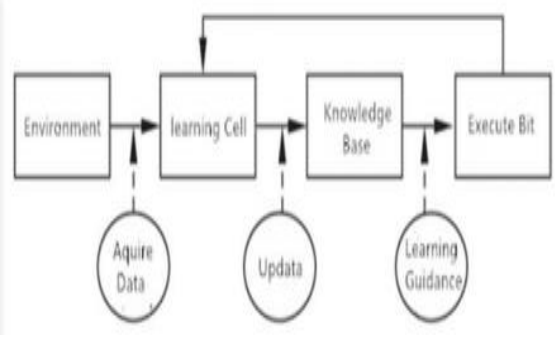

FIG. 1. Structure Diagram of Machine Learning

2. RELATED WORK

Data science and the machine learning pipeline are common terms for the series of interconnected phases that make up the data science and machine learning processes [11]. The particular order of processes depends on the nature of available data but below is a high-level overview [11] there are many such courses developed by physicists outside of HEP for their respective physics departments from which we could learn a great deal in the spirit of interdisciplinary collaboration [12]. Neural networks also called artificial neural networks are models for classification and prediction [13]. Neural network algorithms are inherently parallel. Parallelization methods are used to speed up the computation process. This can lead to discrimination and unfair treatment of certain groups of people. Additionally, the increasing use of ML in decision-making raises concerns about transparency and accountability [13]. A recent report by the Algorithmic importance of ensuring that ML models It consists of set of Algorithms used to analyzes large chunks of Data Analysis and makes data prediction in real time without human aid. A Data model is built automatically using Machine learning procedure and the system are been trained for real time prediction [14]. There are many algorithms for machine learning classification, such as decision tree, naive Bayesian classification algorithm, support vector SVM and artificial neural network.

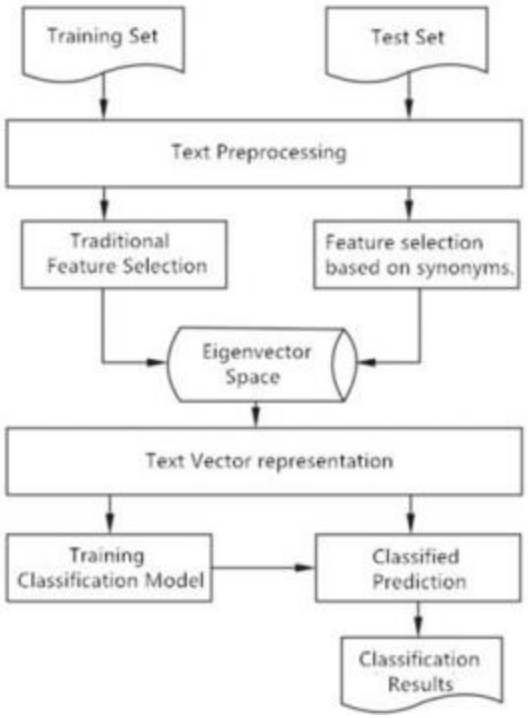

FIG. 2. Classification task flow chart in machine learning

3. PROCESS OF DATA SCIENCE MACHINE LEARNING

The data science and the machine learning process are common terms for the series of interconnected phases that make up the data science and machine learning processes [15]. The particular order of these processes depends on the nature of the project and the available data, Big data as a new hotspot industry needs to be equipped with a set of relatively scientific reasonable machine learning algorithms is classify data effectively decrease the difficulty of data processing analysis to further improve the ability of machine learning to constantly adapt to the needs of society [16].Representation learning algorithms supervised learning techniques to achieve high classification accuracy with computational efficiency. They transform the data while preserving the original characteristics of the data to another domain so that the classification algorithms is improve accuracy reduce computational complexity and increase processing speed [17].

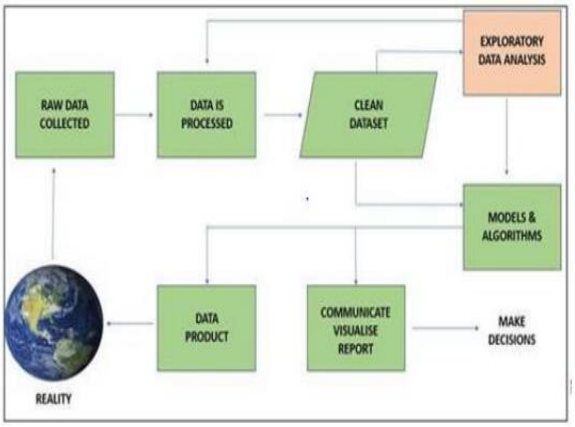

Figure 3: Process of Data Science

4. RESEARCH METHODOLOGY

The real-world examples of ML are currently being used and its impact on society. These case studies will be selected from various industries such as healthcare, finance, transportation, and manufacturing [18]. They will be chosen based on their relevance to the research objectives and the availability of data. The will focus on attitudes toward ML concerns about the technology and opinions on the potential impact of ML on society. The survey data will be analyzed using statistical methods, such as descriptive statistics and inferential statistic [19]. The findings of this research will be useful for policymakers, researchers, and practitioners in making informed decisions about the development and use of this technology.

Fig. 4. Working Process.

FEATURE ENGINEERING

Machine learning fits mathematical notations to the data in order to derive some insights. A feature is generally a numeric representation of an aspect of real-world data. Mathematical formulas work on numerical quantities and raw data exactly numerical. Feature Engineering is the way of extracting features from data and transforming them into formats that are suitable for Machine Learning algorithms [20].

STEP1: Feature Selection There are certain features is important than other features to the accuracy of the model. The methods of Feature Selection are Chi-squared test correlation coefficient scores, LASSO, Ridge regression.

STEP2: Feature Transformation It means transforming our original feature to the functions of original features. Scaling, discretization, binning and filling missing data values are the most common forms of data transformation.

STEP3: Feature Extraction The data to be processed through an algorithm is too large it's generally considered redundant. Analysis with a large number of variables uses a lot of computation power and memory. It is a term for constructing combinations of the variables for tabular data we use PCA to reduce features [21].

MACHINE LEARNING ALGORITHM

The dataset can be classified under 4 major categories:

• Classification
• Regression
• Clustering
• Time Series Analysis

STEP1: Classification is used want to find which category the data belongs to Support Vector Machines, Neural Networks, Naive Bayes, Logistic Regression, and the K Nearest Neighbor

STEP2: Regression Regression works on the Curve-Fitting Techniques intercept formula as "y=mx+c" where the slope value of y when x=0. The data points fall in the curve are used to predict the output values.. Regression Algorithms are Linear Regression, Perceptron, and Neural Networks.

STEP3: Clustering is to group the data based on the similar characteristic, without labels. Ideally, the similar data points are grouped together in the same cluster based on different definitions of similarity. Regression and Classification come under the Supervised Learning Model of Machine Learning while Clustering comes under the Unsupervised Learning Model.

STEP4: Time Series Analysis: A time-series contains sequential data mapped market forecasting use time series analysis. Broadly specified time series models are Autoregressive (AR), Integrated (I) Moving Average (MA) and some other models are the

combination of these models such as Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) models [22].

DOCUMENTATION AND KNOWLEDGE SHARING

Preprocessing to model specifications to final findings in your documentation the team might benefit from knowing this information. Data science and machine learning are iterative processes and this fact should not be overlooked [23]. As you learn more about the project or obtain new insights, you may find that you need to go back and adjust prior decisions.

Many data science make use of machine learning algorithms ti introduction to many widely-used machine learning methods in the field of data science. Based on the specifics of the problem they were created to solve many classes of algorithms [24].

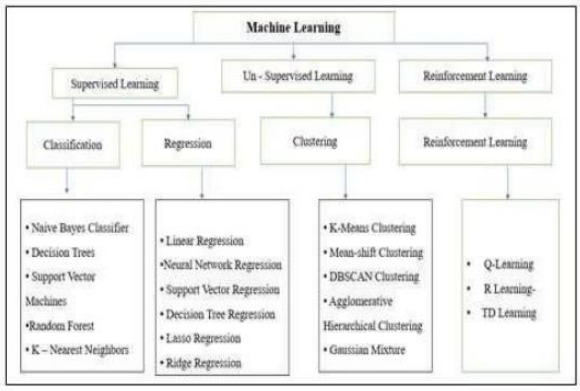

Figure 5: Machine Learning Algorithms

5. RESULT AND DISCUSSION

The model can be deployed across a range of different environments and will often be integrated with apps through API. Deployment is a key step in an organization gaining operational value from machine learning. These are fundamental abilities the relative weight of which can shift based context. Professional data scientists and machine learning experts typically possess a range of these abilities and hone them over their career. The survey results indicate that the majority of respondents (60%) are somewhat familiar with the concept of Machine Learning (ML) (Q1). This suggests that the majority of respondents have a basic understanding of ML but may not have a deep understanding of the technology and its implications. When asked about the potential benefits of ML the majority of respondents (70%) believe that ML has the potential to benefit society

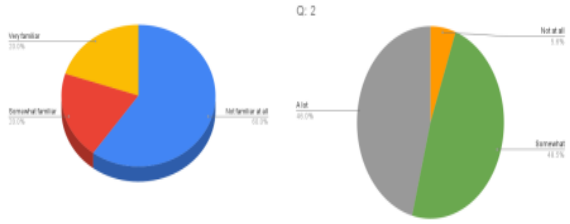

Fig. 6. Survey result

6. CONCLUSION

Data Science and Machine Learning is going to be used prominently to analyze a humongous amount of data. Data Scientists must be equipped with in-depth knowledge of Machine Learning to boost their productivity. The recent developments in the field of machine learning have brought about significant transformations, profoundly impacting many aspects of our lives and professional endeavors. machine learning algorithms which can classify data effectively, decrease the difficulty of data processing analysis to further improve the ability of machine learning, to constantly adapt to the needs of society. Tree pruning is performed to remove anomalies in the training data due to noise or outliers. Logistic regression is computationally inexpensive, but it is prone to under fitting and may have low accuracy . Additionally stronger regulations and laws to protect the privacy and rights of individuals when it comes to ML should be developed. Transparency and accountability in ML decision-making processes should be increased. Finally public education and awareness about ML should be enhanced to increase understanding of the technology and its potential impact on society.

7. FUTURE DIRECTIONS AND RECOMMENDATIONS

We describe some future directions based on the experiences to described that would help HEP researchers interested in physics education that includes data science and machine learning pedagogy file. We must optimize the traditional machine learning algorithms make it have strong vitality in the era of big data, it will need more in-depth studies of

machine learning to deal with huge data information and get the useful information in large data.

REFERENCE

[1] Cao, L.: Data science: a comprehensive overview. ACM Computing Survey (2017). https://doi.org/10.1145/3076253

[2] Matthew J. Graham. 2012. The art of data science. In Astro statistics and Data Mining, Springer Series in Astro statistics, Vol. 2. 47–59.

[3] Donoho, D.: 50Years of Data Science. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience .pdf (2015)

[4] Dyk, D.V., Fuentes, M., Jordan, M.I., Newton, M., Ray, B.K., Lang, D.T., Wickham, H.: ASA Statement on the Role of Statistics in Data Science. http://magazine.amstat.org/blog/2015/10/01/asastatement-on-therole-of-statistics-in-data-science/ (2015)

[5] Nathan Yau. 2009. Rise of the Data Scientist. Retrieved from http://flowingdata.com/2009/06/04/rise-of-thedata-scientist/

[6] Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science 349(6245), 255–260 (2015).

[7] Guo, P.: Python is Now the Most Popular Introductory Teaching Language at Top U.S. Universities, July 2014. http://cacm.acm.org/blogs/blog-cacm/176450-python-is-now-themost-popular-introductory-teaching-language-at-top-u-s - universities

[8] McKinney, W.: Python for data analysis. O'Reilly (2012).

[9] Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. (2016). Exploratory Data Analysis. 10.1007/978- 3-319-43742-2_15.

[10] Galli, Soledad. (2021). Feature-engine: A Python package for feature engineering for machine learning. Journal of Open-Source Software. 6. 3642. 10.21105/joss.03642.

[11] Behera, Rabi & Das, Kajaree. (2017). A Survey on Machine Learning: Concept, Algorithms and Applications. International Journal of Innovative Research in Computer and Communication Engineering. 2.

[12] Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. Big data, 1(2), 85-99.

[13] Aggarwal, C. C. (2011). An introduction to social network data analytics(pp. 1-15).SpringerUS.

[14] Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A.,& Caporaso, J.

[15] G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.Nature biotechnology, 37(8), 852-857.

[16] Aggarwal, C. C. (2011). An introduction to social network data analytics(pp. 1-15).SpringerUS.

[17] Hayashi, C., Yajima, K., Bock, H., Ohsumi, N., Tanaka, Y., & Baba, Y. (1996). ―DataScience, Classification, and Related Methods.‖ N.p.: springer.

[18] Donoho, D. (2015). ―50 years of Data Science.‖

[19] Data Science Association. (2020). ―About Data Science.‖ In . (Ed.).

[20] O'Neil, C., & Schutt, R. (2013). ―Doing Data Science.‖ N.p.: O'Reilly Media, Inc.

[21] S. Shukla Shubhendu and J. Vijay, Applicability of artificial intelligence in different fields of life, International Journal of Scientific Engineering and Research 1 (2013) 28.

[22] "NSF Research Experience for Undergraduates program." https://www.nsf.gov/crssprgm/reu.

[23] "International Masterclasses hands-on particle physics."

[24] Hertzmann A, Fleet D. Machine Learning and Data Mining Lecture Notes. Computer Science Department, University of Toronto. 2010.

[25] Karatzoglou A. Machine Learning in R. Workshop, Telefonica Research, Barcelona, Spain. 2010 December 15.

[26] Viña A. Data Virtualization Goes Mainstream, White Paper, Denodo Technologies, 2015.

[27] Curry E, Kikiras P, Freitas A. et al. Big Data Technical Working Groups, White Paper, BIG Consortium, 2012.

[28] Suthaharan S., Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning, Performance Evaluation Review, 41 (4), March 2014, 70-73.

[29] S. Hido, S. Tokui, S. Oda, Jubatus: An Open Source Platform for Distributed Online Machine Learning, Technical Report of the Joint Jubatus project by Preferred Infrastructure Inc., and NTT Software Innovation Center, Tokyo, Japan, NIPS 2013

Workshop on Big Learning, Lake Tahoe. December 9, 2013. Pp. 1-6.
[30] C.L. P. Chen, C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences, Vol. 275, No. 10, pp. 314-347, 2014