IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 9, September 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

Data Science: The Way How to Use Data

Srinjoy Saha[1], Shreya Bhar[2], Sukalyan Porey[3], Kusum Majhi[4], Debrupa Pal[5]

U.G. Student, Department of Computer Application, Narula Institute of Technology, West Bengal, India[1]

U.G. Student, Department of Computer Application, Narula Institute of Technology, West Bengal, India[2]

U.G. Student, Department of Computer Application, Narula Institute of Technology, West Bengal, India[3]

U.G. Student, Department of Computer Application, Narula Institute of Technology, West Bengal, India[4]

Assistant Professor, Department of Computer Application, Narula Institute of Technology, West Bengal, India[5]

ABSTRACT: Data Science, a new discovery illustration, is potentially one of the most significant advances of the early 21st century. Derived from scientific discovery, it is being applied to every human effort for which there is sufficient data. Significant and remarkable successes have been achieved; even greater claims have been made. Along with benefits, challenge and risks are abound. The science underlying data science has yet to emerge. This claim is based on observing the centuries-long developments of its predecessor paradigms - empirical, theoretical, and Jim Gray's Fourth Paradigm of Scientific Discovery (Hey, Tansley& Tolle, 2009) (aka eScience, data-intensive, computational, procedural). This paper is mainly focuses on essential questions for data science- what is data science, importance of data science in our everyday life, data wrangling, data science algorithms.

KEYWORDS: Data Science, AI (Artificial Intelligence), Data Wrangling, Data Science Algorithms.

I. INTRODUCTION

Data science is all about gathering data, analysis and making decisions. It is an interdisciplinary research field which uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. It unites all the information, statistics, data analysis, and their related methods in order to understand and analyze the actual program with data. Data science contains of various types of techniques and theories which are taken from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. People who willing to work in those mentioned fields and also good at these with enough knowledge called themselves as data scientist [1]. Data scientists create various types of programming codes and combine them with statistical knowledge to create insights from data. It's not an easy thing to become a data scientist but not impossible too.

II. CATEGORIES OF DATA

There are two categories of data and each of them have more two sub-categories [2]. Those are –

Qualitative or Categorical Data: -

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person- male, female, or others, is qualitative data [3].

Two sub-categories of Qualitative data, are -

Nominal Data: The name "nominal" comes from the Latin name "nomen," which means "name." Nominal Data is used to label variables without any order or quantitative value. The colour of hair can be considered nominal data, as one colour can't be compared with another colour.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

Ordinal Data: Ordinal data have natural ordering where a number is 5e5ee5e5e55eq present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them [4].

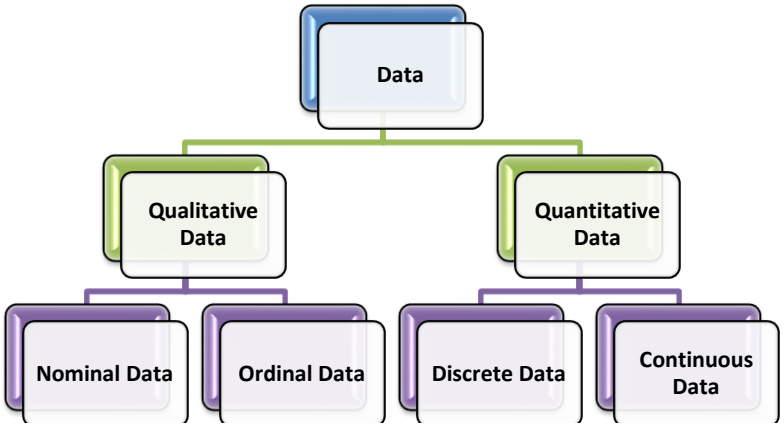

Figure 1: Categories of Data

Quantitative Data: Quantitative data can be expressed in numerical values, which makes it countable and includes statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like, "how much," "how many," and "how often." For example, the price of a phone, the computer's ram, the height or weight of a person, etc., falls under the quantitative data.

Two sub-categories of Quantitative data, are -

Discrete Data: The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

Continuous Data: Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range [5].

III. DATA WRANGLING

In Data Science, a data wrangling process, also known as a data munging process, consists of rearranging, modifying and surveying data from one raw material form into another in order to make it more usable and valuable for a kind of subsequential uses including analytics.
It can be defined as the process of wiping out, arranging, and modifying raw data into the desired format for analysts to use for decision-making. In business, data wrangling helps to perform complex tasks in much less time, generate correct results and help in taking correct decisions. The exact process varies from project to project depending upon the data and the goal have to achieve.

The Importance of using Data Wrangling:

- Properly wrangled data (have to be usable data) ensures that quality data has been inserted into subsequential analyses.
- We have to save all the data from different sources in one centralized location so that we can use further.
- Integrating raw data into the required format and understanding the business context of data.
- To clean the data from the noise or flawed, missing elements.
- Helping business users make concrete, timely decisions.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

IV. DATA SCIENCE ALGORITHMS

Linear Regression: In data science algorithm, Linear Regression is a method of measuring the relationship between two continuous variables. The two variables are –

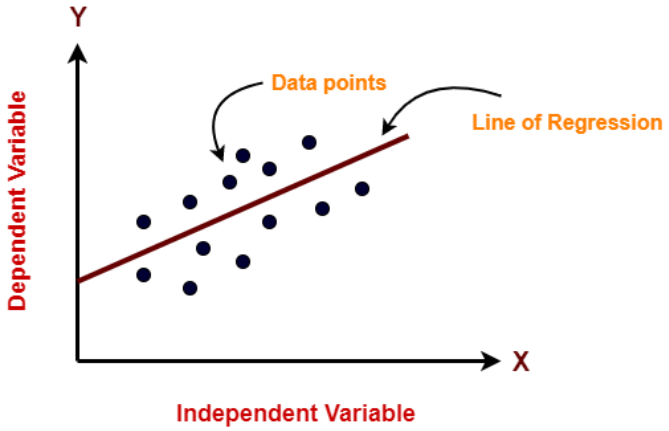

Figure 2: Diagram of Linear Regression

Independent Variable – "**x**"
Dependent Variable – "**y**"

In this case of a simple linear regression, the independent value is the predictor value and it is only one. The relationship between x and y can be described as:
y = mx + c
Where m is the slope and c is the intercept.

Logistic Regression: There is another algorithm for data science. Logistic Regression is used for binary classification of data-points. It performs hierarchical classification whose results belong to either of two classes (1 or 0). An example of logistic regression is predicting whether or not it will rain based on weather condition.

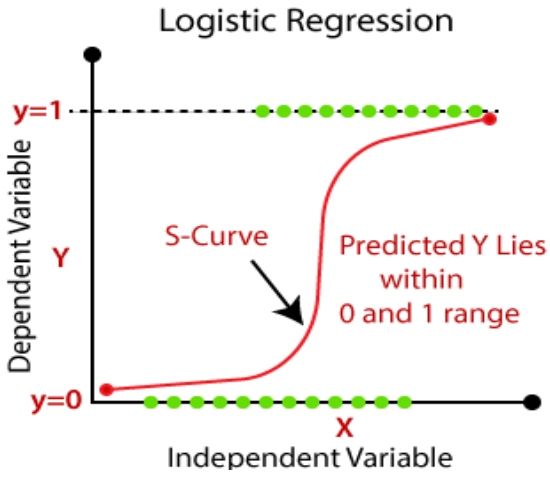

Figure 3: Diagram of Logistic Regression

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

The two important parts of Logistic Regression are–

- Hypothesis
- Sigmoid Curve

Generating from our 'Hypothesis', we fit the data with a log function that ultimately produces an S-shaped curve called a 'Sigmoid'. Based on this log function, we can determine the category of the class.
We generate this with the help of logistic function –
1 / (1 + e^-x)
Here, **e** represents base of natural log and we obtain the S-shaped curve with values between 0 and 1. The equation for logistic regression is written as:
y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x))
Here, b0 and b1 are the coefficients of the input x. These coefficients are estimated using the data through "maximum likelihood estimation" [6].

K-Means Clustering: In this algorithm, K-means clustering is a recursive algorithm that partitions a group of data containing n values into k subgroups. Each of the n value belongs to the k cluster with the nearest mean.
K-means clustering is the most popular form of an unsupervised learning algorithm. It is easy to understand and implement.

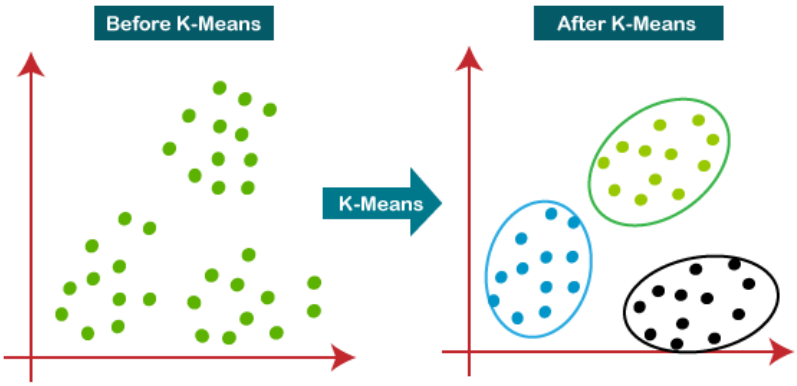

Figure 4: Diagram of K-Means Clustering

The objective of the K-means clustering is to minimize the Euclidean distance of each point from the centroid of the cluster.

The algorithm for K-means clustering –

- At first, we start randomly initialize and select the k-points. These k-points are the means.
- Using the Euclidean distance to find data-points that are closest to their centerr of the cluster.
- Then we calculate the mean of all the points in the cluster which is finding their centroid.
- We recursively repeat step 1, 2 and 3 until all the points are assigned to their respective clusters.

Support Vector Machines: In data science algorithm, Support Vector machines are the powerful classifiers for classification of binary data. They are also used in facial recognition and genetic classification.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

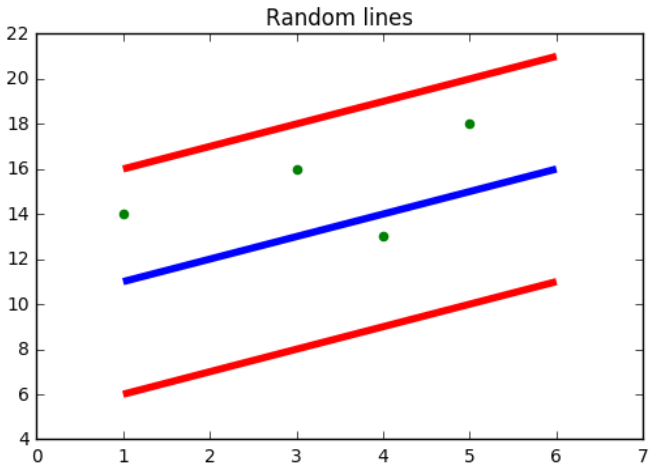

Figure 5: Diagram of Support Vector Machines

Support Vector Machines is able todepict the input vectors to n-dimensional space. SVM's are formed by structure risk minimization.

Artificial Neural Networks: There is another algorithm for data science called Artificial Neural Networks. These are modeled after the neurons in the human brain. It consists of many layers of neurons which are structured to transmit information from the input layer to the output layer and there are hidden layers presentbetween the input and the output layer.

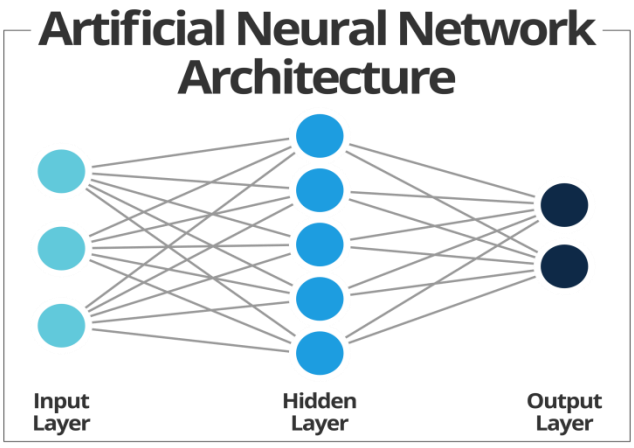

Figure 6: Diagram of Artificial Neural Networks

For example, given the images of cats and dogs, our hidden layers perform various mathematical operations to find the maximum probability of the class our input image falls in. This is an example of binary classification where the class, that is, dog or cat, is assigned its appropriate place.

Decision Trees: In data science, with the help of DecisionTrees, we can perform both prediction and classification. Using Decision Trees, we are able to make decisions with a given set of input.
For example, suppose someone goes to the market to buy a product. First, he/she assesses if he/she really needs the product, that is, he/she will go to the market only if he/she does not have the product. After assessing it, he/she will determine if it is raining or not [8].

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

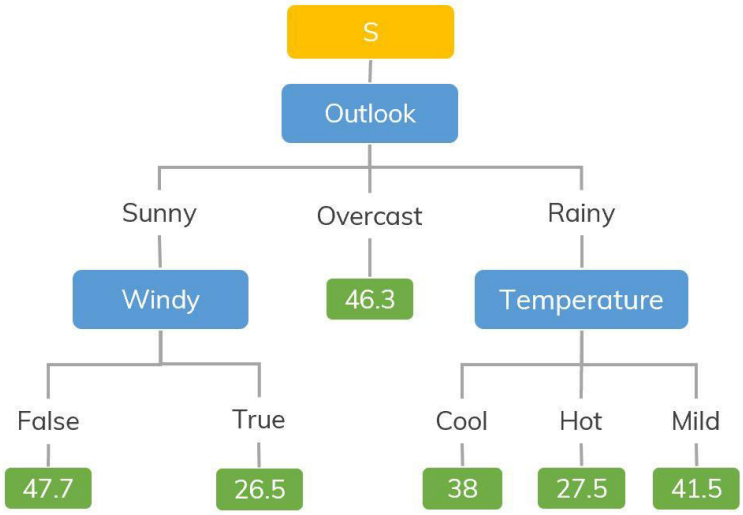

Figure 7: Diagram of Decision Tree

Using the same principle, we build a hierarchical tree to reach a result through a process of decisions. There are two steps to build a tree:

- Induction: Induction is the process in which we build the tree.
- Pruning:In pruning, we simplify the tree by removing complexities.

V. IMPORTANCE OF DATA SCIENCE

Over the past few years, data science has overtaken nearly every industry on the planet. It helps in facilitating the organizations with the potential to process large amount of data. As a result, data science has become a source of energy for business. Also, the role of data science is being leveraged for reducing traffic accidents, law enforcement and much more. So, that's why the contribution of data science is really important in our modern era [7].
Here's how we experience the role of data science in our daily life –

Entertainment: Data science algorithms are also used for entertainment purpose. Because of this, the entertainment platforms are growing up day by day. In the part of personalized marketing, real-time analytics, object detection and classification, these algorithms play a significant role. For example, we can take the name of Netflix and YouTube, both are relied on data to determine what shows are greenlit and promoted [8].

Internet Browsing: All search engines like Safari, Firefox, Opera, Google etc. leverage data science algorithms to deliver the best result for our searched query in fraction of seconds. If this tech didn't exist, these search engines wouldn't have been the same we know today.

Healthcare Sector: The healthcare industry is transforming rapidly, thanks to the implementation of data science technologies. To file, store and manage patient records, data science plays a vital role. Data science is also used for solving the problem of image analysis, genetics and genomics, drug development in the medical sector. We can also make virtual assistants and health bots with the help of data science and AI.

E-Commerce: Various e-commerce websites like Amazon and Flipkart use data science algorithms for a better customer experience. When someone is looking to buy, these sites can offer popular recommendations or something interesting based on user's search history. With this tech, these platforms can enhance the user experience and increase their selling [9].

Face and Speech Recognition: In the field of face and speech recognition data science and AI plays a vital role. Some speech recognition products like Google Voice, Siri, Cortana etc. rely on data science for producing the best result for

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 9, September 2022||

| DOI:10.15680/IJIRSET.2022.1109010 |

us. These days, concept of data science is also used for face recognition by consuming massive amounts of data about what's a face or not, what's a crack on something and what's a smile, just to name a few.

Digital Marketing: Data science identifies, collects, segments and interprets the data in digital marketing. Using data science algorithms digital marketers can target digital advertisements based on user's past behaviour, which is probably the main reason why digital marketing field is growing day by day. Basically, data science gives digital marketers the knowledge they need to succeed [9].

VI. CONCLUSION

In education, Data Science is into a constructive stage of development. It started to evolve from past few years and now producing professionals with complementary and remarkable skills in computer, information and statistical science. Industrial demand on Data Science is growing day by day and also students are getting interest in it. There will be a huge number of students taking courses in data science as the value of data skills becomes even more popular and acknowledged. Now a days Data Scientists have become a major part of many industries as they enrich the researches and development. In future there will be more holistic and nuanced roles for data scientists in industry. Academic institutions should comprehend data science as an important new field and should provide an evolve range of educational pathways to prepare students for data science roles in the workplace.

REFERENCES

[1] Igual, L., &Seguí, S. (2017). Introduction to data science. In Introduction to data science (pp. 1-4). Springer, Cham.

[2] Timbers, T., Campbell, T., & Lee, M. (2022). Data science: A first introduction. CRC Press.

[3] Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publications.

[4] Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

[5] Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., & Li, L. (2016). Big data analytics and big data science: a survey. Journal of Management Analytics, 3(1), 1-42.

[6] Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. Information systems research, 25(3), 443-448.

[7] Goyal, D., Goyal, R., Rekha, G., Malik, S., & Tyagi, A. K. (2020, February). Emerging Trends and Challenges in Data Science and Big Data Analytics. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-8). IEEE.

[8] Cady, F. (2017). The data science handbook. John Wiley & Sons.

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details