An Overview of Data Science Methodologies and Instruments

Dr E. SriDevi[1] **,**

Asst. Professor, Department of C.S.E, Koneru Lakshmaiah Education Foundation, Guntur,
A.P, India – 522502.

V.PremaLatha[2]

Assoc. Professor, Department of C.S.E, Koneru Lakshmaiah Education Foundation, Guntur,
A.P, India – 522502.

Abstract:

The term "data science" is relatively new. We had statisticians before we had data science. These statisticians were proficient in qualitative record evaluation, and companies employed them to investigate their average revenue and performance. The field of computer science and information merged with the introduction of computing techniques, cloud storage, and analytical tools. The science of statistics was born out of this. With a broad range of applications across all industries and organizations, data science is a rapidly expanding field of study.Many scientific approaches, including machine learning, artificial intelligence, statistics, and mathematics, are combined in data science to solve complex problems. It uses data that has been analysed and predictions based on it to provide a variety of information on new trends and patterns in a particular model. The goal of this paper is to give a general overview of data science techniques and open source data science tools.

Keywords : Data, Data mining, Machine Learning, Data science, Open source, Data science Tools.

INTRODUCTION:

With a broad range of applications across all industries and organizations, data science is a rapidly expanding field of study. Many scientific approaches, including machine learning, artificial intelligence, statistics, and mathematics, are combined in data science to solve complex problems. It uses data that has been analysed and predictions based on it to provide a variety of information on new trends and patterns in a particular model.

ISSN PRINT 2319 1775 Online 2320 7876

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

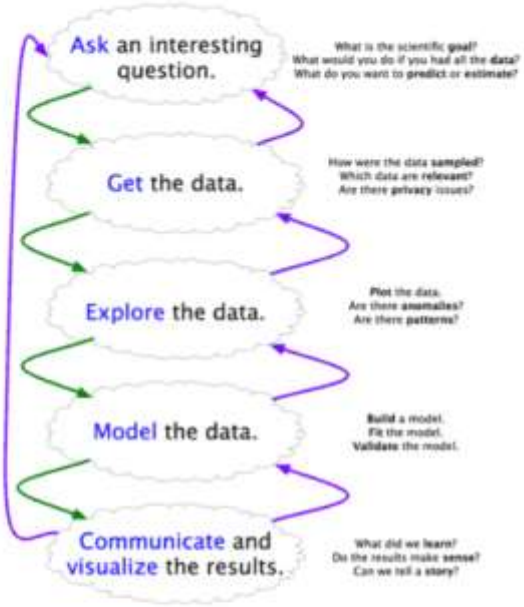
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Fig -1: Data Science WorkFlow

The goal of this paper is to give a general overview of data science techniques and open source data science tools. Thus, we can conduct research and extract valuable information from this massive data set. It takes specialised knowledge to be a "data scientist," who will explore and analyse data using a variety of machine learning algorithms and statistical techniques. A specialised individual in data science, known as a data scientist, uses machine learning algorithms in addition to data analysis to predict future events. It is a synthesis of computer science, statistics, and mathematics as three distinct disciplines.

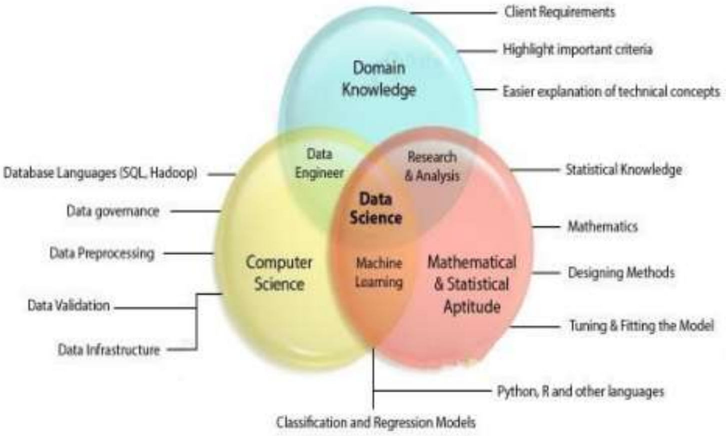

Fig -2: Data Science Overview

ISSN PRINT 2319 1775 Online 2320 7876

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

2. OBJECTIVE OF DATA SCIENCE

The most important thing is to use data effectively in order to meet the ever-growing business needs of individuals' lives. Making corrections to the errors or data manipulation that were shown in earlier projects is another important priority [3]. The main goal of data science is to identify intriguing patterns in data. Therefore, in order to identify patterns in the data, a data scientist must carefully examine the data using a variety of statistical techniques, such as data extraction, wrangling, and pre-processing, in order to analyse and derive insights from the data. Following that, they will use the data to make predictions. A data scientist's primary goal is to draw insightful conclusions from the data. Businesses can use these findings to make more informed business decisions. In order to make better decisions, data science is predicted to make significant contributions to a number of fields, including applied computing, medical sciences, professionals & social life activities, computing paradigms, data management systems, and many more[4].

3. TECHNIQUES FOR DATA SCIENCE

Depending on the kind of data and volume of data collected, several data analysis techniques are available. The following types of techniques are classified based on the type of data to be analysed:

1. Methods grounded in statistics and mathematics
2. Methodologies founded on machine learning and artificial intelligence
3. Methods utilising Graphs and Visualisation

3.1 Mathematics and Statistics Techniques forData Science:

- ***Descriptive Analysis:*** This type of analysis describes performance using a selected benchmark and is based on historical data and key performance indicators. It examines historical patterns and their potential impact on performance in the future.
- ***Dispersion Analysis:*** A dispersed data set serves as the foundation for dispersion. Data analysts can choose the variety of the elements under investigation with this method.

  Regression analysis examines the connection between one or more independent variables and a dependent variable. Regression analysis can be done using a variety of algorithms, including logistic, non-linear, logistic multiple, linear, and more.

- ***Factor Analysis:*** This method can be used to ascertain how a group of variables relate to one another. It will outline the additional variables or factors that define the patterns in the interactions between the initial variables. Procedures involving clustering and classification benefit from factor analysis.
- ***Discriminant analysis*** : This is a crucial classification technique that uses variable measurements to identify distinct features on various groups. Put another way, it pinpoints the key characteristics that separate the two groups from one another.
- ***Time Series Analysis***: This kind of analysis provides us with a set of prepared statistics known as time collection because measurements are spaced across time.

3.2 Artificial Intelligence and Machine Learning Techniques for Data Science

- ***Artificial Neural Networks:*** A neural network is a programming paradigm inspired by biology that provides a brain metaphor for information processing. An Artificial Neural Network system modifies its structure based on the information that flows through the network. ANNs are very accurate even when dealing with noisy data. ANNs are the foundation for the majority of classification and forecasting applications in business.

- ***Decision trees:*** Based on a structure resembling a tree, decision trees represent regression or classification models. It creates a decision tree by dividing a data set into subsets and organising them based on their relationships.
- ***Evolutionary programming***: Combining several different evolutionary algorithms for data analysis is called evolutionary programming. This method is domain-independent, capable of exploring a sizable search space and effectively handling characteristic interaction.
- ***Fuzzy logic***: Data analysis based on probability, or fuzzy logic, is used to manage the uncertainties in data mining methods.

3.3 Graphs and Visualization Techniques of Data Analysis for Data Science

- Column and Bar Charts: Charts and bar charts show the numerical differences between categories. To reflect the variations, the column chart reaches the top of the columns. Within the bar chart's case, axes switch places.
- Line Chart: Line charts show how data changes over a continuous period of time.
- Area Chart:The area chart is derived from the line chart. Additionally, it represents better trend data by adding colour to the area between the axis and the polyline

ISSN PRINT 2319 1775 Online 2320 7876

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

- Pie Chart: Pie charts show the percentage of various classifications. This represents a single series of data. However, it can display the percentage of data in various categories in a multi-layered format.
- Funnel Chart:The stages are represented by a funnel chart.Each module's proportion and size are correspondingly reflected. Rankings will also be compared.
- Word Cloud Chart: This graphic depicts textual content details. It requires a significant amount of data,and the level of prejudice wants to be too high for users to comprehend which is the best-performing one. It's not a exact and accurate analytical method.
- Gantt chart: It shows the actual timing as well as the growth of interest in assessing the requirements.
- Radar Chart: This tool is employed to assess multiple quantized graphs. It illustrates which informational variables possess higher values and those with lower values. An

radar chart is employed to assess collection and categorization in addition to a proportionate illustration.

- A scatter plot displays the way variables are distributed across a square coordinate system as factors. The correlation between the variables can be seen in the distribution within the information factors.A bubble chart is a scatter plot in miniature. Here, the area of the bubble indicates the third value in a manner akin to the x and y coordinates.
- Gauge: It resembles a chart that has materialised. In this case, the pointer denotes the dimension and the dimensions the metric. It is a suitable method for representing interval comparisons.
- Frame Diagram: It employs an inverted tree structure as a hierarchical visual representation.
- Rectangular Tree Diagram: This method is used to simultaneously show hierarchical relationships degree. It depicts the proportion and makes effective use of space making use of a rectangle.

Map

- Regional Map: The value is represented by colour distribution across a divide in a map.
- Point Map: The data's geographic distribution shown as of points on an earthly backdrop is depicted by with a point map. The same-sized points have no bearing on a single data set, but the size of the data is represented by bubbled points in every area.
- Flow Map: A flow map shows the locations of the inflow and outflow. It is employed to denote a line joining the spatial elements' geometric centres of gravity. It will also lessen visual distractions.

- Heat Map:A heat map is a graphic representation of weighted points within a region. Density is represented by the colours.

4. TOOLS FOR DATA SCIENCE

Data scientists' primary responsibility is to make decisions by handling and analysing large amounts of structured and unstructured data. Many big data technologies have been developed and categorised into data processing concepts, as mentioned in paper [5].Therefore, programming languages and tools are required for data scientists to handle such a large amount of data in order to analyse it and complete their work properly. We will look at a few data science tools in this post that are helpful for producing predictions and data analysis. The available Data Science tools are summarised in Table 1.

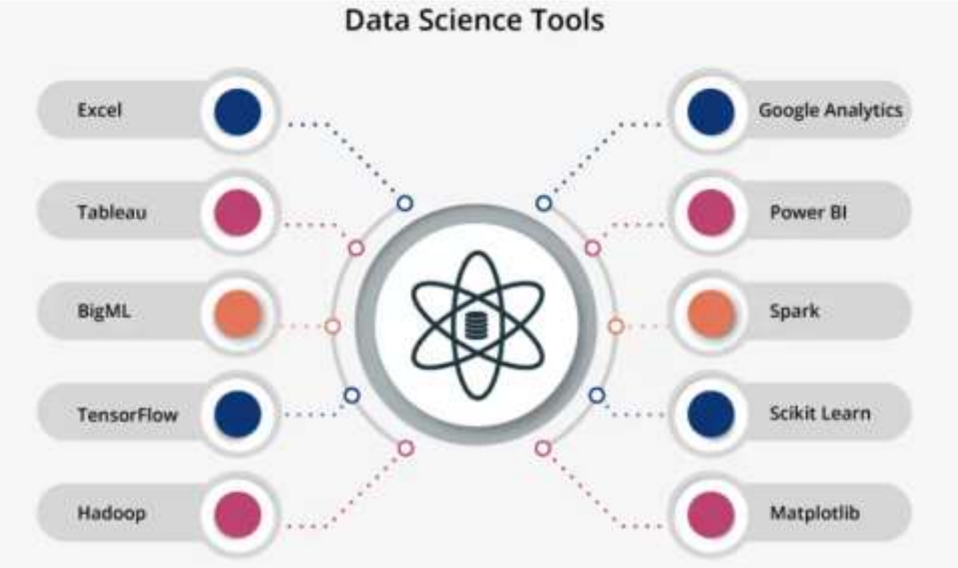

Fig -3: Data Science Overview

The process of gathering and transforming data into a desired and useful form is known as data processing. All that is involved in the operation is processing, which must be authorised manually or automatically according to a set sequence of steps. Data processing tools are required because manual data collection and processing took a lot of time in the past.

A data scientist is a person who primarily uses data to transform it into a valuable or treasured form through logical progression [9].The massive advancement of multiple data forms requires the data scientist to operate the data at multiple or different levels, including data loading, data cleaning, data modelling, data processing, and data evaluation. Since the data is collected from a variety of fields, it is imperative to use the development of skills in a variety of fields, including biotechnology, artificial intelligence, machine learning, robotics, statistical approaches, analytical methods, medical sciences, mathematical procedures, and the Internet of Things.

The viewpoints of data scientists regarding organisations are as follows:

- Making efficient use of data to expand business Developing appropriate methods for gathering data from multiple sources
- Data cleaning
- Data processing and evaluation
- Appropriate A.I algorithms must be implemented
- Include algorithms for deep machine learning
- Create analytical, statistical, and logical reasoning techniques

These days, data science is a required field that bridges several academic disciplines, including mathematics, statistics, mathematical methods, logical reasoning, intelligence algorithms, and machine learning applications. These fields are all related to accessing and effectively using data from different businesses or organisations. When data is used effectively, appropriate decisions can be made to expand the business based on the preferences and satisfaction of the customers. Thus, we may draw the conclusion that as data science becomes more popular, more data scientists will be needed in order for each company to expand. Finally, we turn our attention to the development of successful careers in data science. The primary charm of this industry was its ability to expand all businesses.

5. CONCLUSIONS

By the end of this article, we can say that data scientists have access to a variety of tools and techniques for handling tasks related to data analysis.In order to prepare, analyse, and visualise data in order to create predictive models using various statistical and machine learning algorithms, data scientists need a variety of tools, which we have covered in this article. Many data science tools can carry out intricate data science operations within a single framework, making it simple to implement data science functionalities even for non-programmers.

REFERENCES

[1] Russell, Stuart J., and Peter Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,, 2016.

[2] Nicolae, Bogdan, et al. "Park, Yoonho. Leveraging Adaptive I/O to Optimize Collective Data Shuffling Patterns for Big Data Analytics. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. PP (99) pp: 1-13." (2014).

[3] Islam, Mohaiminul. "Data Analysis: Types, Process, Methods, Techniques and Tools." International Journal on Data Science and Technology 6.1 (2018): 10.

[4] Dhar, Vasant. "Data science and prediction." Communications of the ACM 56.12 (2013): 64-73.

[5] Bejjam, Suvarnamukhi & Seshashayee, M.. (2018). Big Data Concepts and Techniques in Data Processing. International Journal of Computer Sciences and Engineering. 6. 712-714.

[6] Van Der Aalst, Wil. "Data science in action." Process mining. Springer, Berlin, Heidelberg, 2016. 3-23.

[7] Ethem Alpaydin (2004). Introduction to Machine Learning, MIT Press, ISBN 978-0-262-01243-0.

[8] Stuart Russell & Peter Norvig, (2009). Artificial Intelligence – A Modern Approach. Pearson, ISBN 9789332543515.

[9] https://data-flair.training/blogs/data-science-tools/