International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

A Survey of Data Science Techniques and Available Tools

Monali R. Baviskar[1], Priya N Nagargoje[2], Priyanka A. Deshmukh[3], Rina R. Baviskar[4]

[1]*Assistant Professor, MIT Engineering Collage, Aurangabad, Maharashtra*
[2]*Assistant Professor, MIT Engineering Collage, Aurangabad, Maharashtra*
[3]*Assistant Professor, JNEC Engineering Collage, Aurangabad, Maharashtra*
[4]*Research Student, RAIT Engineering Collage, Mumbai, Maharashtra*

Abstract - *Data science is a very recent terminology. Earlier than data science, we had statisticians. These statisticians skilled in qualitative evaluation of records and organizations hired them to research their standard overall performance and income. With the arrival of a computing technique, cloud storage, and analytical equipment, the field of Computer science merged with information. This gave birth to statistics science.*

Data science is a booming field of study which has a multidimensional scope for all organizations and industries. Data Science has lots of scientific methods which are made up of statistical techniques, machine learning, artificial intelligence and mathematics under one framework to solve the once complex problems. It gives various information on emerging trends and patterns in a specific model with the help of analyzed data, and predictions are made on that data. This paper is intended to provide an overview of techniques which are used in data science and the tools which are available as an open source for data science.

Key Words: Data, Data mining, Machine Learning, Data science, Open source, Data science Tools

1. INTRODUCTION

"Data Science is the systematic workflow of extraction, preparation, analysis, visualization, and maintenance of information. It is an interdisciplinary field which is based on scientific methods and processes to gain knowledge from raw data. "

Designing an intelligent system is conceivable by incorporating the computers with learning, processing and decision making ability [1]. With the emergence of the latest technology, there was an exponential growth in data. Data is the basic component in transformation of any individual, organizations and businesses towards development in the future era [2].So by using this big data we can research and get meaningful information from it. It requires unique knowledge of a 'data Scientist' which will use various statistical methods & machine learning algorithms to analyze and explore data. A data Scientist is a specialized person in data science, which will not only analyze the data but also makes use of machine learning algorithms for the future prediction of events. It is a Combination of three different fields that is mathematics, statistics, and computer science.

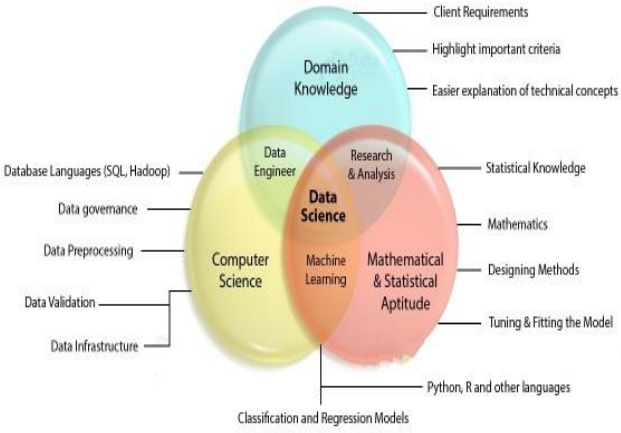

Fig -1: Data Science Overview

2. OBJECTIVE OF DATA SCIENCE

To meet the growing business needs of individuals life it is very much mandatory to make use of data in effective means is the primary concern. Another major concern is to correct the drawbacks depicted in the previous projects or mishandling of data [3]. The principal objective of Data Science is to find interesting patterns within data. So, for finding patterns, a Data Scientist must scrutinize the data thoroughly by using various statistical techniques like data extraction, wrangling and pre-processing to analyze and draw insights from the data. After doing that they will make predictions from the data. The main Objective of a Data Scientist is to make meaningful conclusions from the data. By using these conclusions, companies are able to make smarter business decisions. Data science is expected to do a lot of innovations in the areas like applied computing, medical sciences, professionals & social life activities, computing paradigms, Data management systems and many more to have a better decision making[4].

3. TECHNIQUES FOR DATA SCIENCE

Different types of techniques are available for Data Analysis depending on the type of data, and the amount of data collected. According to the type of data to be analyzed, the techniques are categorized into following types:

1. Techniques based on Mathematics and Statistics
2. Techniques based on Artificial Intelligence and Machine Learning
3. Techniques based on Visualization and Graphs

3.1 Mathematics and Statistics Techniques for Data Science:

Descriptive Analysis: Descriptive Analysis is based on the historical data, Key Performance Indicators, and it will describe the performance based on a chosen benchmark. It considers past trends and how they might affect future performance.

Dispersion Analysis: Dispersion is based on a data set which is spread. This approach permits data analysts to decide the variety of the elements under study.

Regression Analysis: Regression works on the relationship between a dependent variable and one or more independent variables. Various algorithms are available for regression analysis e.g. linear, multiple, logistic, ridge, non-linear, and more.

Factor Analysis: Factor analysis will determine the relationship between a set of variables. It will describe the other factors or variables that specify the patterns in the relationship among the original variables. Factor Analysis is beneficial for clustering and classification procedures.

Discriminant Analysis: This is one of the important classification techniques which identifies the different features on different groups based on variable measurements. In simple words, it identifies major features of two groups which make them different from one another.

Time Series Analysis: In this type of evaluation, measurements are spanned throughout time, which offers us a set of prepared statistics called time collection.

3.2 Artificial Intelligence and Machine Learning Techniques for Data Science

Artificial Neural Networks: A Neural network presents a brain metaphor for processing information which is a biologically-inspired programming paradigm . According to the information flowing on the network, an Artificial Neural Network system changes its structure. ANN can work with noisy data and are highly accurate. Most of the business applications of classification and forecasting are based on ANN.

Decision Trees: Decision trees represent classification or regression models based on tree-like structure. It subdivides a data set in subsets and according to their relation builds a decision tree.

Evolutionary Programming: It is a Combination of different evolutionary algorithms for data analysis. This technique is independent of domain, which could explore sufficient seek area and manages characteristic interaction very efficiently.

Fuzzy Logic: It is a probability based data analysis technique which will handle the uncertainties in data mining techniques.

3.3 Graphs and Visualization Techniques of Data Analysis for Data Science

Column Chart, Bar Chart: Numerical differences between categories are represented by charts and bar charts. The column chart takes to the peak of the columns to mirror the variations. Axes interchange within the case of the bar chart.

Line Chart: Changing data over a continuous interval of time is represented by Line charts.

Area Chart: Area chart is based on the line chart. It moreover fills the region among the polyline and the axis with color, for this reason representing better trend data.

Pie Chart: The proportion of different classifications is represented by Pie charts. Only one series of data is represented by this. But it can represent the proportion of data in different categories in multi-layered form .

Funnel Chart: Funnel chart represents the each stage proportion and the size of each module is reflected accordingly. It will also compare rankings.

Word Cloud Chart: It is a visible illustration of textual content information. It calls for a big amount of information, and the degree of discrimination desires to be excessive for users to understand the most outstanding one. It isn't a completely correct analytical approach.

Gantt Chart: It indicates the real timing and the development of interest in evaluation to the necessities.

Radar Chart: It is used to evaluate more than one quantized charts. It represents which variables within the information have better values and which have decrease values. A radar chart is used for evaluating classification and collection together with proportional illustration.

Scatter Plot: It shows the distribution of variables within the form of factors over a square coordinate system. The distribution within the information factors can display the correlation between the variables.

Bubble Chart: It is a version of the scatter plot. right here, similarly to the x and y coordinates, the area of the bubble represents the 3rd value.

Gauge: It is a sort of materialized chart. here the dimensions represents the metric, and the pointer represents the dimension. it is an appropriate technique to represent interval comparisons.

Frame Diagram: It uses hierarchical visual representation in the form of an inverted tree structure.

Rectangular Tree Diagram: This technique is used to represent relationships of hierarchical format at the same level. It efficiently uses space and represents the proportion using a rectangular area.

Map:

Regional Map: The Color representation is used for value distribution over a map partition.

Point Map: The geographical distribution of data in the form of points on a geographical background is represented by using Point map. Same sized points are meaningless for single data, but bubbled points represent the size of the data in each region.

Flow Map: An inflow area and an outflow area is represented by a flow map. It is used to indicate a line connecting the geometric centers of gravity of the spatial elements. It will also reduce visual clutter.

Heat Map: This represents weighted points in a geographic area. To represent density the colors are used.

4. TOOLS FOR DATA SCIENCE

The major role of Data Scientists Is to make decisions which are done by analyzing and handling lots of unstructured data and structured data. As stated in paper [5] there are numerous big data technologies that have been advanced and classified into data processing concepts.So to handle such a large amount of data programming languages and tools are needed for data scientists to analyze that data and do their work appropriately. In this article we will explore some available tools for data science which is useful for analyzing data and generate predictions. Table 1 summarizes available tools of Data Science.

5. CONCLUSIONS

At the end of this article we can conclude that there are a number of techniques and tools Available for performing data analysis related tasks by data scientists.

And for performing such a data analysis task the data scientist need various tools which we have discussed in this article for preprocessing, analyzing and visualizing data in order to make predictive models using some statistical and machine learning algorithms. Lots of data science tools can perform Complex data science operation in one framework so that it is easy to implement the functionalities of data science without having prior knowledge of coding.

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." (2020).

[3] Islam, Mohaiminul. "Data Analysis: Types, Process, Methods, Techniques and Tools." International Journal on Data Science and Technology 6.1 (2020): 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/

Table -1: Available tools for Data Science

Tool Name	Type	Features	Applications
SAS	Closed source Proprietary software	• It has strong analysis ability • It is flexible 4 generation programming languages. • It has Interactive SAS studio and Support for various types of data formats • It is available with various data encryption algorithms.	• It performs statistical modeling. • Beneficial for multivariate analysis. • Useful for creating safe drug and Clinical Research and forecasting
Apache spark	Open source software	• It is a High speed software • Good integration with the Hadoop ecosystem and data sources. • It has an advanced analytical engine. • Provides in-memory computing. • Handles real time stream processing. • It is dynamic in nature. • It has high fault tolerance.	• Used in Predictive analysis, customer segmentation And sentiment analysis • Useful for Financial, Security and Health Organization.
BigML	Open source	• It provides cloud based environment • BigML specializes in predictive modeling. • It provides interactive visualization of data with ability to export on mobile for IOT devices	• Useful for sales forecasting. • Risk Analytics. • Product innovation.
D3.js	Open source	• It is a client side scripting language Based on JavaScript. • Useful for client side interactions in IOT. • It is useful for making interactive visualizations. • It can be used with CSS.	• Useful for creating interactive web applications. • It can create animated transitions. • Implement customized graph on web pages.
Matlab	Closed source proprietary software	• It provides a numerical computing environment . • It can process Complex mathematical operations. • Matlab has a powerful graphics library. • Matlab is very useful for deep learning. • It provides easy integration with embedded systems	• Useful for creating AI systems. • Good for Image processing, signal processing , Text Analytics. • Useful for industrial decision making.

Tool Name	Type	Features	Applications
Excel	Closed source proprietary software	• It is highly popular for small scale data analysis. • It is mainly used for spreadsheet calculations and visualization. • Excel provides easy connection with SQL. • XL tool pack use for Complex data analysis.	• Data scientists use Excel for data cleansing operations. • Beneficial for business analytics.
ggplot 2	Open source	• It has advanced visualization techniques for programming language. • It allows customizing visualizations. • It allows Data scientists to create interactive graphs by using a text label to the data points.	• Beneficial for creating Complex plots.
Tableau	Open source	• It has powerful Graphics for interactive visualizations. • It has the ability to visualize the geographical data and latitude and longitude plotting. • It can do interfacing with databases, OLAP cubes and spreadsheets. • Provide subscription to others • It maintains revision history. • It has powerful Analytics tool to analyze data.	• Useful in business intelligence. • Useful for working with maps.
Jupyter	Open source	• It supports multiple programming languages like Julia, Python and R. • Web based live code writing, visualizations and presentations are possible. • It provides cleaning operations statistical computations visualizations and prediction algorithms of machine learning. • It can run on cloud.	• Useful for various responsibilities of data science. • Powerful tool for storytelling.
Matplotlib	Open source	• It is dedicated for plotting and visualization functions. • It provides matlab interfacing through pyplot module. • For data visualization with Python new learners can use this tool.	• Useful for various graphics models.

Tool Name	Type	Features	Applications
NLTK	Open source	• It is mainly used for text Analytics. • Useful for Natural Language Processing task. • It has a rich collection of machine learning algorithms.	• Useful for tokenization, steaming, tagging, passing and machine learning techniques • Useful for human language understanding
Scikit-learn	Open source	• This library is simple and easy to implement. • It supports number of machine learning techniques. • It is useful for the situation of rapid prototyping.	• Useful for data analysis in data science.
TensorFlow	Open source	• It can run on CPUs, GPUs and TPUs. • It has high processing ability. • It is easily trainable and has shared components. • It has high availability of statistical distributions and visualization.	• Useful for multidimensional data.
Weka	Open source	• It is written in Java so that it is platform independent. • It has rich collection of machine learning algorithms. • This tool is coding free.	• Useful for classification, clustering, regression, visualization and data preparation.