JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A SURVEY OF DATA SCIENCE TECHNIQUES, TOOLS, AND REAL-WORLD APPLICATIONS

Rakesh Kumar[1], Mukesh Kumar[2], Ruchi Patira[3]
[1-2]MTech Student, [3]Assistant Professor
Department of Computer Science & Engineering,
World College of Technology and Management Gurugram, India

Abstract: Data science is a rapidly growing field and has become essential in many industries. As data availability increases, the need for advanced data analysis techniques and tools becomes increasingly important. This paper provides an overview of data science techniques and tools that are widely used in the industry. It also explores real-world data science applications in various fields such as healthcare, finance, marketing, and social media. The paper begins with an introduction to data science and its importance in modern society. Then, the essential principles and techniques of data science, such as data cleaning, exploratory data analysis, statistical inference, machine learning, and deep learning. Overall, this review paper gives a complete overview of data science methodologies, tools, and applications, making it an invaluable resource for researchers and practitioners.
IndexTerms **– Data Science, Machine Learning, Deep Learning, Data Visualization.**

I. INTRODUCTION

Data science has emerged as a key discipline that spans multiple disciplines, including statistics, computer science, and domain-specific expertise. This involves extracting insights and knowledge from large and complex datasets using various techniques such as statistical modeling, machine learning, and data visualization.

In this section, we provide an overview of some of the fundamental techniques used in data science:

1. Data Cleaning: Data cleaning involves identifying and correcting errors and inconsistencies in the data. This step is essential before any analysis can be performed on the data.
2. Statistical Inference: Statistical inference involves using statistical models to make predictions or draw conclusions about the data. Hypothesis testing, confidence intervals, and regression analysis are all part of the process.
3. Machine Learning: Machine learning involves using algorithms to learn patterns and relationships in the data and make predictions or decisions based on that learning. Decision trees, random forests, support vector machines, and neural networks are some common machine-learning techniques [2].
4. Deep Learning: Deep learning is a subset of machine learning that involves training neural networks with multiple layers to learn complex patterns in the data. It has been utilized for picture and audio recognition, natural language processing, and a variety of other applications [6].

According to the type of data to be analyzed, the techniques are categorized into the following types:

1. Techniques based on Mathematics and Statistics
2. Techniques based on Artificial Intelligence and Machine Learning [12]
3. Techniques based on Visualization and Graphs

II. MATHEMATICS AND STATISTICS TECHNIQUES FOR DATA SCIENCE

Many data science models and algorithms are based on these techniques. Understanding the underlying mathematics and statistical concepts is essential for developing robust and accurate models that can make predictions and provide insights based on data. Here are some additional details on some of the techniques based on mathematics and statistics used in data science:

1. Linear Regression: Linear regression is a statistical approach for modeling the connection between one or more independent variables and a dependent variable. It assumes a linear relationship between the variables and uses least squares regression to estimate the coefficients.
2. Logistic Regression: Logistic regression is used to model the relationship between a binary dependent variable and one or more independent variables. It uses a logistic function to estimate the probability of the dependent variable given the independent variables.
3. Clustering Analysis: Clustering analysis is used to group data points into clusters based on similarity. It is used for segmentation, pattern recognition, and anomaly detection.

4. Support Vector Machines (SVMs): SVMs are used for classification and regression analysis. They determine the best hyperplane for classifying the data based on the largest margin.
5. Random Forests: Random forests are an ensemble learning technique used for classification and regression analysis. They use several decision trees to improve accuracy and avoid overfitting.

These techniques are just a few more examples of the many mathematical and statistical techniques used in data science. They are often combined with other techniques, such as data preprocessing and feature engineering, to provide valuable insights and predictions based on data.

III. Artificial Intelligence and Machine Learning Techniques for Data Science

Artificial Intelligence (AI) and Machine Learning (ML) techniques have become increasingly popular in data science for their ability to analyze large volumes of data and generate accurate predictions. Here are some AI and ML techniques commonly used in data science:

1. Supervised Learning: Supervised learning is a type of ML technique that involves training a model on labeled data to predict the outcome for new, unseen data. It includes techniques like linear regression, logistic regression, and decision trees.
2. Unsupervised Learning: Unsupervised learning is a type of ML technique that involves training a model on unlabeled data to find patterns and groupings in the data. Techniques such as clustering, anomaly detection, and principal component analysis (PCA) are included.
3. Reinforcement Learning: Reinforcement learning is a type of ML technique that involves training a model to make decisions based on feedback from the environment. It is used for applications like game playing, robotics, and recommendation systems.

These strategies demonstrate the wide range of AI and ML techniques accessible for data science, as well as the need of being up to date on the latest breakthroughs in order to remain at the forefront of the industry.

IV. Graphs and Visualization Techniques of Data Analysis for Data Science

Graphs and visualization techniques are important tools for data science, as they enable data scientists to better understand their data and communicate their findings to others. Here are three examples of graphs and visualization techniques used in data analysis:

1. Scatter Plots: Scatter plots are used to visualize the relationship between two variables. Each data point is plotted on the graph based on its values for the two variables, and the resulting pattern can reveal important insights about the data. For example, a scatter plot might be used to explore the relationship between a person's age and their income level [9].
2. Line Chart: A line chart is a basic graph that displays data points as a series of connected dots or lines, often used to show trends over time.
3. Bar Chart: A bar chart is a simple graph that uses bars to represent different categories or groups of data, often used to compare quantities.
4. Histogram: A histogram is a graph that displays the distribution of a continuous variable, typically shown as a series of vertical bars that represent the frequency of data within a certain range [12].
5. Heat Map: A heat map displays data as a grid of colored squares, with each square representing a data point and the color representing its value.
6. Pie Chart: A pie chart is a circular chart that displays data as a series of wedges, with each wedge representing a different category or variable.

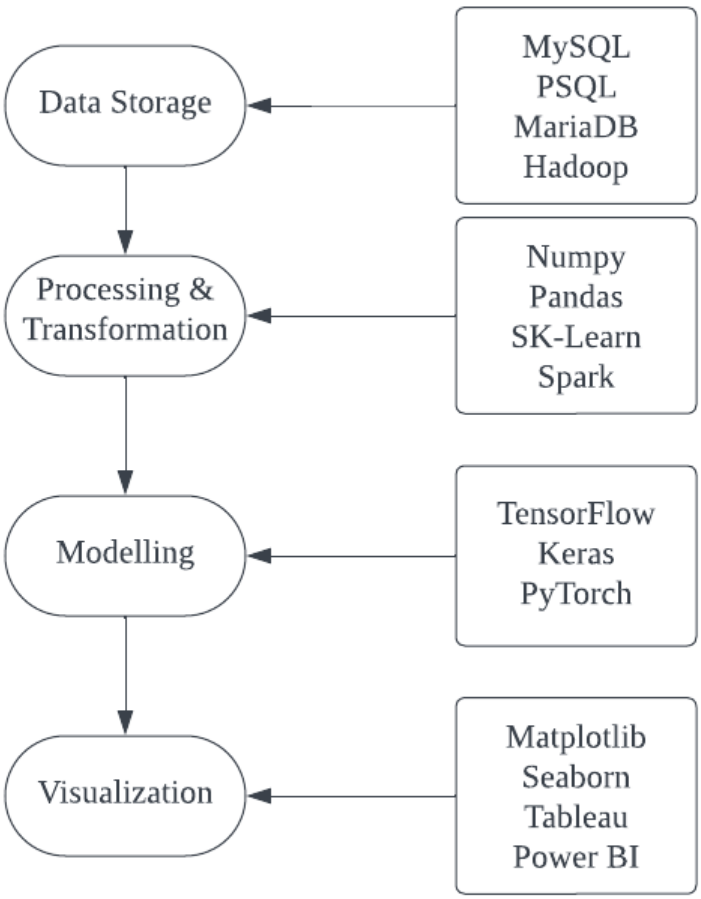

Figure 1: Data Science Tools and Framework

V. TOOLS FOR DATA SCIENCE

S. No	TYPE	Feature	Launch Year	Last update	Number of users
1	Programming Language	Data manipulation, Machine Learning, Deep Learning, Visualization [1]	1991	2022	Millions
3	Query Language	Data storage and retrieval	1974	Ongoing	Millions
4	Spreadsheet Software	Data manipulation, Visualization	1985	2021	Millions
5	Visualization Tool	Interactive dashboards, Data connection, Collaboration	2003	2021	Millions
7	Machine Learning Library	Deep Learning, Neural Networks, Natural Language Processing	2015	2022	Millions
8	Data Analysis Library	Data manipulation, Data cleaning, Data exploration [3]	2008	2021	Millions

9	Data Visualization Library	Line plots, Scatter plots, Bar plots, Histograms	2003	2021	Millions

Table 1: Feature and Launch Years of Data Science Tools

VI. Real-world application with details and amount of users

1. Netflix - Netflix is a popular online streaming platform that uses data science algorithms to recommend TV shows and movies to users based on their viewing history, ratings, and other data. Over 200 million people use the platform worldwide.
2. Amazon - Amazon uses data science techniques to personalize recommendations for users, optimize product pricing, and improve supply chain management. Globally, the platform has around 300 million active users [10].
3. Spotify - Spotify uses data science algorithms to personalize music recommendations for users based on their listening history, user-generated playlists, and other data. The platform has over 365 million active users worldwide.
4. Google - Google uses data science algorithms to provide relevant search results, personalize ads, and optimize user experiences across its suite of products, including Gmail, Google Maps, and YouTube. The platform has over 1 billion active users worldwide.
5. LinkedIn - Data science is used by LinkedIn to promote job listings, link job seekers with potential employers, and tailor information and adverts for users. The platform has over 740 million users worldwide.
6. YouTube (video-sharing platform) - YouTube is a video-sharing platform that allows users to upload, share, and view videos on a wide range of topics. By 2021, YouTube will have more than 2 billion monthly active users [4].

VII. Conclusion

In summary, data science is an interdisciplinary field that uses advanced computational and statistical techniques to analyze and extract insights from large and complex data sets. From mathematical and statistical models to machine learning algorithms and data visualization tools, there are a variety of techniques and tools available to data scientists.

E-commerce platforms and social media networks, as well as healthcare, banking, and transportation, are all examples of data science applications.. By harnessing the power of data science, organizations can make more informed decisions, improve operational efficiency, and create personalized experiences for their users.

As the volume and complexity of data continue to grow, the demand for skilled data scientists will continue to grow. By staying current with the latest data science techniques and tools, subject matter experts can drive innovation and positively impact a wide range of industries and applications.

Data visualization is an important part of data science because it aids in the communication of insights derived from data research. Python and R are two well-known programming languages for data science.

References

[1] R for Data Science - Hadley Wickham: https://r4ds.had.co.nz/
[2] Machine Learning Mastery - Jason Brownlee: https://machinelearningmastery.com/
[3] Big Data Analytics - IBM: https://www.ibm.com/analytics/hadoop/big-data-analytics
[4] Real-World Machine Learning - Henrik Brink, Joseph Richards, and Mark Fetherolf: https://www.oreilly.com/library/view/real-world-machine-learning/9781491964548/
[5] Jessica Davis.2016.10 Programming Languages And Tools Data Scientists Used. Retrieved from http://www.informationweek.com/devops/programming- languages/10-programming-languages-and-tool s-data-scientists-use-now/d/d-id/1326034.
[6] Witten, I., H., Frank, E., Hall, M., A. (2011). Data mining: practical machine learning tools and techniques.
[7] Judge, Peter (2012-10-22). "Doug Cutting: Big Data Is No Bubble". silicon.co.uk. Retrieved 2018-03-11.
[8] Hemsoth, Nicole (2014-10-15). "Cray Launches Hadoop into HPC Airspace". hpcwire.com. Retrieved 2018-03-11.
[9] [Online]. Available: https://www.tibco.com/solutions/business-activity-monitoring [Accessed 5-July-2019].
[10] Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745-766.
[11] DataRobot. 2016. DataRobot. Retrieved from https://www.datarobot.com/.
[12] https://data-flair.training/blogs/data-science-tools/ Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.
[13] Longbing Cao.2017. Data science: A comprehensive overview. ACM Comput. Surv. 50, 3, Article 43 (June 2017), 42 pages. DOI: http://dx.doi.org/10.1145/3076253.