DATA SCIENCE FOR DECISION MAKING IN 21^{ST} CENTURY

Dr. Sampada Gulavani, Associate Professor,
Bharati Vidyapeeth (Deemed to be University), Institute of Management, Kolhapur.
sampada.gulavani@bharatividyapeeth.edu

Dr. Rajesh Kanthe, Director,
Bharati Vidyapeeth (Deemed to be University), Institute of Management, Kolhapur,
rajesh.kanthe@bharatividyapeeth.edu

Dr. Mukund Kulkarni, Asst. Professor,
Bharati Vidyapeeth (Deemed to be University), Institute of Management, Kolhapur,
mukund.kulkarani@bharatividyapeeth.edu

ABSTRACT

In the 21st century, emerging fields in computer science are Data Science & Machine Learning. Data Science analyse the given data using statistical analysis and identifies hidden patterns among the data. The purpose of data science is to extract meaningful and logical patterns from the given data. Nowadays there is much more demand for data scientists whose job is to obtain results using statistical analysis and communicate the findings to the end user to assist them to make better business decisions. Data science includes different steps like capturing of data from different sources, maintenance of data, processing on data, and analysis of data and communicating the data to the end user in the form of chart, graph or reports. Data science uses different technologies like decision trees, classification, and clustering and dimensionality reduction. Data science has a lot of applications in the field like healthcare, gaming, image recognition, recommendation systems, fraud detection etc.

Keywords: Data science, Machine learning, regression, correlation, dimensionality reduction.

Introduction

Due to tremendous use of social media, it is observed that more data is being created in today's world. The traditional business intelligence tools cannot be applicable to process massive amounts of unstructured data. Around 2008, the data science field emerged due to the need of organizing, analysing and predicting large volumes of data collected from different organizations. The aim of data science is to discover hidden patterns from the raw data with the help of different algorithms and machine learning techniques. It uses different techniques of machine learning, classification, data mining etc. Data science deals with different types of data like noisy, structured or unstructured and uses different algorithms to extract knowledge from the stored data. The aim of data science is to reveal the hidden structure of data from different points of view. Prerequisites for data science include the knowledge of statistics, machine learning, programming and how to manage the database and extract data from them. Once the data is collected, there is a need to clean the data, prepare it and align the data as per the need. By considering this, the paper consists of important steps in data science, technologies used in data science and applications of data science in various fields.

Literature Review

Shrestha (2019) in his study found that Data science consists of quantitative and qualitative aptitudes and nonmathematical abilities. Data science transmits the data for inspection purposes. The aim of Data science is to investigate the collected data and extract the data. To control autonomous gadget machines are being used by different methods of technology and the Internet of Things.

Lakhani (2022) in his study found that data Science depends upon the decision which is gathered from data analysis. It totally depends upon the experience and intuition of the decision maker. Data science and Artificial Intelligence caused the tremendous change in decision making of medium and large scale businesses. For this there is always a need for a large amount of data to finalize the best policies with different types of patterns.

Crocetta (2021) in his study found that the use of the web has increased which has caused tremendous effect on data production and consumption on a daily basis. With the increasing use of IoT and social networks, a large amount of data is being generated. Once the data is consolidated within the different disciplines, there becomes a problem of data revolution. It arises from a need for data science which deals with this type of data. It assists to make better policy maker decisions for business organizations.

Varshney(2017) found that Data science deals with large amount of data to extract proper results or patterns. It uses different activities like data mining, classification and regression techniques. It considers techniques from different fields like mathematical science, advanced statistics and computer science. From computer science it uses different techniques like machine learning, classification & cluster analysis for prediction purposes.

Nandhini (2018), found the working nature of Big data with traditional technologies by considering three different ways like the volume of data , the rate of data generation and transmission and variety of data. In the last two years, due to the use of social media, tremendous data is being generated. To deal with such a large amount of data, there is a need for new technology where the data science field has emerged.

Krishna (2018) in his study found that the aim of data science is to predict accurate and trustworthy solutions. The purpose is to process data and prepare predictive modelling. There is a need for Sought Skill-sets where analysts must focus and work in it. In today's different applications, there is more scope for artificial intelligence. The aim is to make human life simple and without any complication. For this artificial intelligence algorithms are preferred for prediction purposes. Use of the recommender system helps e-commerce sites to recommend the products which are based on their recent search.

Dumontier (2017) In his study on Data Science Methods, infrastructure, and applications, found that Data Science is a new paradigm which aims to study problems and questions in the existing disciplines. It tries to explore different possibilities for analysis of data. Once the data is analysed and prediction is done, different problems may arise like proper treatment for sensitive data, transparency of scientific data and data gathering methods. The aim is that data should be properly accessible, reliable, and reusable.

Wang (2019) in his study found that Data scientists are in search to find multiple causal relationships among data. Here they will not overcome strong assumptions of univariate causal inference. They always consider different factors that are interrelated and interact with each other which affects the data. For this a method of feature construction is preferred which uses relational databases along with deep learning.

Wing (2018) in his study found that the focus of data science is on different methods and data being collected. Here a specific domain is considered for the inclusion of the data science field. In the context of other domains, different disciplines used are computer science, mathematics, and statistics.

Kanter & Veeramachaneni (2015) in his study found that relational database and deep feature synthesis is being used by data science. For this role automated feature construction is considered along with machine learning. In this paper, different machine learning methods are considered to achieve expert level performance.

Objectives of the Study

1. To study different steps included in Data science
2. To study the different areas where Data science is applicable.

Steps in Data Science:

i. Capture: Structured and unstructured raw data is collected from different sources by using methods like manual entry or web scraping. Data is captured from different sources where data is in structured or unstructured form.
ii. Prepare and Maintain: In this step first put the raw data in consistent format which is further required for analytics purpose. It includes different steps like cleansing, duplicating and reformatting the data. Once data is clean and transform, then it is combined into a data warehouse. The aim is proper analysis of the collected data.
iii. Process: In this step data is checked for the suitability purpose which is determined by examining different patterns and distribution of values within the data. For processing on the data different algorithms are used like deep learning or machine learning.
iv. Analyse: In this step, data scientists perform different type of analysis like statistical analysis, predictive analytics, regression, machine learning and deep learning algorithms. The aim is to extract insights from the prepared data.
v. Communicate: In this step final results are communicated to the end users in the form of reports, charts, and other data visualizations techniques. It helps decision makers to make proper decisions. It uses different programming languages like R or Python. These languages include components for generating visualizations. Common tools used for modelling SQL Analysis services, SAS/Access and R

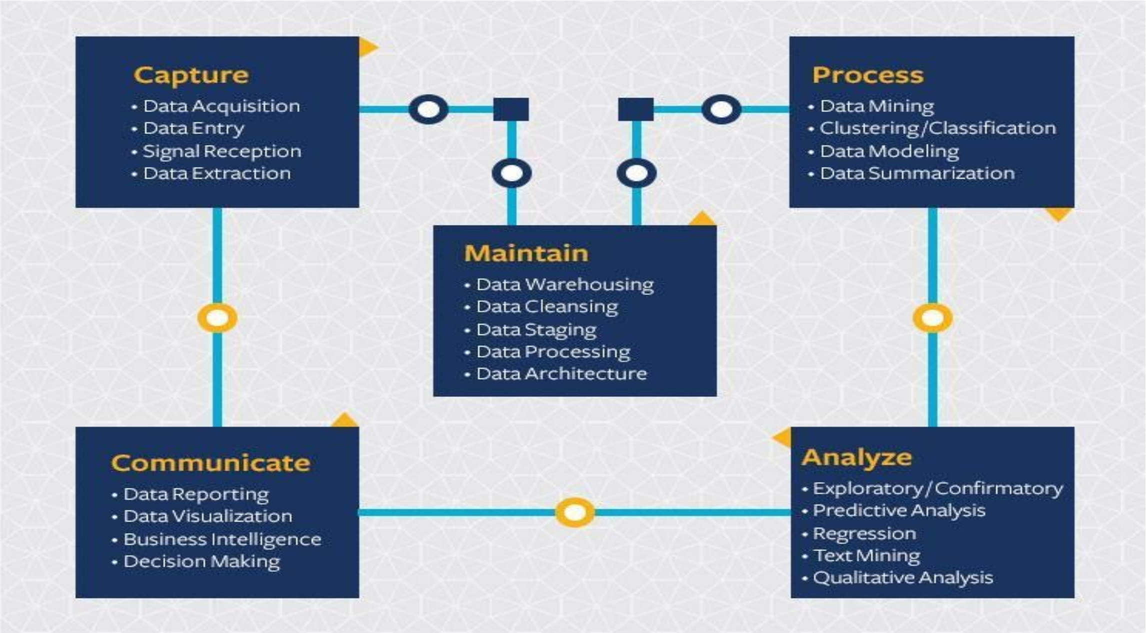

Figure 1: Steps in Data Science (Source:www.slideshare.com)

Technologies used in Data Science:
Commonly used technologies in data science are :

i) Linear Regression**:** In this method, a value of dependent variable (y) is predicted with given independent variable (x). It provides a linear relationship between x and y. This method is based on supervised learning.
ii) Logistic Regression: Logistic regression considers the relation between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It uses independent variables to predict the categorical dependent variable. The output value lies between 0 and 1. Generally it is considered in classification problems.
iii) Decision Trees: These are used for classification of data and the aim is for prediction of the future. It is used to generate rules to predict target variables with the help of observed variables. In decision tree variables are used in hierarchical order and provide answers in step by step manner.
iv) Naive Bayes Classifiers : It uses bayes theorem to classify the data. Generally, it is preferred when the dataset is large in volume and provides accurate results.
v) Clustering: In clustering, the different clusters are prepared. This type of learning is called unsupervised learning. Based on the similarity measure criterion, cluster analysis has various models like hierarchical clustering, nearest cluster centre and density models.
vi) Dimensionality Reduction: In dimensionality reduction, complexity of data is reduced while computation.

Data Science Programming Models:

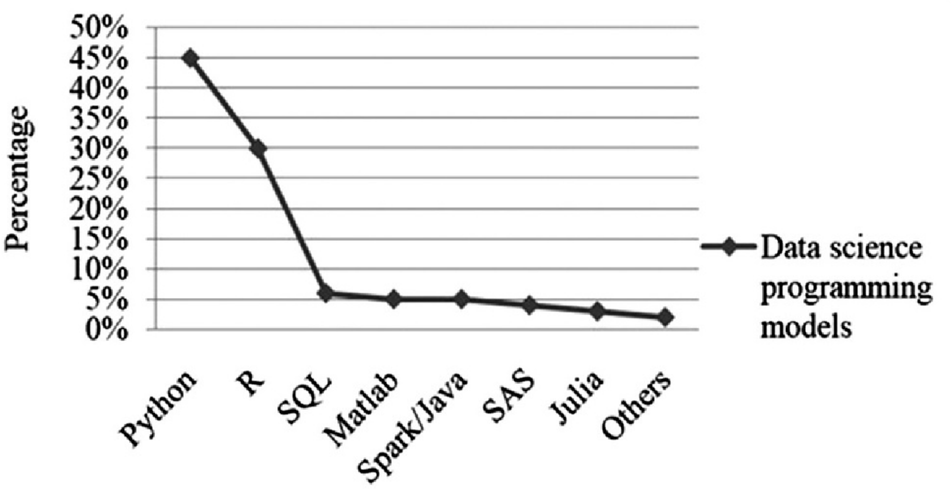

Figure 2: Data Science Programming Models (www.slideshare.com)

i) Python: Python is an easy and open-source language. It has a collection of libraries for data manipulation and data analysis. Its agility and productivity make python the commonly used programming language for data science. The future of python depends on the number of service providers that allow for SDKs in python. It is also necessary to consider the expansion of python apps.
ii) R Language: R is commonly used for statistical purposes. It is compatible with Windows, Macintosh, UNIX, and Linux platforms. It offers extensible, source language and computing environment along with effective data handling and storage facility. It has a large collection of intermediate tools for data analysis. It also provides graphical facilities for data analysis purposes.
iii) Hadoop: It is used for big data and is an open source software framework. It is used for distributed storage of very large datasets on different clusters. It assists in the massive amount of storage of the data. It provides enormous processing power to handle concurrent tasks.
iv) Visualization Tools: It considers the creation of visual representation of data. It uses Tableau with the aim to generate attractive plots and charts with animation. It uses Visualization tools like D3 to build the visualization frame-work. Here we can use a tool Data wrapper to create different maps.
v) Tensor Flow: It is generally used for computation purposes and preferred when computations need to be visualized in a data flow graph. It has interfaces with C++ and CUDA support. It is available on embedded platforms.

Applications of Data Science:
Data science has found its applications in different industry

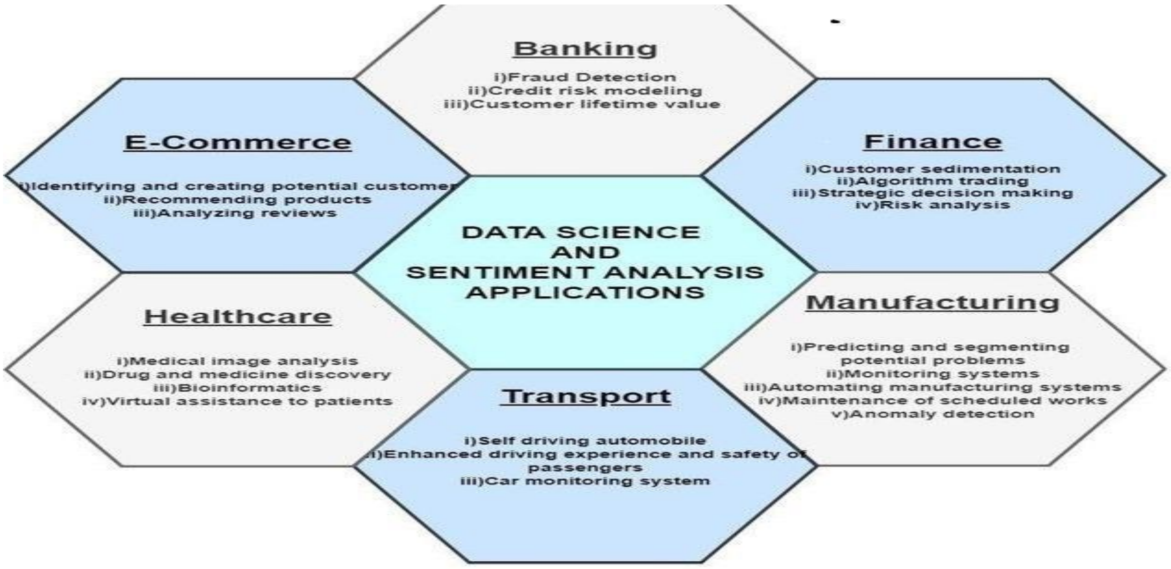
Banking
i)Fraud Detection
ii)Credit risk modeling
iii)Customer lifetime value
E-Commerce
Identifying and creating potential customer
ii)Recommending products
iii)Analyzing reviews
Finance
i)Customer sedimentation
ii)Algorithm trading
iii)Strategic decision making
iv)Risk analysis
DATA SCIENCE AND SENTIMENT ANALYSIS APPLICATIONS
Healthcare
i)Medical image analysis
ii)Drug and medicine discovery
iii)Bioinformatics
iv)Virtual assistance to patients
Manufacturing
i)Predicting and segmenting potential problems
ii)Monitoring systems
iii)Automating manufacturing systems
iv)Maintenance of scheduled works
v)Anomaly detection
Transport
i)Self driving automobile
)Enhanced driving experience and safety o passengers
iii)Car monitoring system

Figure 3: Applications of Data Science (Source: www.slideshare.com)
i) Healthcare: Data Science has enabled practitioners to analyse the data, make correlations between the variables of the data. Data Science also enables hospital managers to reduce waiting time and give more attention for care of patients.
ii) Self-Driving Cars: Now-a-days self-driving cars can adjust speed limits and avoid different routes. It uses machine learning and predictive analysis. It uses tiny cameras and sensors to relay information in real time. Commonly used self-driving cars are Tesla, Ford and Volkswagen.
iii) Cyber security: Kaspersky, one of the international firms, uses data science and machine learning to detect over 360,000 new samples of malware on a daily basis. Here Data science assists to detect and learn new methods of cyber -crime.
iv) Image Recognition: Data science is used to identify patterns in images and detecting objects in an image.
vi) Recommendation Systems: Data science is commonly used in recommender systems like Netflix and Amazon. It gives movie and product recommendations based on which platform you are browsing.
vii) Fraud Detection: To detect fraud, data science is used in banking and financial institutions. Facts collected from analyst logs are utilized to recognize fraud utilizing information science. It uses information mining and machine learning to counteract instances of fraud.

Conclusion
Data science reveals trends and can be used to make better decisions in business organizations. It uses different tools, technologies and resources. With the help of this, it creates innovative products and services to the

business sector. It is predicted that by 2025, global data will increase to 175 zettabytes. Here data science derives valuable conclusions and predictions. Machine learning models learn from vast amounts of data fed to them with training and tasting data sets. Different application areas where data science is widely used are healthcare, finance, banking, image recognition, recommendation system and more. Today companies are using different techniques of data science to expand their business and to provide proper service to their customers. Data scientists should possess skills like mathematics, advanced statistics and computer science. In the future, there is growing demand for data scientists in different industries. But further it may lead to increased complexity and challenges in the field of data science.

References

Crocetta. C, Carpita M., & Perchinunno P (2021). “Data Science and Its Applications to Social Research”, Springer .

Dumontier M. & Kuhn T. (2017). “Data Science – Methods, infrastructure, and applications”, DOI 10.3233/DS-170013 IOS Press.

Kanter J., & Veeramachaneni K. (2015). “Deep Feature Synthesis: Towards Automating Data Science Endeavors". Proceedings of the International Conference on Data Science and Advanced Analytics. IEEE.

Krishna C., & Mohana H (2018). “A Review of Artificial Intelligence Methods for Data Science and Data Analytics: Applications and Research Challenges”, Proceedings of the Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018) IEEE Xplore Part Number:CFP18OZV-ART; ISBN:978-1-5386-1442-6

Lakhani N. (2022). Applications of Data Science and AI in Business, Applications of Data Science and AI in Business, Vol.10, Issue 5, SSN: 2321-9653.

Nandhini P., .Kalpana V., & Sikkandhar J. (2018). “Data analytics application used in the field of big data for security intelligence”, International Journal of Creative Research Thoughts, Vol.6, Issue 2.

Shreshtha S., Singh A., Sahdev S., Singha M. & Rajput S. (2019) IEEE Conference, 978-1-5386-9346-9.

Varshney M., Garg S., Jyotsna A., & Kiran R. (2017). “A Study on Issues, Challenges and Application in Data Science”, International Journal of Trend in Scientific Research and Development, Vol.1, Issue 5, ISSN: 2456 – 6470.

Wang Y., & Blei D. (2019). “The blessings of multiple causes”. Journal of the American Statistical Association, 114(528), 1574-1596.

Wing J., (2019). “The data life cycle. Harvard Data Science Review*”*, *1*(1).