INTRODUCTION TO DATA SCIENCE (IS1102-1)

Course Learning Objectives

1. Explain the concepts of data mining and types of Analytics
2. Illustrate the use of different data mining algorithm
3. Describe the basic concepts of R programming
4. Apply the Data visualization concepts using R programs
5. Get the idea of lookup functions and Pivot Tables and Illustrate the use of Data validation and Data Visualization

Content

UNIT -I

1. Introduction to Data Science
2. Introduction to Data Mining:
3. Preprocessing
4. Data Warehouse and On-line Analytical Processing
5. Classification:

UNIT – II

1. R Programming Basics:
2. Data Visualization using R
3. Working with R Charts and Graphs

UNIT – III

1. Introduction to Data Analysis using Excel
2. Data Analysis Process
3. Data Quick Analysis:

Course Outcomes

At the end of the course student will be able to

1. Apply the Concepts of data science in various fields
2. Study different data mining algorithm
3. Describe R basics, Variables and Data Types, Control Structures, Array, Matrix, Vectors, Factors,
4. Analysis the data using different R graphs and Charts.
5. Acquire the knowledge of data analysis and carry out the data analysis process.

References

Text Books

1. Jiawei Han, Micheline Kamber, Jian Pei (2012), Data Mining: Concepts and Techniques
2. Microsoft Excel 2019 Data Analysis and Business Modelling (Business Skills), 6th Edition, Wayne L Winston, ISBN-13: 978-1509305889, ISBN-10: 1509305882.
3. Tilman M. Davies, "The Book of R: A First Course in Programming and Statistics", No Starch Press; 1st Edition,2016.

UNIT -1

1.1 Introduction to Data Science

- Evolution of Data Science
- Data Science Roles
- Stages in a Data Science Project
- Applications of Data Science in various fields.

What is Data Science?

- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

- Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results

Evolution of Data Science

The key milestones in the evolution of data science:

1962: American mathematician John W. Tukey was a visionary who foresaw the emergence of data science long before personal computers, as detailed in his influential article "The Future of Data Analysis.“

Similarly, Peter Naur, a Danish computer engineer, recognized the significance of data science and offered an early definition in his book, describing it as the science of managing established data while leaving the interpretation of data representations to other fields and sciences.

Their pioneering insights laid the groundwork for the development of data science as a distinct and transformative field.

1977: The theories and predictions of "pre" data scientists like Tukey and Naur became more concrete with the establishment of The International Association for Statistical Computing (IASC), whose mission was "to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

Evolution of Data Science

1980s and 1990s: Data science began taking more significant strides with the emergence of the first Knowledge Discovery in Databases (KDD) workshop and the founding of the International Federation of Classification Societies (IFCS). These two societies were among the first to focus on educating and training professionals in the theory and methodology of data science

It was at this point that data science started to garner more attention from leading professionals hoping to monetize big data and applied statistics

1990s and early 2000s: Data science has emerged as a recognized and specialized field. Several data science academic journals began to circulate, and data science proponents like Jeff Wu and William S. Cleveland continued to help develop and expound upon the necessity and potential of data science.

2005: Big data enters the scene. With tech giants such as Google and Facebook uncovering large amounts of data, new technologies capable of processing them became necessary. Hadoop rose to the challenge, and later on Spark and Cassandra made their debuts.

Evolution of Data Science

The key milestones in the evolution of data science:

2010 : Data Science as a Distinct Field

- In the early 2010s, data science started gaining recognition as a distinct field, fueled by the need to make sense of vast amounts of data.
- The term "data science" became more widely used as organizations sought professionals with expertise in data analysis, statistics, and programming.
- Data science emerged as an interdisciplinary field, incorporating elements from statistics, computer science, machine learning, and domain knowledge.

2014: Due to the increasing importance of data, and organizations' interest in finding patterns and making better business decisions, demand for data scientists began to see dramatic growth in different parts of the world

Evolution of Data Science

Data Science (Present):

- Today, data science plays a crucial role in decision-making across various sectors, including finance, healthcare, marketing, and more.
- Companies use data science to optimize operations, improve customer experiences, and gain a competitive edge.
- Data scientists started leveraging AI and deep learning to tackle more challenging problems and enhance data analysis.
- Data science has become an integral part of the digital transformation journey for organizations worldwide.

Stages of Data Science life cycle

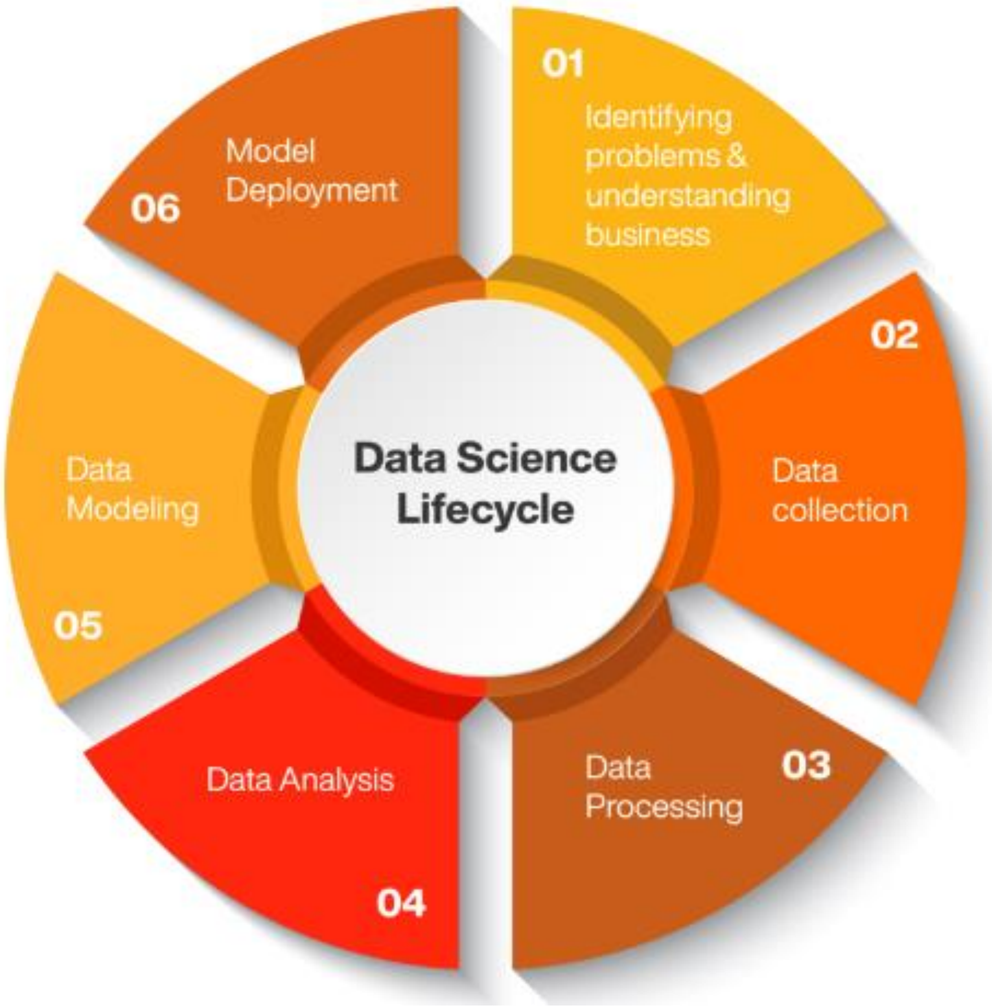
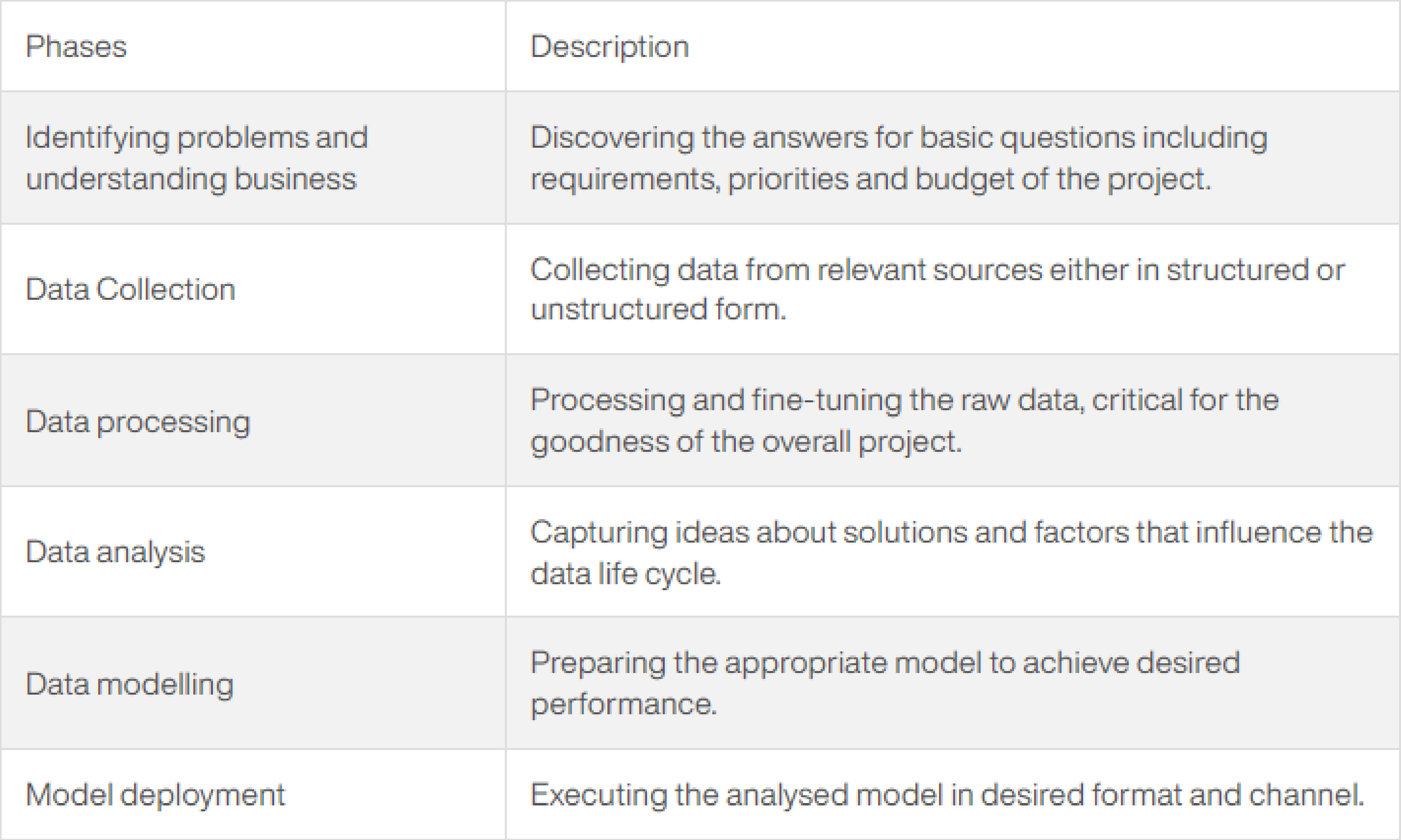


Figure: The different phases of data science life cycle

Stages of Data Science life cycle

The different phases of data science life cycle



Phases	Description
Identifying problems and understanding business	Discovering the answers for basic questions including requirements, priorities and budget of the project.
Data Collection	Collecting data from relevant sources either in structured or unstructured form.
Data processing	Processing and fine-tuning the raw data, critical for the goodness of the overall project.
Data analysis	Capturing ideas about solutions and factors that influence the data life cycle.
Data modelling	Preparing the appropriate model to achieve desired performance.
Model deployment	Executing the analysed model in desired format and channel.

Stages of Data Science life cycle

1. Identifying problems and understanding business

It is important to understand the business objective early since it will decide the final goal of your analysis.

The phase should –

- Clearly state the problem that requires solutions and why it should be resolved at once
- Define the potential value of the business project
- Find risks, including ethical aspects involved in the project
- Build and communicate a highly integrated, flexible project plan

2. Data collection

- In the data collection stage raw data is gathered from relevant sources.
- The data captured can be either in structured or unstructured form.
- The methods of collecting the data might come from – logs from websites, social media data, data from online repositories, and even data streamed from online sources via APIs, web scraping or data that could be present in excel or any other source.

Stages of Data Science life cycle

3. Data processing

- In this phase, data scientists analyze the data collected for biases, patterns, ranges, and distribution of values.
- It also involves the introspection of different types of data, including nominal, numerical, and categorical data.
- Data visualization is also done to highlight the critical trends and patterns of data, comprehended by simple bars and line charts. Data preparation is the most time-consuming but the most critical phase in the entire life cycle of data analytics
- The goodness of the model depends on this data processing stage.

4. Data Analysis or Exploratory Data Analysis

- Once the data is available and ready in the required format, understand and analysis the data by utilizing various statistical tools
- The data is investigated by creating various statistical functions and identifying dependent and independent variables or features.
- Data analysis reveals which data or features are essential, as well as the distribution of data. To aid comprehension, various plots are used to show the data.

Stages of Data Science life cycle

5. Data Modelling

- Modelling Data is one of the major phases of data processes and is often mentioned as the heart of data analysis. A model takes the organized data as input and gives the preferred output.
- This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem.
- After deciding on the model family, cautiously pick the algorithms to put into effect and enforce them.
- Tune the hyperparameters of every model to obtain the preferred performance.

6. Model Evaluation & Model Deployment

- The model is now being tested with real-world data and evaluated on a cautiously thought out set of assessment metrics. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved
- The model is now deployed to real-time data entering the system, and output is generated. The model can be deployed as a web service, an embedded application in an edge application, or a mobile application. This is a critical phase since the model is now exposed to the real world.

Data Scientist Roles and Responsibilities

Data Scientist Roles and Responsibilities

1. Collect data and identify data sources
2. Analyze huge amounts of data, both structured and unstructured
3. Create solutions and strategies to business problems
4. Work with team members and leaders to develop data strategy
5. To discover trends and patterns, combine various algorithms and modules
6. Present data using various data visualization techniques and tools
7. Investigate additional technologies and tools for developing innovative data strategies
8. Create comprehensive analytical solutions, from data gathering to display; assist in the construction of data engineering pipelines

Data Scientist Roles and Responsibilities

9. Supporting the data scientists, BI developers, and analysts team as needed for their projects. Working with the sales and pre-sales team on cost reduction, effort estimation, and cost optimization
10. To boost general effectiveness and performance, stay current with the newest tools, trends, and technologies
11. collaborating together with the product team and partners to provide data-driven solutions created with original concepts
12. Create analytics solutions for businesses by combining various tools, applied statistics, and machine learning
13. Lead discussions and assess the feasibility of AI/ML solutions for business processes and outcomes
14. Architect, implement, and monitor data pipelines, as well as conduct knowledge sharing sessions with peers to ensure effective data use

Data Science Roles

Data science is a multidisciplinary field that requires a diverse set of skills and expertise. As a result, various roles have emerged to cover different aspects of the data science process. Some of the key data science roles include:

1. **Data Scientist**:
    1. Data scientists are responsible for designing and implementing algorithms and models to extract insights and knowledge from data.
    2. They analyze complex datasets, perform statistical analysis, and use machine learning techniques to solve business problems and make data-driven decisions.
    3. Data scientists should have strong programming skills, expertise in data manipulation, and a deep understanding of machine learning algorithms.
2. **Data Analyst**:
    1. Data analysts focus on interpreting data, preparing reports, and generating visualizations to communicate insights to stakeholders.
    2. They work with structured and unstructured data to identify trends, patterns, and correlations.
    3. Data analysts often use tools like SQL, Excel, and data visualization platforms to analyze and present data.

Data Science Roles

3. Machine Learning Engineer:

1. Machine learning engineers are responsible for developing and deploying machine learning models into production systems.
2. They work closely with data scientists to implement and optimize algorithms and ensure the models are scalable and efficient.
3. Machine learning engineers also handle model deployment, monitoring, and maintenance.

4. Data Engineer:

1. Data engineers are involved in building and maintaining the infrastructure required for data storage, processing, and retrieval.
2. They design and develop data pipelines to collect, clean, and transform data, ensuring that it is available and accessible for analysis.
3. Data engineers use technologies like Hadoop, Spark, and databases to handle large-scale data processing.

Data Science Roles

5. Business Intelligence (BI) Analyst:

1. BI analysts focus on analyzing historical data to support business decision-making.
2. They create dashboards, reports, and data visualizations to convey insights to business stakeholders.
3. BI analysts often work with tools like Tableau, Power BI, or other business intelligence platforms.

6. Data Architect:

1. Data architects are responsible for designing and managing the overall data infrastructure and architecture of an organization.
2. They ensure that data is organized, integrated, and available for analysis and decision-making.
3. Data architects also define data governance policies and best practices.

7. Big Data Engineer:

1. Big data engineers specialize in managing and processing large volumes of data, often in distributed computing environments.
2. They work with technologies like Hadoop, Spark, and NoSQL databases to handle big data challenges efficiently.

Applications of Data Science in various fields.

Data science has a wide range of applications across various fields and industries. Some of the prominent applications include:

1. **Healthcare**:
    1. Data science is used to analyze electronic health records, medical imaging, and patient data to improve diagnostics, personalized medicine, and treatment outcomes.
    2. Predictive modeling is applied to forecast disease outbreaks and identify high-risk patients for preventive care.
2. **Finance**:
    1. In finance, data science is used for fraud detection, credit risk assessment, and algorithmic trading.
    2. Sentiment analysis is used to gauge market sentiment and make investment decisions.
3. **Marketing and Advertising**:
    1. Data science is employed to analyze customer behavior, preferences, and buying patterns to create targeted marketing campaigns.
    2. Customer segmentation and recommendation systems are used to enhance customer experience and increase sales.

Applications of Data Science in various fields.

4. E-commerce:

1. Data science drives product recommendations, supply chain optimization, and inventory management in e-commerce platforms.
2. It is also used to analyze customer reviews and feedback to improve product offerings.

5. Manufacturing and Industry:

1. Data science is used in predictive maintenance to identify potential equipment failures and optimize production processes.
2. It helps in quality control and optimization of manufacturing processes.

6. Transportation and Logistics:

1. Data science is used to optimize routes, improve logistics and supply chain management, and predict demand in transportation and delivery services.

7. Energy and Utilities:

1. Data science is applied in predicting energy consumption, optimizing energy distribution, and identifying patterns to reduce waste and increase efficiency.

Applications of Data Science in various fields.

8. Education:

- Data science helps in personalized learning by analyzing student data to tailor educational content and identify at-risk students who need additional support.

9. Social Media and Sentiment Analysis:

- Data science is used to analyze social media data to understand trends, customer sentiment, and public opinions.

10. Climate and Environmental Sciences:

- Data science plays a vital role in analyzing climate data, weather patterns, and environmental factors to understand and predict changes in the environment.

11. Government and Public Policy:

- Data science is used in various government sectors to analyze public data, improve policy-making, and optimize resource allocation.

12. Sports Analytics:

- Data science is employed in sports to analyze player performance, optimize team strategies, and enhance fan engagement.