# Principles of Data Science

Vasudevan T V

# Course Contents

- Module 1 - Introduction to Data Science
- Module 2 - Data Mining and Preprocessing
- Module 3 - Classification models
- Module 4 - Introduction to Association Mining , Clustering and Evaluation metrics

# Module 1
# Introduction to Data Science

- Data science is a collection of techniques used to extract value from data
- It has become an essential tool for any organisation that collects, stores, and processes data as part of its operations
- Data science techniques find useful patterns, connections, and relationships within data
- It can be used for decision making
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining
- However, each of these terms have a different meaning depending on the context

# Importance of Data Science

- Example - Retail Clothing Shop
- Consider a retail clothing shop that wants to increase its sales.
- It can use the following types of analysis using data science techniques
- sales analysis
- By analysing past sales data, the store can identify which items were bestsellers during certain seasons
- For instance, if winter jackets sold really well last year, the store can stock more of these for the upcoming winter
- Customer Behaviour Analysis
- Data science can classify customers based on their purchasing behaviour
- For example, if data shows that younger people tend to buy casual wear more than formal wear, then the shop can start marketing campaigns suitable for the corresponding segment

# Importance of Data Science

- Inventory Management
- Using predictive analytics, the store can forecast demand for different products
- If data indicates that sales of workout gear will increase in the spring, the store can ensure it has enough stock to meet that demand
- Targeted Marketing
- Data science enables personalized marketing strategies
- If a customer frequently buys shoes, the store can send them advertisements for new footwear, which is likely to increase the chances of a sale

# History of Data Science

| Decade | Key Development | Impact on Data Science |
|---|---|---|
| 1950s-1960s | Emergence of Computers | Automated data processing |
| 1970s | Development of statistical software | Broader access to data analysis tools |
| 1980s | Introduction of Personal Computers | Spread of data skills among the populace |
| 1990s | Introduction of World Wide Web | Massive increase in data availability |
| 2000s | Machine Learning and Predictive Analytics | Enhanced decision-making capabilities |
| 2010s-2020s | AI integration | Advanced predictive models |

# Types of Data

1. Structured data
- It is a kind of data that has a well defined structure
- It is stored in tabular form, organised into rows and columns
- Examples - Relational Databases, Spreadsheets, CSV files
2. Semi-structured data
- It is a kind of data that is less structured
- It uses tags, key-value pairs, headers etc. to separate data
- Examples - HTML, XML, JSON, Emails
3. Unstructured data
- It is a kind of data that has no fixed format
- Examples - Text documents, image files, audio files, video files

# Types of Data

1. Categorical data

- It is a kind of data that can be divided into different categories or groups
- It is of two types - ordinal and nominal

(a) Ordinal data

- It is a kind of data that has a meaningful order
- Examples
- Feedback Ratings - {Poor, Satisfactory, Fair, Good, Very Good, Excellent, Outstanding}
- Education Levels - {Higher Secondary School, Bachelor's Degree, Master's Degree}
- Class Grades - {S, A+, A, B+, B, C+, C, D, P, F}

# Types of Data

1. Categorical data
   - It is a kind of data that can be divided into different categories or groups
   - It is of two types - ordinal and nominal

(b) Nominal data
   - It is a kind of data that does not have any inherent order
   - Examples
   - Gender - {Male, Female, Transgender}
   - Fruits - {Mango, Apple, Banana}
   - Eye Colour - {Blue, Black, Brown}

# Types of Data

2. Non-Categorical data ( Numeric Data )
   - It contains numerical values that have a meaningful order
   - It is of two types - continuous and discrete

(a) Continuous data
   - It can take any value within a given range
   - Examples
   - Temperature
   - Height
   - Weight

# Types of Data

2. Non-Categorical data ( Numeric Data )

- It contains numerical values that have a meaningful order
- It is of two types - continuous and discrete

(a) Discrete data

- It consists of distinct values, which are often counted, not measured
- Examples
- Number of students in a class
- Number of cars in a parking lot
- Number of players in a team

# Real World Applications of Data Science

1. Healthcare
   - To forecast patient admissions and optimise resource allocation
2. Finance
   - Analyse transaction patterns to identify fraudulent activities
3. Education
   - To predict the performance of a student
4. Retail Shops
   - Classify customers based on their purchasing behaviour
5. Transportation
   - To predict failures of vehicles before they occur
6. Sports
   - Analyse player statistics to optimise performance

# References

1. Fundamentals of data science, Wagh, S. J., Bhende, M. S., Thakare, A.D., Chapman and Hall/CRC, Ist Edition 2021 ( Module 1 )