NARSIMHA REDDY ENGINEERING COLLEGE
UGC-AUTONOMOUS INSTITUTION
Your roots to success...
An Autonomous Institute
NAAC Accreditation 'A' Grade
Accredited by NBA
Approved by AICTE, Affiliated to JNTUH

UNIT-I

Data Management: Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/Signals/GPS etc. Data Management, Data Quality(noise, outliers, missing values, duplicate data) and Data Processing & Processing

Data Management

Data analytics is the process of analyzing raw data to find trends and answer questions. It has a broad scope across the field. This process includes many different techniques and goals that can shift from industry to industry.

The data analytics process has components that can help a variety of initiatives. By combining these components, a successful data analytics initiative can help answer business questions related to historical trends, future predictions and decision making.

The data used was not as much of as it is today, the data then could be so easily stored and managed by all the users and business enterprises on a single computer, because the data never exceeded to the extent of 19 exabytes but now in this era, the data has increased about 2.5 quintillion per day. Most of the data is generated from social media sites like Facebook, Instagram, Twitter, etc, and the other sources can be e-business, e-commerce transactions, hospital, school, bank data, etc. This data is impossible to manage by traditional data storing techniques. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analysed to benefit yourself from it. But how do we do it? Well, that's where the term 'Data Analytics' comes in.

Data Analytics important :

Data Analytics has a key role in improving your business as it is used to gather hidden

insights, Interesting Patterns in Data, generate reports, perform market analysis, and improve business requirements

Role of Data Analytics

Gather Hidden Insights – Hidden insights from data are gathered and then analyzed with respect to business requirements.

Generate Reports – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.

Perform Market Analysis – Market Analysis can be performed to understand the strengths and weaknesses of competitors.

Improve Business Requirement – Analysis of Data allows improving Business to customer requirements and experience.

Data and architecture design:

Data architecture in Information Technology is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations.

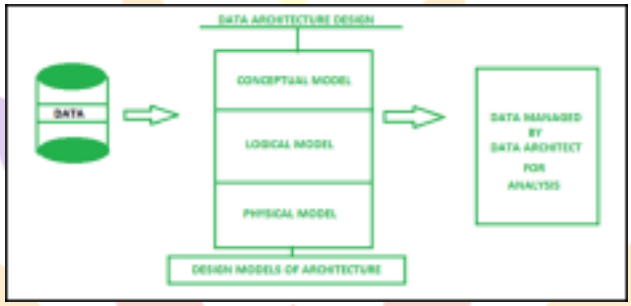
A data architecture should set data standards for all its data systems as a vision or a model of the eventual interactions between those data systems .

Data architectures address data in storage and data in motion; descriptions of data stores, data groups and data items; and mappings of those data artifacts to data qualities, applications, locations etc. Essential to realizing the target state, Data Architecture describes how data is processed, stored, and utilized in a given system. It provides criteria for data processing operations that make it possible to design data flows and also control the flow of data in the system.

The Data Architect is typically responsible for defining the target state, aligning during development and then following up to ensure enhancements are done in the spirit of the original blueprint. The Data Architect breaks the subject down by going through 3 traditional architectural processes: Conceptual model: It is a business model which uses Entity Relationship (ER) model for relation between entities and their attributes.

Logical model: It is a model where problems are represented in the form of logic such as rows and column of data, classes, xml tags and other DBMS techniques

. Physical model: Physical models holds the database design like which type of database technology will be suitable for architecture.



Various constraints and influences will have an effect on data architecture design. These include enterprise requirements, technology drivers, economics, business policies and data processing need.

Enterprise requirements: These will generally include such elements as economical and effective system expansion, acceptable performance levels (especially system access speed), transaction reliability, and transparent data management. In addition, the conversion of raw data such as transaction records and image files into more useful information forms through such features as data warehouses is also a common organizational requirement, since this enables managerial decision making and other organizational processes. One of the architecture techniques is the split between managing transaction data and (master) reference data. Another one is splitting data capture systems from data retrieval systems (as done in a data warehouse).

Technology drivers: These are usually suggested by the completed data architecture and database architecture designs. In addition, some technology drivers will derive from existing organizational integration frameworks and standards, organizational economics, and existing site resources (e.g. previously purchased software licensing).

Economics: These are also important factors that must be considered during the data architecture phase. It is possible that some solutions, while optimal in principle, may not be potential candidates due to their cost. External factors such as the business cycle, interest

rates, market conditions, and legal considerations could all have an effect on decisions relevant to data architecture.

Business policies: Business policies that also drive data architecture design include internal organizational policies, rules of regulatory bodies, professional standards, and applicable governmental laws that can vary by applicable agency. These policies and rules will help describe the manner in which enterprise wishes to process their data.

Data processing needs These include accurate and reproducible transactions performed in high volumes, data warehousing for the support of management information systems (and potential data mining), repetitive periodic reporting, ad hoc reporting, and support of various organizational

initiatives as required (i.e. annual budgets, new product development) The General Approach is based on designing the Architecture at three Levels of Specification.

Understand various sources of the Data:

Data can be generated from two types of sources namely Primary and Secondary Sources of Primary Data. Data collection is the process of acquiring, collecting, extracting, and storing the voluminous

amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis. In the process of big data analysis, "Data collection" is the initial step before starting to analyze the patterns or useful information in data.

The data which is to be analyzed must be collected from different valid sources. The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as "knowledge". Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.

Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection. Most of the data collected are of two types known as qualitative data which is a group of non-numerical data such as words, sentences mostly focus on behaviour and actions of the group and another one is quantitative data which is in numerical forms and can be calculated using different scientific tools and

sampling data.

Tools used in data analytics :

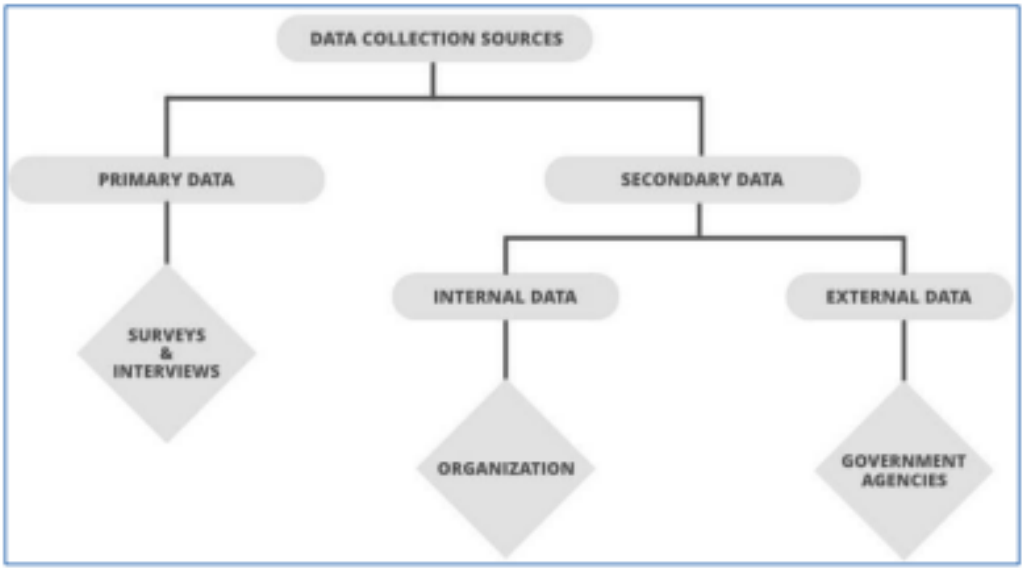
Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open-source or user-friendly, the top tools in the data analytics market are as follows.

- R programming
- Python
- Tableau Public
- Qlik View
- SAS
- Microsoft Excel
- Rapid Miner
- KNIME
- Open Refine
- Apache Spark

Missing data : (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data

Data duplication : is the process of creating one or more identical versions of data, either intentionally, such as for planned backups, or unintentionally

Data collection :



1. Primary data
2. Secondary data

1.Primary data:

The data collected must be according to the demand and requirements of the target audience on which analysis performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. Interview method:

· The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee.

· Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing.

· These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

· The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video.

· The survey method can be obtained in both online and offline model like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

3. Observation method:

· The observation method is same method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats.

· In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behaviour towards the products. The data obtained will be sent for processing.

Internal source:

These types of data can easily be found within the organization such as market record, as sales record, transactions, customer data, accounting resources, etc. The cost and time

consumption is less in obtaining internal sources.

- **Accounting resources**- This gives so much information which can be usedby the marketing researcher. They give information about internal factors.
- **Sales Force Report**- It gives information about the sales of a product. The information provided is from outside the organization.
- **Internal Experts**-These are people who are heading the various departments. They can give an idea of how a particular thing is working.
- **Miscellaneous Reports**-These are what information you are getting from operational reports. If the data available within the organization are unsuitable or in adequate,the marketer should extend the search to external secondary data sources.

External source:

The data which can't be found at internal organizations and can begained through external third party resources is external source data. The cost and time consumption are more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labour bureau, syndicate services, and other non-governmental publications.

1. **Government Publications-**
   - Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data.

It is an office which generates demographic data. It includes details of gender, age, occupation etc.

2. **Central Statistical Organization-**
   - This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO.
   - It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.
3. **Director General of Commercial Intelligence-**
   - This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

Data Management process :

Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost- effectively. The goal of data management is to help people, organizations, and connected things optimize the use of data within the bounds of policy and regulation so that they can make decisions and take actions that maximize the benefit to the organization.

Managing digital data in an organization involves a broad range of tasks, policies, procedures, and practices.

Cloud Computing means storing and accessing the data and programs on remote servers that are hosted on the internet instead of the computer's hard drive or local server. Cloud computing is also referred to as Internet-based computing, it is a technology where the resource is provided as a service through the Internet to the user. The data that is stored can be files, images, documents, or any other storable document.

The following are some of the Operations that can be performed with Cloud

Computing · Storage, backup, and recovery of data

· Delivery of software on demand

· Development of new applications and services

· Streaming videos and audio

Software as a Service is the most common cloud service used by organizations. A 2020 Virayo study found that 80 percent of organizations use one or more SaaS applications in their business. While using SaaS services, you don't have to install any software on your computer. Instead, you can easily access them on the cloud where they are stored. So, if you want to do some urgent work and do not have your laptop with you, all you need is an internet connection and a browser to access the required tools.

Features of PaaS

PaaS has several features. Some of them include:

· **Auto-scaling:** Since it is based on virtualization technology, resources can be scaled up and down at your convenience.

· **Resource sharing:** PaaS allows resource sharing amongst different development teams. · **Accessibility:** It also allows several users to access the platform with the same development application.

· **Time-saving:** PaaS offers developers pre-coded components and several development tools, which saves their time and resources.

· **Integrations:** PaaS allows the integration of databases and web services

Features of IaaS include:

· **Dynamic scaling:** IaaS allows dynamic and flexible scaling of resources as they are available in an as-a-service model.

· **Platform virtualization:** IaaS uses platform virtualization technology to provide cloud computing infrastructure.

· **Costs:** The services are available on a pay-as-you-go basis. Therefore you pay only for the resources you use.

· **Control:** IaaS users have complete control over their infrastructure and IT platform. Benefits of IaaS:

· **Scaling:** The IaaS services are available 24*7*365. You can easily scale globally and enhance application performance.

· **Enhanced security:** The data is secure and can only be accessed by authorized people. IaaS also allows you to keep backups in case of data loss.

· **Automation:** IaaS easily automates the deployment of various resources, such as networks, and servers.

· **Saves time and cost:** Users save a lot of time and cost since all the hardware maintenance is done by the vendor.

· **Flexibility:** IaaS vendors allow users to purchase the features they need and scale up and down at their convenience.

Data quality refers to the reliability, accuracy, completeness, and consistency of data. High-quality data is free from errors, inconsistencies, and inaccuracies, making it suitable for reliable decision making and analysis.

Data quality encompasses various aspects, including correctness, timeliness, relevance, and adherence to predefined standards. Organizations prioritize data quality to ensure that their information assets meet the required standards and contribute effectively to business processes and decision-making. Effective data quality management involves processes such as data profiling, cleansing, validation, and monitoring to maintain and improve data integrity.

Data Quality vs Data Integrity

Oversight of data quality is only one component of data integrity, which includes many other elements as well. Keeping data valuable and helpful to the company is the main

objective of data integrity. To achieve data integrity, the following four essential elements are necessary:

· Data Integration: The smooth integration of data from various sources is very much essential.

· Data Quality: A vital aspect of maintaining data integrity is verifying that the information is complete, legitimate, unique, current, and accurate.

· Location Intelligence: when location insights are included in the data, it gains dimension and therefore becomes more useful and actionable.

· Data Enrichment: By adding more information from outside sources, such customer, business, and geographical data, data enrichment may improve the context and completeness of data.

Outlier :

Outliers are data points that lie outside the majority of the data in a particular data set. These values might be much higher or lower in value than other points and may impact the results of the data analysis in ways that misrepresent the data sample. By learning how to identify and handle outliers, data analysts can increase the likelihood that their analysis will accurately reflect the validity and reliability of their results.

The 3 Different Types of Outliers

In statistics and data science, there are three generally accepted categories which all outliers fall into:

· Type 1: Global Outliers (aka Point Anomalies)

· Type 2: Contextual Outliers (aka Conditional Anomalies)

· Type 3: Collective Outliers

1. Global Outliers

1. Definition: Global outliers are data points that deviate significantly from the overall distribution of a dataset.

2. Causes: Errors in data collection, measurement errors, or truly unusual events can result in global outliers.

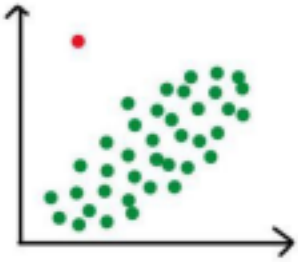
3. Impact: Global outliers can distort data analysis results and affect machine

learning model performance.

4. Detection: Techniques include statistical methods (e.g., z-score, Mahalanobis distance), machine learning algorithms (e.g., isolation forest, one-class SVM), and data visualization techniques.

5. Handling: Options may include removing or correcting outliers, transforming data, or using robust methods.

6. Considerations: Carefully considering the impact of global outliers is crucial for accurate data analysis and machine learning model outcomes



2. Collective Outliers

1. Definition: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.

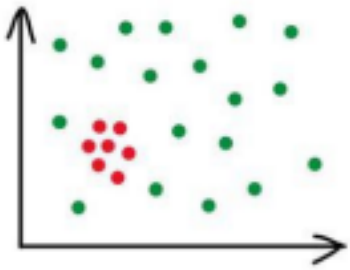
2. Characteristics: Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.

3. Detection: Techniques for detecting collective outliers include clustering algorithms, density-based methods, and subspace-based approaches.

4 Impact: Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.

5. Handling: Handling collective outliers depends on the specific use case and may involve further analysis of the group behavior, identification of contributing factors, or considering contextual information.

6. Considerations: Detecting and interpreting collective outliers can be more complex than individual outliers, as the focus is on group behavior rather than individual data points. Proper understanding of the data context and domain knowledge is crucial for effective handling of collective outliers



3. Contextual Outliers

1. Definition: Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup.

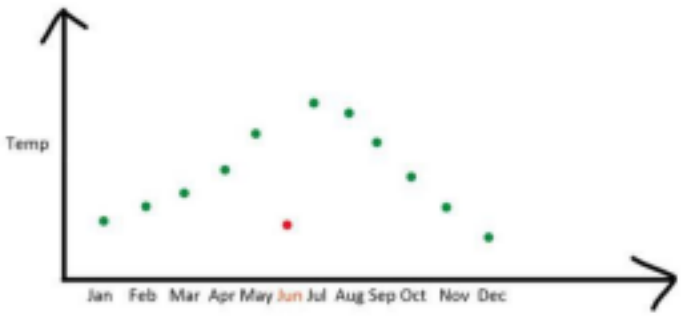
2. Characteristics: Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.

3. Detection: Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.

4. Contextual Information: Contextual information such as time, location, or other relevant factors are crucial in identifying contextual outliers.

5. Impact: Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.

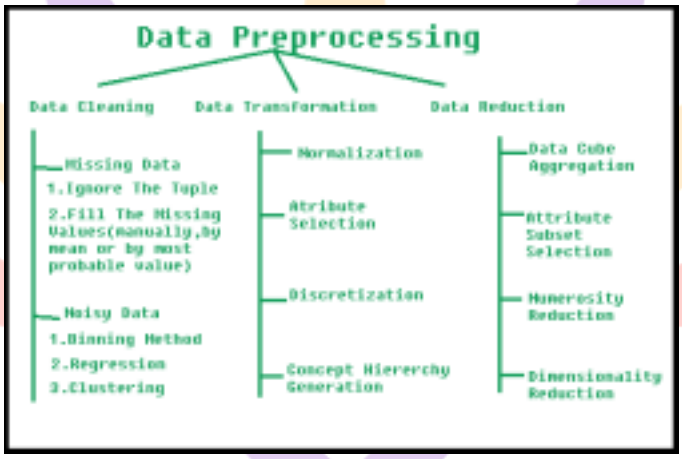
6. Handling: Handling contextual outliers may involve considering the contextual information, contextual normalization or transformation of data, or using context-specific models or algorithms. **7. Considerations:** Proper understanding of the context and domain-specific knowledge is crucial for accurate detection and interpretation of contextual outliers, as they may vary based on the specific context or subgroup being considered.



NARSIMHA REDDY ENGINEERING COLLEGE
UGC-AUTONOMOUS INSTITUTION
An Autonomous Institute
NAAC Accreditation 'A' Grade
Accredited by NBA
Approved by AICTE, Affiliated to JNTUH

Data pre processing:

Preprocessing in Data Mining: Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Data Preprocessing

1. Data Cleaning: The data can have many irrelevant and missing parts. To handle thispart, data cleaning is done. It involves handling of missing data, noisy data etc.

· Missing Data: This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

o Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

o Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

· Noisy Data: Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

o Binning Method: This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

o Regression:Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

o Clustering: This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

· Normalization: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

· Attribute Selection: In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

· Discretization: This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

· Concept Hierarchy Generation: Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

3. Data Reduction: Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

· Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual

information, and principal component analysis (PCA).

· Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

· Sampling: This involves selecting a subset of data points from the dataset. Sampling is often

used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

· Clustering: This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

· Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gif compression.

How is Data Preprocessing Used?

This we have earlier noted is one of the reasons data preprocessing is important in the earlier stages of the development of machine learning and AI applications. While in AI context data preprocessing is applied in order to optimize the methods used to cleanse, transform and structure the data in a way that will enhance the accuracy of a new model with less computing power used.

An excellent data preprocessing step will help develop a set of components or tools that can be utilized to quickly prototype on a set of ideas or even run experiments on improving business processes or customer satisfaction. For instance, preprocessing can enhance the manner in which data is arranged for a recommendation engine by enhancing the age ranges of customers that are used for categorisation.

It can also make the process of developing and enhancing data easier for more enhanced BI which is beneficial to the business. For instance, small size, category or regions of the customers may have different behaviors across regions. Backend processing the data into the correct formats might enable BI teams to integrate such findings into BI dashboard.

In a broad concept, data preprocessing is a sub-process of web mining which is used in customer relationship management (CRM). There's usually the possibility of pre-processing of the Web usage logs in order to arrive at meaningful data sets referred to as user transactions which are actually a set of groups of URL references. Sessions may be stored to make user identification possible as well as the websites requested and their sequence and time of use.

NARSIMHA REDDY ENGINEERING COLLEGE
UGC-AUTONOMOUS INSTITUTION
Your roots to success...
An Autonomous Institute
NAAC Accreditation 'A' Grade
Accredited by NBA
Approved by AICTE, Affiliated to JNTUH

UNIT -II

Data Analytics: Introduction to Analytics, Introduction to Tools and Environment, Application of Modeling in Business, Databases & Types of Data and Variables, DataModeling Techniques, Missing Imputations etc. Need for Business Modeling

Introduction to Analytics :

Data has been the buzzword for ages now. Either the data being generated from large-scale enterprises or the data generated from an individual, each and every aspect of data needs to be analyzed to benefit yourself from it.

Why is Data Analytics important?

1. Data Mining

Most simply stated, data mining is a process used to extract usable data from a large dataset. Data mining involves data collection, warehousing and computer processing. In order to segment and evaluate the data, data mining uses advanced algorithms.

Real-Life Scenario: Data mining is often used in the health care industry during patient clinical trials. The algorithms can evaluate behavioral patterns of large amounts of data for interpretation, knowledge building and decision making.

2. Text Analytics

Text analytics is the process of drawing meaning out of written communication. Usually, text analytics software relies on text mining and natural language processing (NLP) algorithms to find patterns and meaning.

Real-Life Scenario: Text analytics is used to build the auto-correct function on your mobile device. It will not only correct your spelling, but also predict what you're going to type next based on linguistic analysis and data pattern recognition.

3. Data Visualization

Data visualization presents a clear picture of what the data actually means. Using bar graphs, pie charts, tables and other visuals, data visualization makes the data easier for those making business decisions to comprehend.

Real-Life Scenario: Data visualizations are part of our everyday lives on IoT devices – and you probably don't even realize it. Think about the exercise rings on your smartwatch, the energy-use trends from your smart thermostat and the weekly screen time charts on your phone.

4. Business Intelligence

Business intelligence (BI) is the end game. It leverages analytics tools to convert data to actionable insights. Often paired with data visualization techniques, BI provides decision makers with detailed intel about the state of the business

Data Analytics Tool	**How It's Used**
Artificial Intelligence	Makes decisions that can provide a plausible likelihood in achieving a goal
NoSQL Database	Delivers a method for accumulation and retrieval of data
R Programming	Assists data scientists in designing statistical software
Data Lakes	Accumulates data without transforming it into structured data
Predictive Analytics	Predicts future behavior via prior data
Apache Spark	Generates big data transformation via Python, R, Scala and Java
Prescriptive Analytics	Provides guidance about what to do to achieve a desired outcome

In-Memory Database	Saves time by omitting the requirements to access hard drives
Hadoop Ecosystem	Ingests, stores, analyzes and maintains large data sets
Blockchain	Distributed ledger technologies have proven valuable in managing data challenges
Microsoft Excel	Aggregates data to create reports and easy-to-use dashboards

Different Components of Data Analytics

Generally, there are three stages of data analytics: collection and storage, process and organization, and finally, analysis and visualization. In other words, it starts with identifying the data, then progresses to organizing it in a way that makes sense, and ends with identifying patterns and trends that mean something.

But when it comes to business, we can take these stages a bit further. To start, before we begin sourcing data, we need to engage in some business analytics. We need to ask questions about our objectives and desired outcomes before we identify the type of data we need to gather.

We also need to consider the people and the processes making this analysis happen. Do we need more qualified people? Do we need more training? And how will we share our findings internally and externally?

As businesses are continuing to make digital transformations, the components of data analytics can be seen more as a comprehensive data strategy, with the following components:

1. Address the specific business needs.
2. Determine where the data exists and how it will be gathered.
3. Take inventory of the technical infrastructure needed to support the sourcing

of data. 4. Identify how to turn data into actionable insights.

5. Look at the necessary processes and required skillsets of your people.

6. Ensure the right people have access to the right data.

7. Define the business value by creating a roadmap.

Data Analytics Applications

Data in business :

In Data Analytics there are many advantages of data, but without the proper data analytics tools and processes, you can't access these benefits. Raw data is also very important and you need data analytics to unlock the potential of raw data and converted into useful information for the business.

Example –

Record of the potential customer, records of customers like name, address.

Data in healthcare :

Data is extremely useful in this field of medical and healthcare. Most of the medical devices are big data-oriented. In Data Analytics uses of data has gone to such an extent that in the healthcare sector each record or you can say data is very essential where doctors can check person through the heart and temperature monitoring watch which is critical information of any patients and kept to be as data fitted on patient's hand and prescribe him with related medicines. Example –

Patient records like name, address, contact no. etc., treatment records, Records of Doctor's profile are the examples in healthcare.

Data in media and entertainment :

The business model runs on collecting and creating the content, further analyzing it, then marketing and distribution of the content. We can run through customer's data along with observable data and gather even minute information to create a customer's detailed profile. The benefits of big data in the media and entertainment industry include forecasting what the target audience wants, planning, optimization, expanding acquisition, and retention suggest content on-demand and new. Example –

Records of the team, the time duration of media project, location, etc.

Data in transportation :

Data in transportation is very crucial. For proper communication and for proper synchronization of transport medium you need data and to analyze the information you need data analytics. Data potential is to analyze how many passengers traveled from any source to destination and with the help of data analytics it can be processed in real-time for the smooth functioning of transportation. Example –

feedback of customer, transport time, source and destination records, customer traveled history, etc.

Data in banking :

Banking is a very crucial sector. Data here is very beneficial and helps in fraud detection in the banking system. Using big data, we can search for all the illegal activities that

have taken place and can identify the misuse of credit and debit cards, business precision, you can say for customer statistics modification, and in public analytics for business.

Example –

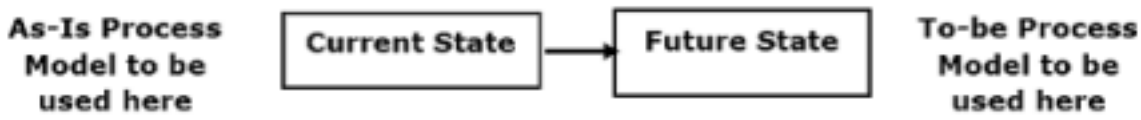
Employee records, Bank name address, and branch name, customer account records, transaction history, etc.

Application of Modelling in Business :

Purpose of Business Modelling

Business modelling is used to design current and future state of an enterprise. This model is used by the Business Analyst and the stakeholders to ensure that they have an accurate understanding of the current "As-Is" model of the enterprise.

It is used to verify if, stakeholders have a shared understanding of the proposed "To-be of the solution.



Analyzing requirements is a part of business modelling process and it forms the core focus area. Functional Requirements are gathered during the "Current state". These requirements are provided by the stakeholders regarding the business processes, data, and business rules that describe the desired functionality which will be designed in the Future State.

Databases & Types of Data and variables

Relational Database Management System: RDBMS is a software system used to maintain relational databases. Many relational database systems have an option of using the SQL.

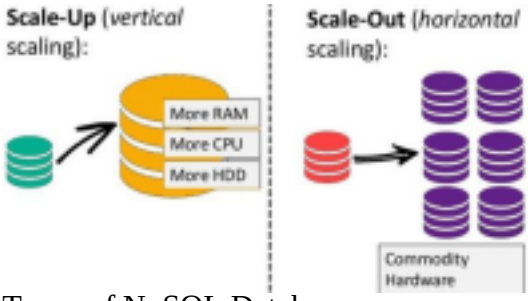
NoSQL:

· NoSQLDatabaseisanon-relationalDataManagementSystem,thatdoesnotrequire a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs. NoSQL is used for Big data and real-time web apps. For example, companies like Twitter, Facebook and Google collect terabytes of user data every single day.

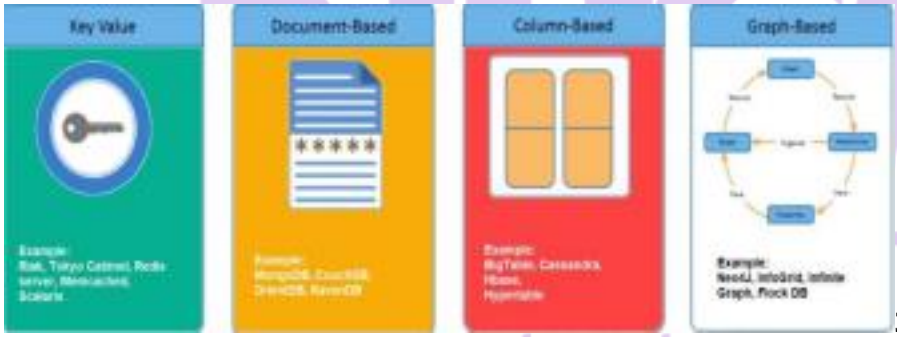
· **NoSQL database** stands for "Not Only SQL" or "Not SQL." Though a better term would be "NoREL", NoSQL caught on. Carl Strozz introduced the NoSQL concept in 1998. · Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.

· The concept of NoSQL databases became popular with Internet giants like Google, Facebook, Amazon, etc. whodealwith hugevolumesof data. The systemresponse time becomes slow when you use RDBMS for massive volumes of data.

· To resolve this problem, we could"scale up"our systems byupgradingour existing hardware. This process is expensive. The alternative for this issue is to distribute databaseloadonmultiplehostswhenevertheloadincreases.Thismethodisknown as "scaling out."



Types of NoSQL Databases:



30

Differences between SQL and NoSQL :

1. SQL	2. NoSQL

3. RELATIONAL DATABASE MANAGEMENT SYSTEM (RDBMS)	**4.** Non-relational or distributed database system.
5. These databases have fixed or static or predefined schema	**6.** They have a dynamic schema
7. These databases are not suited for hierarchical data storage.	**8.** These databases are best suited for hierarchical data storage.
9. These databases are best suited for complex queries	**10.**These databases are not so good for complex queries
11. Vertically Scalable	**12.**Horizontally scalable
13. Follows ACID property	**14.**Follows CAP(consistency, availability, partition tolerance)
15. Examples: MySQL, PostgreSQL, Oracle, MS-SQL Server, etc	**16.**Examples: MongoDB, HBase, Neo4j, Cassandra, etc

Data Modelling Techniques in Data Analytics :

Importance of Data Modeling in Data Warehouses

· Improved Data Quality: A well-structured data model helps ensure data consistency, accuracy, and reliability, which are critical for generating meaningful insights.

· Efficient Data Retrieval: By organizing data into logical structures, data modeling enables faster and more efficient data retrieval, which is essential for timely decision-making.

· Scalability: A robust data model allows for easy scaling of the data warehouse as the volume of data grows, ensuring that performance remains optimal.

· Reduced Redundancy: Proper data modeling helps eliminate data redundancy, reducing storage costs and simplifying data management.

Types of Data Models

Data modeling for data warehouses typically involves three main types of models:

1. Conceptual Data Model

· Purpose: Provides a high-level overview of the business entities and their relationships without going into technical details.

· Components: Entities, relationships, and attributes.

· Example: A conceptual model might define entities like "Customer," "Product," and "Sales" and illustrate the relationships between them.

2. Logical Data Model

· Purpose: Represents the logical structure of the data, including the relationships between entities and the data types for each attribute, without considering physical storage.

· Components: Tables, columns, relationships, and constraints.

· Example: A logical model might define tables such as "Customer," "Product," and "Sales" with their respective columns like "CustomerID," "ProductID," and "SaleDate."

3. Physical Data Model

· Purpose: Specifies how the data will be physically stored in the database, including indexing, partitioning, and data storage mechanisms.

· Components: Tables, indexes, partitions, and storage settings.

· Example: A physical model might define storage settings for the "Sales" table, such as partitioning by date to improve query performance

Data Modelling Techniques :

Hierarchical Data Models

Hierarchical data models represent data in a tree-like structure, where information is organized in levels with parent-child relationships. A simplified example is a family tree: the highest level is the parent, followed by branches for children, then grandchildren, and so on. Each element in the hierarchy is called a node, and connections between them are called links.

Hierarchical data models excel at representing data with a natural parent-child hierarchy, providing an intuitive and efficient way to organize, store, and access information.

Relational Data Models

Relational data models, the foundation of relational databases (RDBMS), offer a structured and organized way to represent and manage data. They organize data into relations, also known as tables, where each relation holds records (rows) related to a specific entity or concept.

Relational data models provide a well-established and versatile approach for managing data, particularly for structured and well-defined information. Their advantages make them a popular choice for a wide range of applications across various sectors.

Object-oriented Data Models

Object-oriented data models (OODMs) offer a unique perspective on organizing and managing data, taking inspiration from the principles of object-oriented programming (OOP).

OODMs are similar to graph data models. The main difference is that OODMs focus on individual objects and their internal coherence, with relationships emerging through object interactions.

Need for Business Modelling :

The main purpose of Business Model is to assist the company in developing a plan which will establish and validate critical points of line in the business. This includes activities such as resources, customer relationships, revenue and expenses.

Business Models are important because it helps because

1. The target market is clear

Business Model guides you through the process for determining value preposition and will make you understand how your product can satisfy the customer. The clear and simple business model helps to determine target which will prioritize.

2. The product created is fixed

By following a precise model, there is a lot of clarity in the business model and creating products. The system becomes transparent.

3. Preparing a strategy becomes easier

The business model helps to determine the business strategy automatically. The system does not attract consumers but it teaches how to significantly develop close relationships with producers.

4. Anticipating Competition

Without a business model, the companies will find difficult to find a business position in the market. Due to this the company faces competition. With proper business plans , companies can make strategies for acquiring resources and selling best products to the consumers. So it is very important to determine the right business model.

NARSIMHA REDDY ENGINEERING COLLEGE
UGC-AUTONOMOUS INSTITUTION
An Autonomous Institute
NAAC Accreditation 'A' Grade
Accredited by NBA
Approved by AICTE, Affiliated to JNTUH
Your roots to success...

UNIT - Ill

Regression - Concepts, Blue property assumptions, Least Square Estimation, Variable Rationalization, and Model Building etc. Logistic Regression: Model Theory, Model fit Statistics, Model Construction, Analyticsapplications to various Business Domains etc.

THE SIMPLE LINEAR REGRESSION MODEL

We consider the modeling between the dependent and one independent variable. When there is only one independent variable in the linear regression model, the model is generally termed as simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

The linear model

Consider a simple linear regression model

$y = \beta_0 + \beta_1 X + \varepsilon$

where y is termed as the dependent or study variable and X is termed as independent or explanatory variable. The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as intercept term and the parameter β_1 is termed as slope parameter. These parameters are usually called as **regression coefficients**. The unobservable error component ε accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y . There can be several reasons

for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherit randomness in the observations etc. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant variance σ^2 . Later, we will additionally assume that ε is normally distributed.

The independent variable is viewed as controlled by the experimenter, so it is considered as non stochastic whereas y is viewed as a random variable with

$E(y) = \beta_0 + \beta_1 X$

and

$Var(y) = \sigma^2$.

$Var(y \mid x) = \sigma^2$.

When the values of

β_0, β_1 and σ^2 are known, the model is completely described. The parameters β^0 , β_1 and

σ^2 are generally unknown in practice and ε is unobserved. The determination of the statistical model

$y = \beta_0 + \beta_1 X + \varepsilon$ depends on the determination (i.e., estimation) of β_0, β_1 and σ^2. In order to know the

values of these parameters, n pairs of observations $(x_i, y_i)(i = 1,..., n)$ on (X, y) are observed/collected and are used to determine these unknown parameters.

Various methods of estimation can be used to determine the estimates of the parameters.

$\bar{x} =$ Among them, the methods of least squares and maximum likelihood are the popular methods of estimation.

BLUE PROPERTY ASSUMPTIONS

- B-BEST

L-LINEAR

U-UNBIASED

E-ESTIMATOR

An estimator is BLUE if the following hold:

It is linear (Regression model)

It is unbiased

It is an efficient estimator(unbiased estimator with least variance)

Linearity

An estimator is said to be a linear estimator of (f3) if it is a linear function of the sample observations $+X_2 + + X_n$

N Sample mean is a linear estimator because it is a linear function of the X values.

UNBIASEDNESS

A desirable property of a distribution of estimates is that its mean equals the true mean of the variables being estimated formally, an estimator is an u n b i a s e d e s t i m a t o r if its Sampling distribution has as its expected value equal to the true value of population.

We also write this as follows:

Similarly, if this is not the case, we say that the estimator is biased

TWO TYPES OF ESTIMATORS

A point estimate of a population parameter is a single value of a statistic.

POINTER ESTIMATORS

For example, the sample mean x is a point estimate of the population mean p. Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.

Interval Estimators

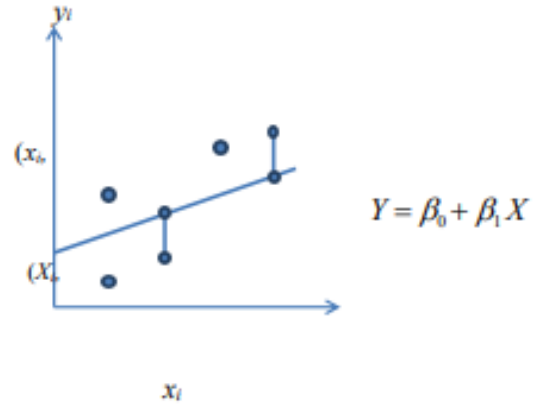
An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, *a* < x < *b* is an interval estimate of the population mean p. It indicates that the population mean is greater than *a* but less than *b*.

LEAST SQUARES ESTIMATION

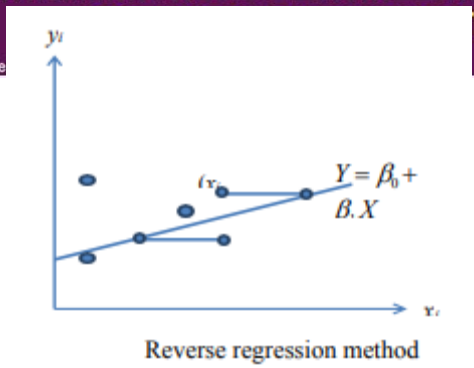
Suppose a sample of n *sets* of paired observations (x_i, y_i) (i = 1, 2,..., n) are available. These observations are assumed to satisfy the simple linear regression model and so we can write $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (i = 1, 2,..., n).

The method of least squares estimates the parameters β_0 and β_1 by minimizing the sum of squares of Difference between the observations and the line in the scatter diagram. Such an idea is viewed from different perspectives. When the **vertical difference** between the observations and the line in the scatter

Diagram is considered and its sum of squares is minimized to obtain the β_0 and β_1 , the meth estimates of is known as **direct regression**.



Alternatively,the sum of squares of difference between the observations and the line in horizontal direction in the scatter diagram can be minimized to obtain the estimates of β0 and β1 . This is known as reverse (or Inverse) regression method.



Reverse regression method

Instead of horizontal or vertical errors, if the sum of squares of perpendicular distances between the observations and the line in the scatter diagram is minimized to obtain the estimates of method is known as **orthogonal regression** or **major axis regression method.** y_i

β_0 and β_1 , theMajor axis regression method

Instead of minimizing the distance, the area can also be minimized. The **reduced major axis regression method** minimizes the sum of the areas of rectangles defined between the observed data points and the nearest point on the line in the scatter diagram to obtain the estimates of regression coefficients. This is shown in the following figure:

y_i

$(x_i y_i)$

$Y = \beta_0 + \beta_1 X$

(X_i, Y_i)

x_i

Reduced major axis method

The method of **least absolute deviation regression** considers the sum of the absolute deviation of the observations from the line in the vertical direction in the scatter diagram as in the case of direct regression to obtain the estimates of β_0 and β_1 .

No assumption is required about the form of probability distribution of ε_i in deriving the least

squares estimates. For the purpose of deriving the statistical inferences only, we assume that ε_i' *s* are random variable with $E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j (i, j = 1, 2,..., n)$. This assumption is

needed to find the mean, variance and other properties of the least squares estimates. The assumption that ε_i' *s* are normally distributed is utilized while constructing the tests of hypotheses and confidence intervals of the parameters.

LOGISTIC REGRESSION

Logistic regression, or Logistic regression, or Logistic model is a regression model where the dependent variable (DV) is categorical. Logistic regression was developed by statistician David Cox in 1958.

> Ordinary Least Square

> Maximum Likelihood Estimation

The ordinary least squares, or OLS, can also be called the linear least squares.

OLS and MLE:

OLS -> Ordinary Least Square

MLE -> Maximum Likelihood Estimation

The ordinary least squares, or OLS, can also be called the linear least squares. This is a method for approximately determining the unknown parameters located in a linear regression model. According to books of statistics and other online sources, the ordinary by minimizing the total of squared vertical distances between the observed responses within the dataset and the responses predicted by the linear approximation. Through a simple formula, you can express the resulting estimation of the linear regression model.

ANALYTICS APPLICATIONS TO VARIOUS BUSINESS DOMAINS

Predictive Analytics is an art of predicting future on the basis of past trend.

It is a branch of Statistics which comprises of Modelling Techniques, Machine Learning & Data Mining.

Predictive Analytics is primarily used in Decision Making.

What and Why analytics:

Analytics is a journey that involves a combination of potential skills, advanced technologies,

applications, and processes used by firm to gain business insights from data and statistics. This is done to perform business planning.

Reporting Vs Analytics:

Reporting is presenting result of data analysis and Analytics is process or systems involved in analysis of data to obtain a desired output.

Introduction to tools and Environment:

Analytics is now days used in all the fields ranging from Medical Science to Aero science to Government Activities.

Data Science and Analytics are used by Manufacturing companies as well as to develop their business and solve various issues by the help of historical data base.

Tools are the software that can be used for Analytics like SAS or R. While techniques are the procedures to be followed to reach up to a solution.

Various steps involved in Analytics:

Access
Manage
Analyze Report

Various Analytics techniques are:

Data Preparation
Reporting, Dashboards & Visualization
Segmentation Icon
Forecasting
Descriptive Modelling
Predictive Modelling

APPLICATION OF MODELLING IN BUSINESS:

A statistical model embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population. A model represents, often in considerably idealized form, the data-generating process. Signal processing is an enabling technology that encompasses the fundamental theory, applications, algorithms, and implementations of processing or transferring information contained in many different physical, symbolic, or abstract formats broadly designated as signals. It uses mathematical, statistical, computational, heuristic, and linguistic representations, formalisms, and techniques for representation, modelling, analysis, synthesis, discovery, recovery, sensing, acquisition, extraction, learning, security, or forensics. In manufacturing statistical models are used to define Warranty policies, solving various conveyor related issues, Statistical Process Control etc.

Databases & Type of data and variables:

A data dictionary, or metadata repository, as defined in the IBM Dictionary of Computing, is a

"centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format".

Data can be categorized on various parameters like Categorical, Type etc.

Data can be categorized on various parameters like Categorical, Type etc.

Data is of 2 types – Numeric and Character. Again numeric data can be further divided into sub group of – Discrete and Continuous.

Again, Data can be divided into 2 categories – Nominal and ordinal.

Also based on usage data is divided into 2 categories – Quantitative and Qualitative

Data Modelling Techniques Overview:

Regression analysis mainly focuses on finding a relationship between a dependent variable and one or more independent variables.

Predict the value of a dependent variable based on the value of at least one independent variable.

It explains the impact of changes in an independent variable on the dependent variable. Y = f(X, β) where Y is the dependent variable unknown coefficient

Missing Imputations:

In R, missing values are represented by the symbol NA (not available). Impossible values (e.g., dividing by zero) are represented by the symbol NaN (not a number). Unlike SAS, R uses the same symbol for character and numeric data.

NARSIMHA REDDY ENGINEERING COLLEGE
UGC-AUTONOMOUS INSTITUTION
NRCM Your roots to success...
An Autonomous Institute
NAAC Accreditation 'A' Grade
Accredited by NBA
Approved by AICTE, Affiliated to JNTUH

UNIT-IV

Object Segmentation: Regression Vs Segmentation – Supervised and Unsupervised Learning, Tree Building – Regression, Classification, Overfitting, Pruning and Complexity, Multiple Decision Trees etc. Time Series Methods: Arima, Measures of Forecast Accuracy, STL approach, Extract features from generated model as Height, Average Energy etc and Analyze for prediction

Regression Vs Segmentation:

Regression:

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression refers to average relationship between two or more variables. One of these variables is called the dependent or the explained variable and the other variable is independent or the explaining variable. Regression is one of the statistical method that is used to estimate the unknown of one variable from the known value of the related variable.

For Example regression helps finance and investment professionals and other business professionals. It helps in predicting company sales depending upon weather, previous sales, GDP growth etc., Regression considers a set of random variables to find the mathematical relationship between them. The relationship is mostly in the form of straight line which approximates the separate data points. There are two types of regression namely simple linear regression and multiple linear regression.

Simple linear regression represents the relationship between two variables where one of them is independent variable X and other variable is dependent variable Y.

Multiple linear regression model that contains multiple independent variables is referred to as multiple regression model. An equation

Segmentation is the process of segmenting the data according to the company's requirement to refine the analyses based on the defined context through certain tools. The Main objective is to understand the customers better and then obtain actionable data to improve the outcome. It allows to filter the analyses depending upon some elements.

Object Segmentation is the process of segmenting the moving objects. That means locating the objects and its boundaries from background. The frames that are captured contain noises that are extracted from captured frames. The objects after filtering are passed to optical flow vectors for thresholding operation. The unwanted objects are then removed. In filtering process, the holes are created. They are closed through a morphological operation.

NARSIMHA REDDY ENGINEERING COLLEGE
UGC-AUTONOMOUS INSTITUTION
An Autonomous Institute
NAAC Accreditation 'A' Grade
Accredited by NBA
Approved by AICTE, Affiliated to JNTUH
Your roots to success...

Segmentation:

Segmentation is the Process of segmenting the data according to the company's requirement to refine the analyses based on the defined context through certain tools.

Objective segmentation is the process of segmenting the moving objects. Locating the objects and its boundaries from background. The frames that are captured contain noise that are extracted from captured frames. The Objects after filtering are passed to optical flow vectors for thresholding operation. The unwanted objects are then removed. In the filtering process the holes are created. They are closed through morphological operation.

Supervised Learning:

- Supervised learning allows you to collect data or produce a data output from the previous experience.
- Helps you to optimize performance criteria using experience
- Supervised machine learning helps you to solve various types of real-world computation problems

Example:

For example, you want to train a machine to help you predict how long it will take you to drive home from your workplace. Here, you start by creating a set of labeled data. This data includes

- Weather conditions
- Time of the day
- Holidays

All these details are your inputs. The output is the amount of time it took to drive back home on that specific day.

Types of Supervised Machine Learning Techniques

Regression:

Regression technique predicts a single output value using training data.

Example: You can use regression to predict the house price from training data. The input variables will be locality, size of a house, etc.

Classification:

Classification means to group the output inside a class. If the algorithm tries to label input into two distinct classes, it is called binary classification. Selecting between more than two classes is referred to as multiclass classification.

Example: Determining whether or not someone will be a defaulter of the loan.

Strengths: Outputs always have a probabilistic interpretation, and the algorithm can be regularized to avoid overfitting.

Weaknesses: Logistic regression may underperform when there are multiple or non-linear decision boundaries. This method is not flexible, so it does not capture more complex relationships.

Unsupervised Learning

 Unsupervised machine learning finds all kind of unknown patterns in data. Unsupervised methods help you to find features which can be useful for categorization.
 It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
 It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

Example:

She knows and identifies this dog. A few weeks later a family friend brings along a dog and tries to play with the baby.

Baby has not seen this dog earlier. But it recognizes many features (2 ears, eyes, walking on 4 legs) are like her pet dog. She identifies a new animal like a dog. This is unsupervised learning, where you are not taught but you learn from the data (in this case data about a dog.) Had this been supervised learning, the family friend would have told the baby that it's a dog.

Unsupervised learning problems further grouped into clustering and association problems.

Clustering

Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters (groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

Association

Association rules allow you to establish associations amongst data objects inside large databases. This unsupervised technique is about discovering exciting relationships between variables in large databases. For example, people that buy a new home most likely to buy new furniture.

Other Examples:
 A subgroup of cancer patients grouped by their gene expression measurements Groups of shopper based on their browsing and purchasing histories Movie group by the rating given by movies viewers

Supervised Learning Unsupervised Learning

Supervised learning algorithms are trained using labeled data.

Supervised learning model takes direct feedback to check if it is predicting correct output or not.
Unsupervised learning algorithms are trained using unlabeled data.

Unsupervised learning model does not take any feedback.

Supervised learning model predicts the output. Unsupervised learning model finds the hidden patterns in data.

In supervised learning, input data is provided to the model along with the output.

The goal of supervised learning is to train the model so that it can predict the output when it is given new data.

Supervised learning needs supervision to train the model.

Supervised learning can be categorized in Classification and Regression problems.
In unsupervised learning, only input data is provided to the model.

The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.

Unsupervised learning does not need any supervision to train the model.

Unsupervised Learning can be classified in Clustering and Associations problems .

Supervised learning can be used for those cases where we know the input as well as corresponding outputs.
Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.

Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.

Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.
Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.

It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

Supervised and Unsupervised Learning:

[redacted]
Process Input and Output Variables are Provided
Input Data The Algorithms are trained through labelled data.
Number of Classes Number of classes is known already
Only Input Data is Provided

The Algorithms are trained through unlabelled data. Number of classes is not known.
[redacted]
Use of Data It uses Training data to learn a link in between Input and Outputs
Algorithms Used Support Vector Machines, Random Forest, Neural Network, Classification Trees, Linear and Logistics Regression
Real Time Learning Learning methods works offline
Accuracy of Results It is Accurate and Trusty worthy method
It does not need any output data.

Unsupervised algorithms are divided into categories like k-means, cluster algorithms, hierarchal clsutering

Learning methods works in real time
It is Less accurate and trusty worthy method.

Main Drawback	Classification of Bigdata is a	Because data used in
	challenging task	unsupervised learning is labelled and unknown it is difficult to obtain precise information regarding the data sorting.

TREE BUILDING

Decision Tree is a diagrammatic representation of alternative courses of action and sequence of states of nature.

The Various steps involved in decision tree analysis are

1. Determine the number of decisions to be taken and the alternative strategies available for each decision in sequential manner.
2. Determine the outcome(or event) which may occur from each alternative strategy. 3. Construct a tree diagram representing the order in which decisions are taken and outcomes are occurring. The decision tree diagram begins from left side and move towards right side.
4. Determine the Probabilities of occurrences of each state of nature.
5. Determine the pay off values for each pair (or combination) of state of nature and course of action.
6. Calculate expected pay off value for each course of action starting from right side of the decision trees.
7. Select the course of action (or alternative strategy) with the best expected pay off value.
8. With backwards from last decision point to first decision point and at each

Time Series Methods:

The time series data is ordered sequence of observations on quantative variable that is measured over equally spaced time interval. The time series are used in statistics, signal Processing, econometrics, pattern recognition, weather forecasting.

The time series analysis is used to analyse the time series data and to forecast the future value of variable under consideration. The data in time series analysis contain set of identifiable components and random errors that make the pattern difficult to identify.

ARIMA is one of the time series methods. ARIMA stands for Autoregressive Integrated Moving Average. In time series analyses it is generalization. This model is fitted to time

series data to understand the data or to predict future points in the series. They are applied in certain cases where the data shows evidence of non stationary and where the initial differencing step is applied to minimize the non stationary.

ARIMA is well known stochastic method that is used to analyze the time series data sets. It contains three time series components such AR (Auto-Regressive), I (Integrated), MA (Moving Average). Every component minimizes the final residuals indicated as p, d, q respectively. Integrated (I) is initial step in ARIMA to extract trend data. This is possible by differencing the data from earlier values. The first step is indicated by (0,1,0) and the second step is indicated by (0,2,0). This continues until data becomes trend less. The time series will become trend less after differencing. The second step in ARIMA is auto regression to analyze the time series data. Once this data becomes stationary the AR component gets activated. The auto regressive step will extract the previous value from present value. This is obtained through simple linear regression by considering independent or predictor variables as it time lagged values.

Y=C+Y+Y+Y…..+,ib

OVERFITTING

The Decision Trees will be at high risk to overfit the training data to a high degree when they are not pruned.

Over fitting might occurs when models select the noise or errors in training data set. Therefore overfitting will be seen viewed as performance gap between training and test data.

A General method to reduce overfitting in decision trees is decision tree pruning. There are two methods namely

1. Post Pruning
2. Pre Pruning

The above figure shows relationship between Tree Depth and Overfitting The following

figure shows the Reduced Error Pruning

STL APPROACH:

STL Stands for Seasonal and Trend Decomposition using Loess. Loess is a method used to estimate the non linear relationships. STL is versatile, robust ad stastical method of decomposing the time series data into three components as Trend, Seasonality and residual.

The purpose of development of STL is to build a decomposition procedure and a computer implementation that can satisfy the below criteria.

1. A Simple design and a straight forward use.

2. Easy computer implementation and a fast computation for long time series.

3. Specification of observations for each cycle of seasonal component to an integer greater than 1.

4. Flexible to specify the variation in trend and seasonal components.

5. Robust trend and seasonal components that cannot be distorted by transient and aberrant behaviour of data.

6. It should be able to decompose the series with missing values.

STL contains a set of smoothing operations that employ same smoother locally weighted regression or loess with only one exception.

STL has various parameters that must be selected by the data analysts.

$n_{(p)}$= Number of observations in every cycle of seasonal component. $n_{(i)}$= Number of Passes through inner loop.

$n_{(o)}$= Number of robustness iterations of outer loop.

$n_{(l)}$= Smoothing parameters for low pass filter

$n_{(t)}$= Smoothing parameter for trend component.

$n_{(x)}$= Smoothing parameter for seasonal component.

STL can be implemented in any environments for the purpose of graphics and data analysis.

Pruning:

Pruning is a technique that is used to reduce the size of decision tress by removing parts of the tree which provide less power for classifying instances. It reduces the complexity of final classifier and even improves the accuracy by reduction of over fitting.

Pruning occurs either in top-down or bottom-up. The top-down pruning traverses the nodes

and trims the sub tree starting at root. The bottom up pruning begins at leaf nodes.

Pruning when applied on decision trees will remove one or more sub trees from it. There are various methods for decision tree pruning .They replaces the sub tree with leaf if classification accuracy is not reduced over pruning data set. Pruning increases number of classification errors on training set but improves classification accuracy on unseen data The Pruning techniques can be divided into two groups. The methods in first group will compute the probability of sub tree misclassification and then make the pruning decision through an independent test set called pruning data set. In Second group the iterative grow and prune method is used while creating a tree. These pruning techniques are as follows.

1. Cost Complexity Pruning
2. Reduced Error Pruning
3. Critical Value Pruning
4. Minimum Error Pruning
5. Pessimistic Error Pruning
6. Error Based Pruning
7. Optimal Pruning
8. Minimum Description Length Pruning

Extract features from generated model as Height, Average Energy etc and Analyze for prediction

Feature extraction is the process of extracting useful characteristics from data. It calculates the values from input images. A feature is also called as descriptor. It is defined as a function of one or more measurements by specifying certain quantifiable property of complete image or sub image or of single object. Three methods of feature extraction are performed that lead to three different results for every classification. They are compared each other. The best feature extraction method for very two classifications is used for the part and complete system together.

For example consider an image model that is generated. The feature extraction focuses around the measurement of geometric properties and surface characteristics of regions. The features extracted from this image model are

[redacted] Y h,w, A, L,R Height, width, area, roundness, perimeter

[redacted]

I, R, G,B	K	Deviation Contrasts

The features provide relevant information for classification. To reduce the computational time required, in pattern recognition process it is necessary to select features suitable for classification.

The extracted features are later on analysed for prediction. A Predictor is used for this purpose. The model is trained on labelled examples can be good and bad quality to retrieve quality classification model. A model is trained on labelled examples of salient and non salient images for retrieving content classification model. While using a model to classify the data the output of image is a class label and a score matrix.

The predictor is selected due to its advantages. The advantage does not have the problem of Overfitting. The Overfitting occurs when model has multiple features related to the number of observations and results in poor predictive performance. This problem is relevant to consider while working with machine learning on images because number of features extracted from an image are large.

Measures of Forecast Accuracy:

Forecast accuracy is a method of deviation of prediction or forecast from the actual outcome. It is generally defined as

Error=Actual Demand-Forecast

Or

e=A-F

Forecast accuracy can be measured by using two methods

1. Mean Forecast Accuracy(MFE)
2. Mean Absolute Deviation (MAD)

Mean Forecast Accuracy (MFE):

Mean forecast error represents the deviation of forecast from actual demand. It is the mean of differences for every period in between number of period forecasts and actual demand for related periods. This error is mostly used as bias for following and adjusting forecasts. If it is positive then the forecasts are low compared to actual demand. And if it is negative then the forecasts will be high. Mean forecast deviation is defined as follows

సభ్య)

If the value of MFE is be greater than zero then model tends to under forecast and if the value of MFE is less than zero then model tends to over forecast.

Mean Absolute Deviation (MAD)

Mean absolute deviation deviates the forecasted demand from actual demand. It is the mean deviation for every period absolute terms and respective period demand.

MAD=∑ |□□□□ □□1□ |

Uses of Forecast Error:

1. It is used to compare alternative forecasting models
2. It is used to forecast the model bias.
3. It is used to generate absolute size of forecast errors.
4. It is used to identify the forecast models which needed to be adjusted.

UNIT-V

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations

. **Data visualization:** Data visualization is the graphical representation of information and **data**. By using visual elements like charts, graphs, and maps, **data visualization** tools provide an accessible way to see and understand trends, outliers, and patterns in **data**.

Uses of data visualization

➢Powerful way to explore data with presentable results.

➢Primary use is the preprocessing portion of the data mining process. ➢Supports in data cleaning process by finding incorrect and missing values.

General Data visualization Techniques:

➢Box plots

➢Histograms

➢Heat maps

➢Charts

➢Tree maps

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.

Data visualization Techniques:

Data visualization aims to communicate data clearly and effectively through graphical representation. Data visualization has been used extensively in many applications—for example, at work for reporting, managing business operations, and tracking progress of tasks.

More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data.

Data visualization Techniques:

1. Pixel Oriented Visualization technique

2. Geometric Projection Visualization technique
a. Scatter plot matrices
b. Hyper slice
c. Parallel coordinates
3. Icon based visualization techniques
4. Hierarchical Visualization techniques

Pixel-Oriented Visualization Techniques:

➢A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value. ➢For a data set of m dimensions, pixel-oriented techniques create m windows on the screen, one for each dimension.

➢The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows. The colors of the pixels reflect the corresponding values.

➢Inside a window, the data values are arranged in some global order shared by all windows.

➢The global order may be obtained by sorting all data records in a way that's meaningful for the task at hand.

Pixel-based visualizations use that approach and are capable of displaying large amounts of data on a single screen.

Case Study:

➢All Electronics maintains a customer information table, which consists of 4 dimensions: income, transaction_volume and age. ➢We analyse the correlation between income and other attributes by visualization.

➢We sort all customers in income in ascending order and use this order to layout the customer data in the 4 visualization windows as shown in fig.

Geometric Projection visualization techniques:

A drawback of **pixel-oriented visualization** techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.

Geometric projection techniques help users find interesting projections of multidimensional data sets.

Geometric projection techniques are a good choice for finding outliers and correlation between attributes in multivariate data. A geometric projection technique does this by using transformations and projections of the data. When using large data sets a clustering algorithm is usually necessary to apply before the visualization technique to avoid cluttered and unclear data caused by the too much information. Some widely used geometric projection techniques are:

Scatter plots:

A scatter plot is one of the most common visualization techniques and can be visualized both in 3D and 2D. The scatter plot visualizes different attributes of the data on the x,y axis for 2D visualizations and also along the z-axis in 3D. Scatter plots are usable to find correlations between attributes in arbitrary small data sets. If the data set gets too big or contains too many attributes the scatter plot gets cluttered and hard to interpret.

HyperSlice:

HyperSlice is a new method for the visualization of scalar functions of many variables. With this method the multi-dimensional function is presented in a simple and easy to understand way in which all dimensions are treated identically. The central concept is the representation of a multi-dimensional function as a matrix of orthogonal two-dimensional slices. These two dimensional slices lend themselves very well to interaction via direct manipulation, due to a one to one relation between screen space and variable space.

Parallel coordinates:

➢To visualize n-dimensional data points, the parallel coordinates technique draws n equally spaced axes, one for each dimension, parallel to one of the display axes.

➢A data record is represented by a polygonal line that intersects each axis at the point corresponding to the associated dimension value. ➢A major limitation of the parallel

coordinate's technique is that it cannot effectively show a data set of many records.

Icon-based visualization techniques:

Icon-based techniques visualize data by changing the properties of an icon or glyph according to the data. An early version was **Chernoff faces** where data is mapped to different face parts as nose, mouth, eyes and more. For example how rich people are can be mapped to the mouth of the Chernoff face. Rich people represented by a happy mouth and and poor people by a sad mouth. Other methods are:

Stick figures:

It maps multidimensional data to five –piece stick figure, where each figure has 4 limbs and a body.

Two dimensions are mapped to the display (x and y) axes and the remaining dimensions are mapped to the angle and/or length of the limbs.

Hierarchical Visualization Techniques:

Hierarchical visualization techniques are techniques, whose domain data structure and type of information are, respectively, tree and hierarchical information. There are two basic branches of **visualization techniques** for **hierarchies**. The first is based on a node-edge graph-layout approach which focuses attention on the structure and relationships, and the second on space-filling approaches, which focus attention on the relative sizes of nodes in the **hierarchy**.

➢The visualization techniques discussed so far focus on visualizing multiple dimensions simultaneously.

➢However, for a large data set of high dimensionality, it would be difficult to visualize all dimensions at the same time.

➢Hierarchical visualization techniques partition all dimensions into subsets (i.e., subspaces).

The subspaces are visualized in a hierarchical manner.

➢"Worlds-within-Worlds," also known as n-Vision, is a representative hierarchical visualization method.

➢Suppose we want to visualize a 6-D data set, where the dimensions are F,X1, : : : ,X5.

Given more dimensions, more levels of worlds can be used, which is why the method is called "worlds-within-worlds."

Worlds - Within – Worlds

As another example of hierarchical visualization methods, tree-maps display hierarchical data as a set of nested rectangles. For example, a tree-map visualizing Google news stories.

Tree Map

All news stories are organized into seven categories, each shown in a large rectangle of a unique color. Within each category (i.e., each rectangle at the top level), the news stories are further partitioned into smaller subcategories.

Data visualization choices:

Five factors that influence data visualization choices:

Audience: It's important to adjust data representation to the specific target audience.

Content: The type of data you are dealing with will determine the tactics. **Context:** You can use different data visualization approaches and read data depending on the context.

Dynamics: There are various types of data, and each type has a different rate of change.

Purpose: The goal of data visualization affects the way it is implemented. In order to make a complex analysis, visualizations are compiled into dynamic and controllable dashboards that work as visual data analysis techniques and tools.

Tools for Data visualization:

Data visualization tools for different types of users and purposes.

Tableau is one of the leaders in this field. A user-friendly interface and a rich library of interactive visualizations, Tableau stands out for its powerful capabilities. The platform provides large integration options including My SQL, Teradata, Hadoop and Amazon Web

Services. This platform to derive meaning from data and use insights for effective storytelling.

R and **Python** are well-equipped for **data visualization**. Customizing graphics is easier and more intuitive in **R** with the help of ggplot2 than in Python with Matplotlib. The Seaborn library helps to overcome this, and offers **good** standard solutions which get by with relatively few lines of code.

Plotly is one of the most popular platforms in this category. It's more complex than Tableau, however, comes with analytics perks. With this visualization tool, you can create charts using R or Python, build custom data analytics

IBM Watson Analytics is known for its NLP capabilities. The platform literally supports conversational data control a longside strong dashboard building and data reporting tools.

Tools for complex data visualization:

The growing adoption of connected technology places a lot of opportunities before the companies and organizations. To deal with large volumes ofmulti source often unstructured data, businesses search for more complex visualization and analytics solutions. This category includes **Power BI, Kibana and Grafana.**

Power BI is exceptional for its highly intuitive drag-and-drop interface, short learning curve and large integration capabilities, including Salesforce and MailChimp.

Kibana is the part of the Elastic Stack that turns data into visual insights. It's built on and designed to work on Elasticsearch data only. This exclusivity, however, does not prevent it from being one of the best data visualization tools for log data.

Grafana a professional data visualization and analytic tool that supports up to 30 data sources, including AWS, Elastic search and Prometheus. Grafana is more flexible in terms of integrations compared to Kibana, each of the systems works best with its own type of data.

Data Visualization Process:

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make **data** easier for the human brain to understand and pull insights from. The main goal of **data visualization** is to make it easier to identify patterns, trends and

outliers in large **data** sets.

Fig: Data visualization Process