Chapter-1

INTRODUCTION

Data: Is a collection of discrete **values** that convey **information,** describing **quantity, quality, fact, statistics,** other basic units of meaning, or simply sequences of **symbols** that may be further **interpreted.**

Data base: A database is an organized collection of data stored and accessed electronically. Small databases can be stored on a file system, while large databases are hosted on computer clusters or cloud storage.

A **database management system** (**DBMS**) is the **software** that interacts with **end users**, applications, and the database itself to capture and analyze the data.

Datum: A **datum** is an individual value in a collection of data.

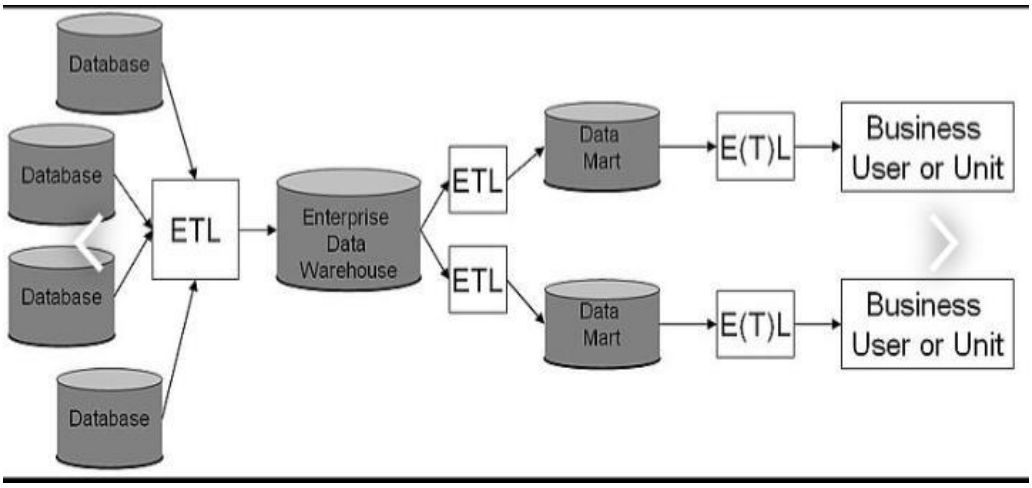
Data are usually organized into structures such as tables that provide additional context and meaning, and which may themselves be used as data in larger structures. Data may be used as variables in a computational process.

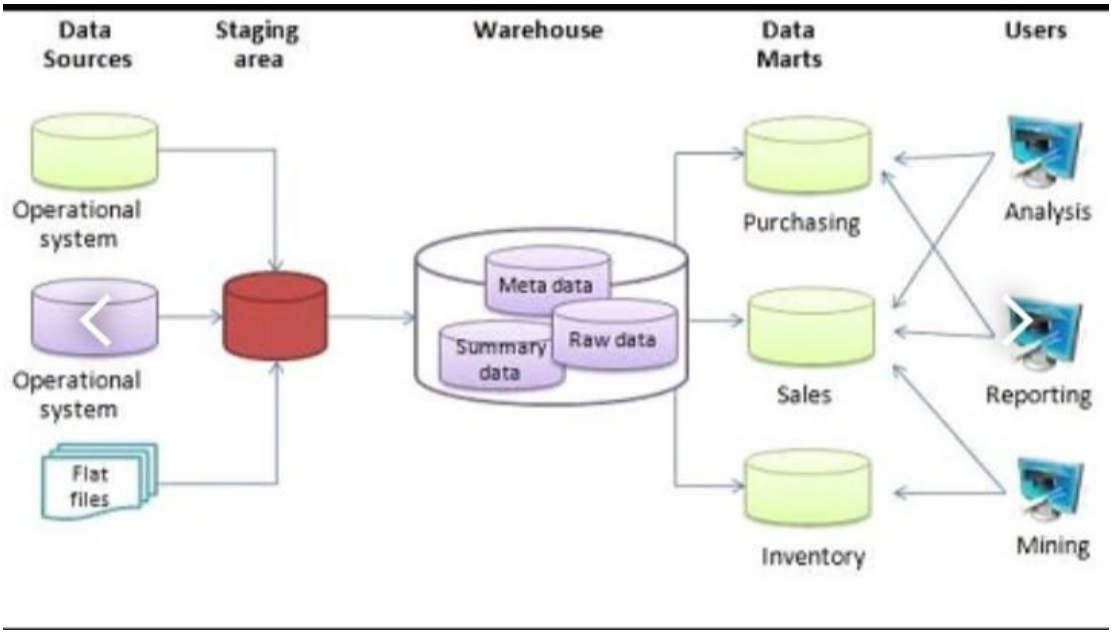
Data set: A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files.

Data warehouse: A data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is considered a core component of business intelligence.

[1] DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place

[2] that are used for creating analytical reports for workers throughout the enterprise.





The data stored in the warehouse is uploaded from the operational systems (such as marketing or sales). The data may pass through an **operational data store** and may require **data cleansing** for additional operations to ensure **data quality** before it is used in the DW for reporting.

Data are commonly used in scientific research, finance, and in virtually every other form of human organizational activity. Examples of data sets include stock prices, crime rates, unemployment rates, literacy rates, and census data.

1.1 Applications and Advantages of data analysis

1. Transportation

Data analytics can be applied to help in improving Transportation Systems and the intelligence around them. The predictive method of the analysis helps find transport problems like Traffic or network congestion. It helps synchronize the vast amount of data and uses them to build and design plans and strategies to plan alternative routes and reduce congestion and traffic, which in turn reduces the number of accidents and misshapenness. Data Analytics can also help to optimize the buyer's experience in the travels by recording the information from social media. It also helps travel companies fix their packages and boost the personalized travel experience as per the data collected.

For **Example** During the Wedding season or the Holiday season, the transport facilities are prepared to accommodate the heavy number of passengers travelling from one place to another using prediction tools and techniques.

2. Logistics and Delivery

There are different logistic companies like DHL, FedEx, etc that use data analytics to manage their overall operations. Using the applications of data analytics, they can figure out the best shipping routes, and approximate delivery times, and also can track the real-time status of goods that are dispatched using GPS trackers. Data Analytics has made online shopping easier and more demandable.

Example of Use of data analytics in Logistics and Delivery:

When a shipment is dispatched from its origin, till it reaches its buyers, every position is tracked which leads to the minimizing of the loss of the goods.

3. Web Search or Internet Web Results

The web search engines like Yahoo, Bing, Duckduckgo, and Google use a set of data to give you when you search a data. Whenever you hit on the search button, the search engines use algorithms of data analytics to deliver the best-searched results within a limited time frame. The set of data that appears whenever we search for any information is obtained through data analytics.

The searched data is considered as a keyword and all the related pieces of information are presented in a sorted manner that one can easily understand.

For **Example**, when you search for a product on **amazon** it keeps showing on your social media profiles or to provide you with the details of the product to convince you by that

product. Even **Google AdSense** collect user interest from browsing history and displays ads which are relevant to it.

4. Manufacturing

Data analytics helps the manufacturing industries maintain their overall work through certain tools like prediction analysis, regression analysis, budgeting, etc. The unit can figure out the number of products needed to be manufactured according to the data collected and analyzed from the demand samples and likewise in many other operations increasing the operating capacity as well as the profitability.

5. Security

Data analyst provides utmost security to the organization, Security Analytics is a way to deal with online protection zeroed in on the examination of information to deliver proactive safety efforts. No business can foresee the future, particularly where security dangers are concerned, yet by sending security investigation apparatuses that can dissect security occasions it is conceivable to identify danger before it gets an opportunity to affect your framework and main concern.

6. Education

Data analytics applications in education are the most needed data analyst in the current scenario. It is mostly used in adaptive learning, new innovations, adaptive content, etc. Is the estimation, assortment, investigation, and detailing of information about students and their specific circumstances, for reasons for comprehension and streamlining learning and conditions in which it happens.

7. Healthcare

Applications of data analytics in healthcare can be utilized to channel enormous measures of information in seconds to discover treatment choices or answers for various illnesses. This won’t just give precise arrangements dependent on recorded data yet may likewise give accurate answers for exceptional worries for specific patients.

8. Military

Military applications of data analytics bring together an assortment of specialized and application-situated use cases. It empowers chiefs and technologists to make associations between information investigation and such fields as augmented reality and psychological science that are driving military associations around the globe forward.

9. Insurance

There is a lot of data analysis taking place during the insurance process. Several data, such as actuarial data and claims data, help insurance companies realize the risk involved in insuring the person. Analytical software can be used to identify risky claims and bring them before the authorities for further investigation.

10. Digital Advertisement

Digital advertising has also been transformed as a result of the application of data science. Data analytics and data algorithms are used in a wide range of advertising mediums, including digital billboards in cities and banners on websites.

11. Fraud and Risk Detection

Detecting fraud may have been the first application of data analytics. They applied data analytics because they already had a large amount of customer data at their disposal. Data analysis was used to examine recent spending patterns and customer profiles to determine the likelihood of default. It eventually resulted in a reduction in fraud and risk.

12. Travel

Data analysis applications can be used to improve the traveller's purchasing experience by analyzing social media and mobile/weblog data. Companies can use data on recent browse-to-buy conversion rates to create customized offers and packages that take into account the preferences and desires of their customers.

13. Communication, Media, and Entertainment

When it comes to creating content for different target audiences, recommending content, and measuring content performance, organizations in this industry analyze customer data and behavioral data simultaneously. Data analytics is applied to collect and utilize customer insights and understand their pattern of social-media usage.

14. Energy and Utility

Many firms involved in energy management use data analysis applications in areas such as smart-grid management, energy distribution, energy optimization, and automation building for other utility-based firms.

1.2 INTRODUCTION TO "WEKA" TOOL

The Weka machine learning workbench is a modern platform for applied machine learning. Weka is an acronym which stands for Waikato Environment for Knowledge Analysis. It is also the **Name of a New Zealand bird the Weka.** Five of the top features for using Weka are:

***Open Source:** Weka is released as open source software under the GNU GPL. It is Dual licensed and Pentaho Corporation owns the exclusive license to use the platform for Business intelligence in their own product.

***Graphical Interface:** It has a Graphical User Interface (GUI). This allows you to Complete your machine learning projects without programming.

***Command Line Interface:** All features of the software can be used from the command Line. This can be very useful if you are looking to take your usage of the platform to the Next level and start automating common tasks.

* **Java API:** It is written in Java and provides a API that is well documented and promotes Integration into your own applications. Note that the GNU GPL means that in turn your Software would also have to be released as GPL.

***Community:** The interface and algorithms are well documented and the large community Using the platform means that there is plenty of support if you need it.

1.2.1 Introduction to the Weka Graphical Interface

The Weka workbench provides two main ways to work on your problem: The Explorer for playing Around and trying things out and the Experiment Environment for controlled experiments.

1.2.1.1 Weka Explorer

The explorer is where you play around with your data and think about what transforms to Apply to your data, what algorithms you want to run as experiments. The Explorer interface is

Divided into 5 different tabs:

***Preprocess:** Load a dataset and manipulate the data into a form that you want to work With.

***Classify:** Select and run classification and regression algorithms to operate on your data.

* **Cluster:** Select and run clustering algorithms on your dataset.

***Associate:** Run association algorithms to extract insights from your data.

***Select Attributes:** Run attribute selection algorithms on your data to select those attributes that are relevant to the feature you want to predict.

* **Visualize:** Visualize the relationship between attributes.

1.2.1.2 Weka Experiment Environment

This interface is for designing experiments with your selection of algorithms and datasets ,Running experiments and analyzing the results. The tools for analyzing results are powerful, Allowing you to consider and compare results that are statistically significant over multiple runs.

The Experiment interface is divided into 3 different tabs:

***Setup for designing your controlled experiments.

* Run for executing your experiments.

*Analyze for analyzing the results from your experiments.

1.3 Why You Should Use Weka

The main reason Weka is so good for beginners is that you can work through the process Of applied machine learning using the graphical interface without having to write a single Line of code. This is a big deal because getting a handle on the process, handling data and Experimenting with algorithms is what a beginner should be learning about, not learning yet Another scripting language.

Here are some additional reasons:

*It provides a simple graphical user interface that encapsulates the process of applied Machine learning.

*It facilitates algorithm and dataset exploration as well as rigorous experiment design and Analysis.

*It is free and open source, meaning you can download it and start using it now.

*It is cross-platform and runs on Windows, Mac OS X and Linux, meaning it will run on Whatever environment you're using.

* It contains state-of-the-art algorithms with an impressive list of Decision Trees, Rule-Based Algorithms and Ensemble methods, as well as others.

1.4 INTRODUCTION TO "TABLEAU" TOOL

Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and Distribute an interactive and shareable dashboard, which depict the trends, variations, and Density of the data in the form of graphs and charts. Tableau can connect to files, relational and Big Data sources to acquire and process data. As a leading data visualization tool, Tableau has Many desirable and unique features. Its powerful data discovery and exploration application Allows you to answer important questions in seconds. You can use Tableau's drag and drop Interface to visualize any data, explore different views, and even combine multiple databases Easily. It does not require any complex scripting. Anyone who understands the business Problems can address it with a visualization of the relevant data. After analysis, sharing with Others is as easy as publishing to Tableau Server.

TABLEAU FEATURES

Tableau provides solutions for all kinds of industries, departments, and data environments.

Following are some **unique features which enable Tableau to handle diverse scenarios.**

*__Speed of Analysis:__ As it does not require high level of programming expertise, any user with access to data can start using it to derive value from the data.

*__Self-Reliant:__ Tableau does not need a complex software setup. The desktop version which is used by most users is easily installed and contains all the features needed to start and complete data analysis.

*__Visual Discovery:__ The user explores and analyses the data by using visual tools like colors, trend lines, charts, and graphs. There is very little script to be written as nearly everything is done by drag and drop.

*__Blend Diverse Data Sets:__ Tableau allows you to blend different relational, semi structured and raw data sources in real time, without expensive up-front integration costs. The users don't need to know the details of how data is stored.

*__Architecture Agnostic:__ Tableau works in all kinds of devices where data flows. Hence, the user need not worry about specific hardware or software requirements to use Tableau.

* __Real-Time Collaboration:__ Tableau can filter, sort, and discuss data on the fly and embed a live dashboard in portals like SharePoint site or Salesforce. You can save your view of data and allow colleagues to subscribe to your interactive dashboards so they see the very latest data just by refreshing their web browser.

* __Centralized Data:__ Tableau server provides a centralized location to manage all of the organization's published data sources. You can delete, change permissions, add tags, and manage schedules in one convenient location. It's easy to schedule extract refreshes and manage them in the data server. Administrators can centrally define a schedule for extracts on the server for both incremental and full refreshes.

There are three basic steps involved in creating any Tableau data analysis report.

These three steps are :

* **Connect to a data source :**It involves locating the data and using an appropriate type of connection to read the data.

* **Choose dimensions and measures :**This involves selecting the required columns from the source data for analysis.

***Apply visualization technique :**This involves applying required visualization methods, such as a specific chart or graph type to the data being analyzed.

Chapter-2

DATA ANALYSIS AND ANALYTICS

2.1 DATA MINING AND DATA CLEANING

Data Analysis: Data analysis is a process of **inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.**

In today's business world, data analysis plays a role in making decisions more **scientific** and helping businesses operate more **effectively**.

Data mining: Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of **machine learning, statistics, and database systems**. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of **extracting information (with intelligent methods) from a data set** and transforming the information into a comprehensible structure for further use.

Data mining is the analysis step of the "knowledge discovery in databases" **process, or KDD.** Aside from the raw analysis step, it also involves :

1)database and data management aspects

2)data pre-processing

3)model and inference considerations

4)interestingness metrics

5)complexity considerations

6)post-processing of discovered structures

7)pre-processing

8)online updating

In statistical applications, data analysis can be divided into **descriptive statistics**, **exploratory data analysis** (EDA), and **confirmatory data analysis** (CDA). EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing **hypotheses**. **Predictive analytics** focuses on the application of statistical models for predictive forecasting or classification, while **text analytics** applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of **unstructured data**. All of the above are varieties of data analysis.

Data integration: It is a precursor to data analysis, and data analysis is closely linked to **data visualization** and data dissemination.

Data analytics: As the process of analyzing raw data to find trends and answer questions, the definition of data analytics captures its broad scope of the field. However, it includes many techniques with many different goals.

2.2 DATA CLEANING

10 Super Neat Ways to Clean Data in Excel Spreadsheets

1) Get Rid of Extra Spaces
2) Select and Treat All Blank Cells
3) Convert Numbers Stored as Text into Numbers
4) Remove Duplicates
5) Highlight Errors
6) Use Find and Replace to Clean Data in Excel
7) Change Text to Lower/Upper/Proper Case
8) Parse Data Using Text to Column
9) Spell Check
10) Delete all Formatting

Chapter-3

REPORT OF FIFA WORLD CUP DATA SET

The **FIFA World Cup**, often simply called the **World Cup**, is an **international association football** competition contested by the senior **men's national teams** of the members of the *Fédération Internationale de Football Association* (**FIFA**), the sport's global governing body. The championship has been awarded **every four years** since the **inaugural tournament in 1930**, except in 1942 and 1946 when it was not held because of the **Second World War**. The current champions are **France**, who won their second title at the **2018 tournament** in Russia.

The data set contains:

1) Instances:23922
2) Attributes:27

1)Day

2) Month

3)Year

4) home team

5)away team

6)home team continent

7)away team continent

8)home team fifa rank

9)away team fifa rank

10)home team total fifa points

11)away team total fifa points

12)home team score

13)away team score

14)tournament

15)city

16)country

17)neutral location

18)shoot-out

19)home team result

20)home team goalkeeper score

21)away team goalkeeper score

22)home team mean defense score

23)home team mean offense score

24)home team mean midfield score

25)away team mean defense score

26)away team mean offense score

27)away team mean midfield score

Data cleaning

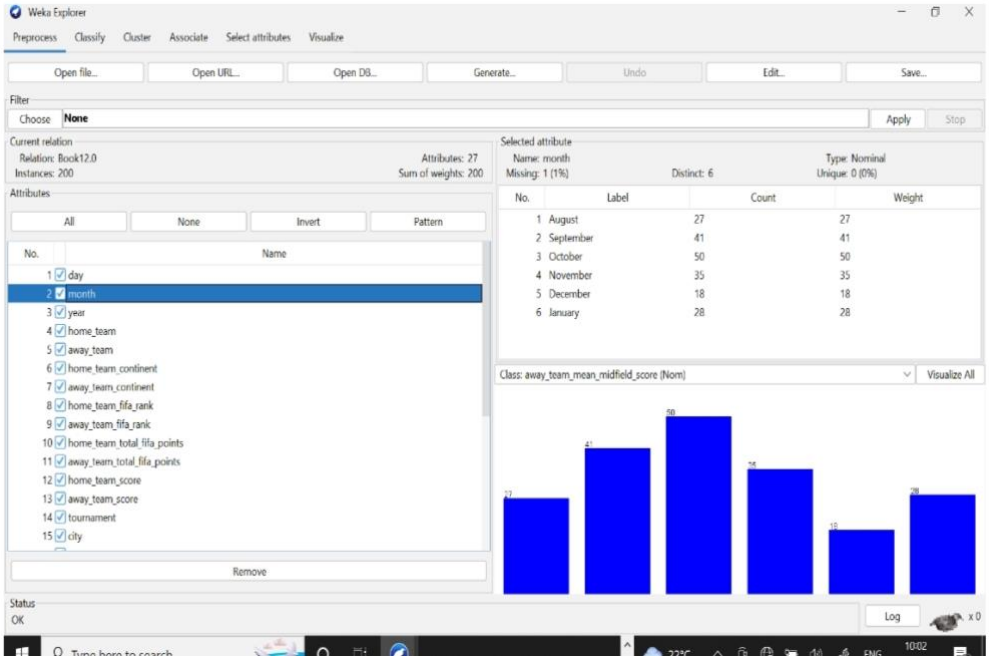
In order to **Boost results and revenue , Save time and increase productivity(qualitative analysis) , Minimize compliance risks**
In order to get qualitative and accurate visualization.

Refer the link for detail data Cleaning process:

https://youtu.be/e0TfIbZXPeA

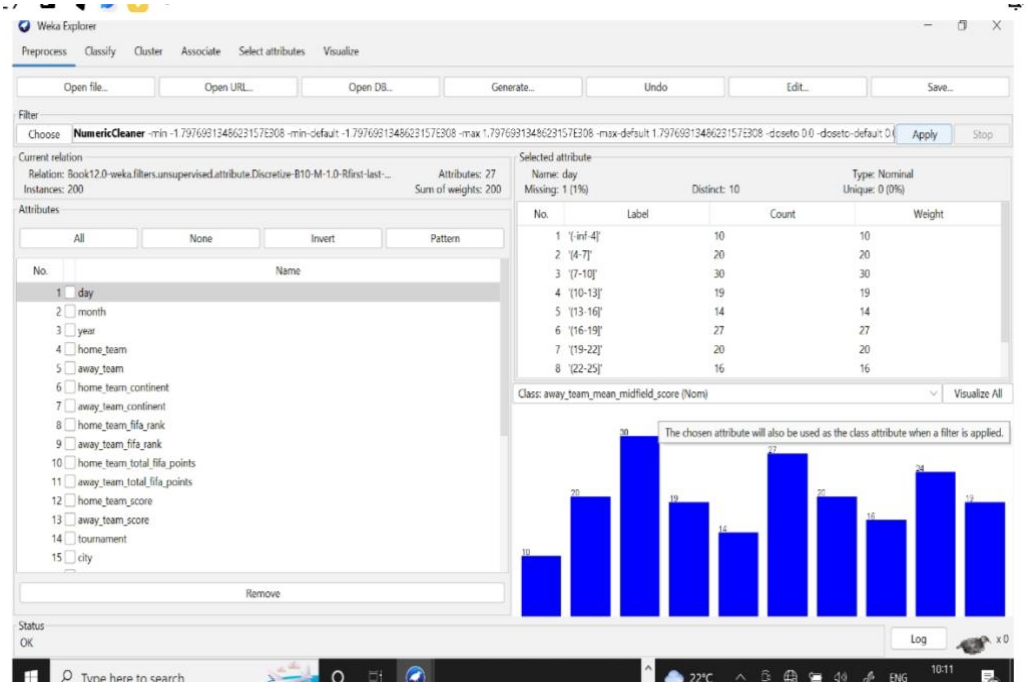
Preprocess: Load a dataset and manipulate the data into a form that we want to work With.

Classify: Select and run classification and regression algorithms to operate on your data.

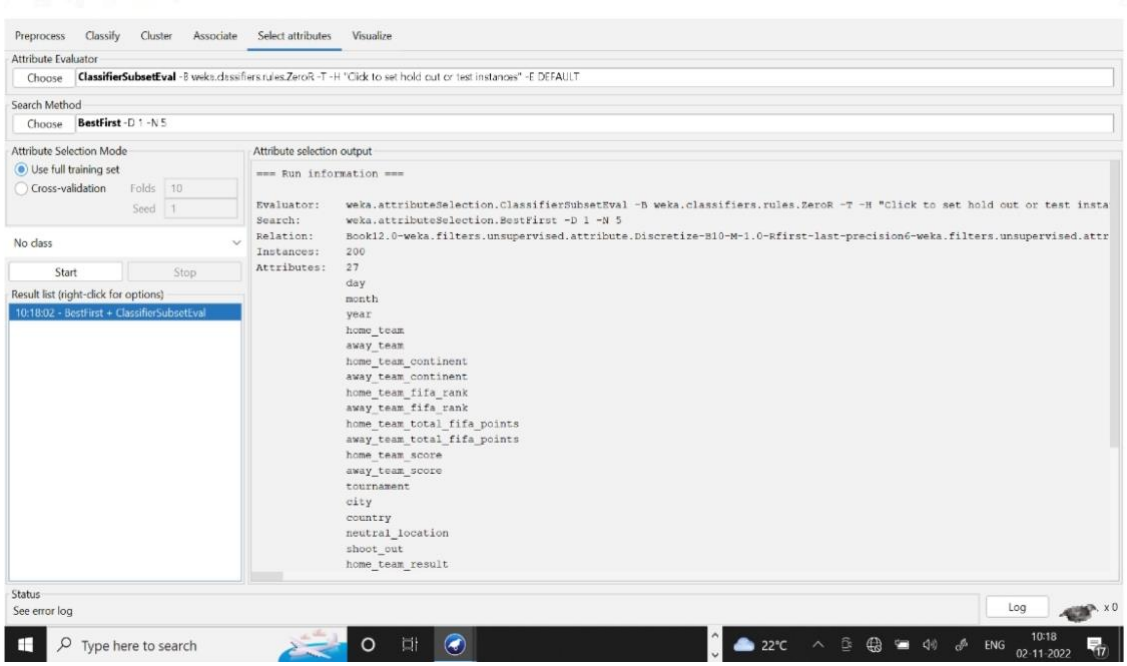


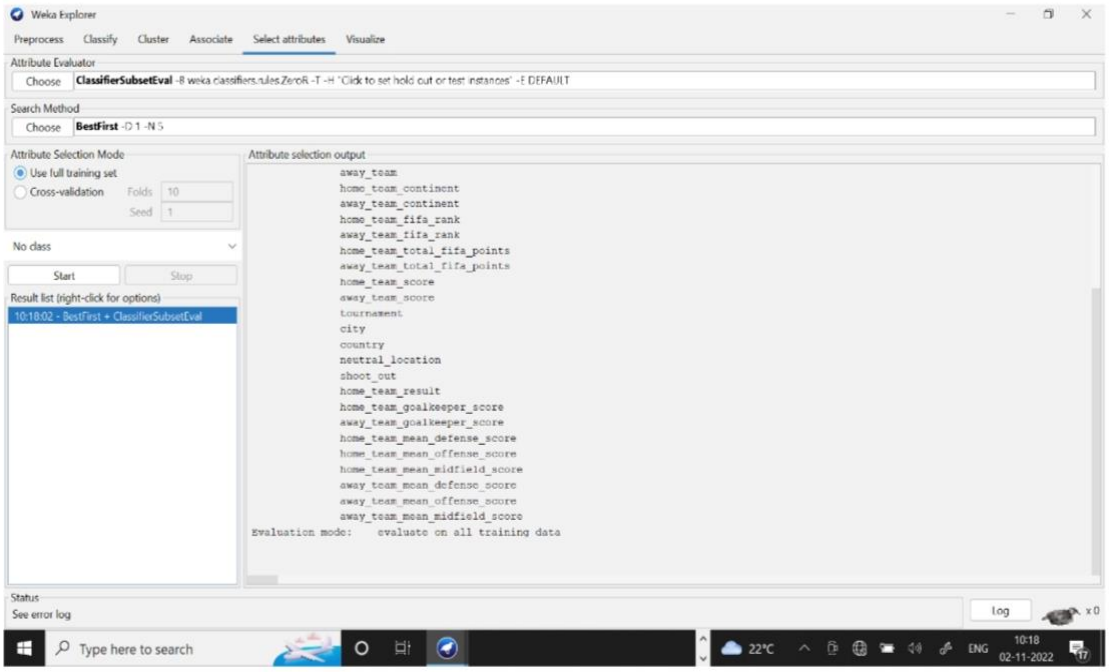
Cluster: Select and run clustering algorithms on your dataset.

Associate: Run association algorithms to extract insights from your data.



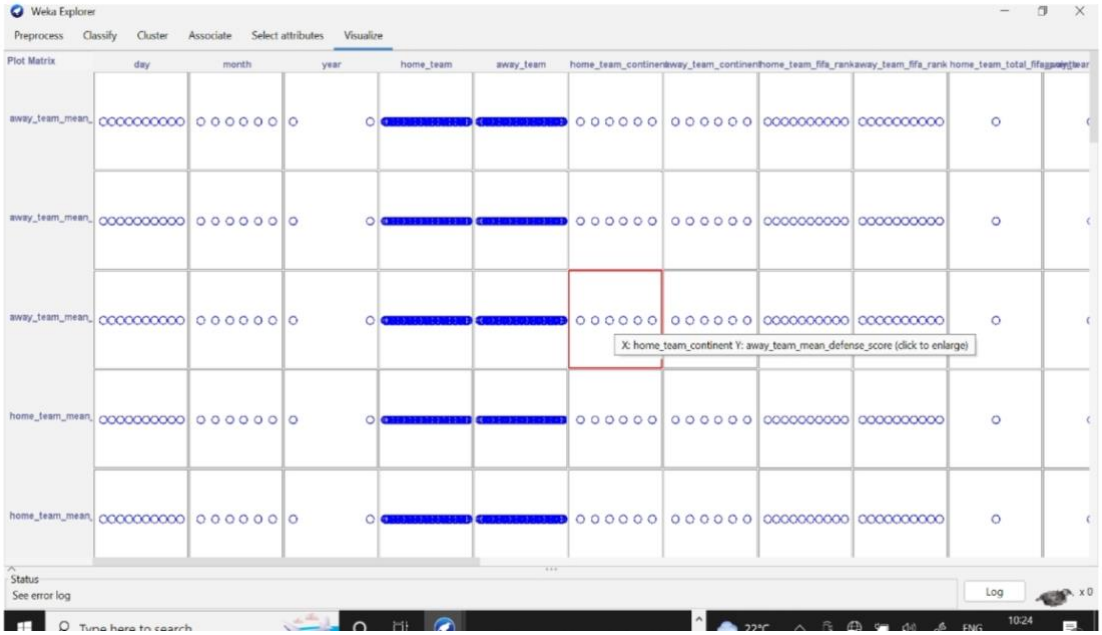
Select Attributes: Run attribute selection algorithms on your data to select those attributes that are relevant to the feature you want to predict.

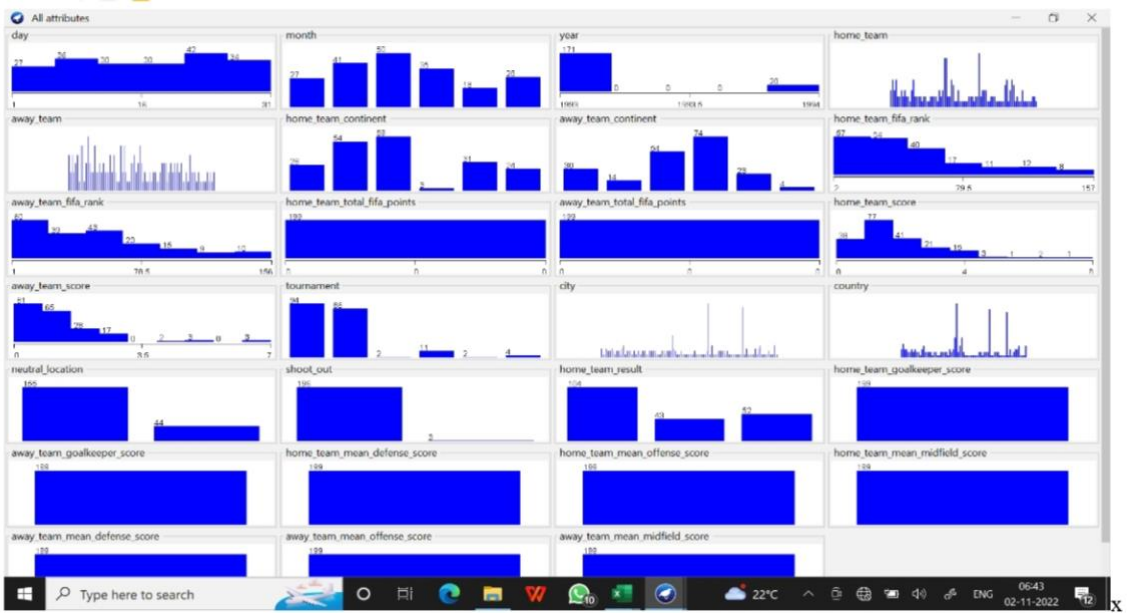




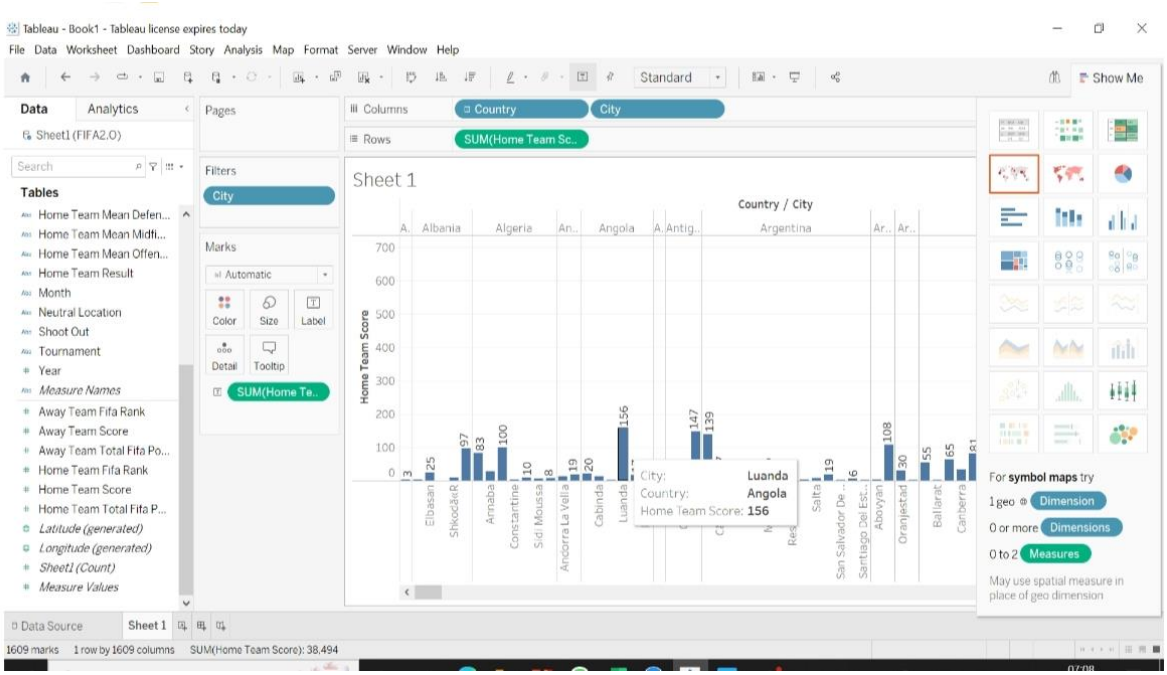
3.1 VISUALIZATION IN WEKA

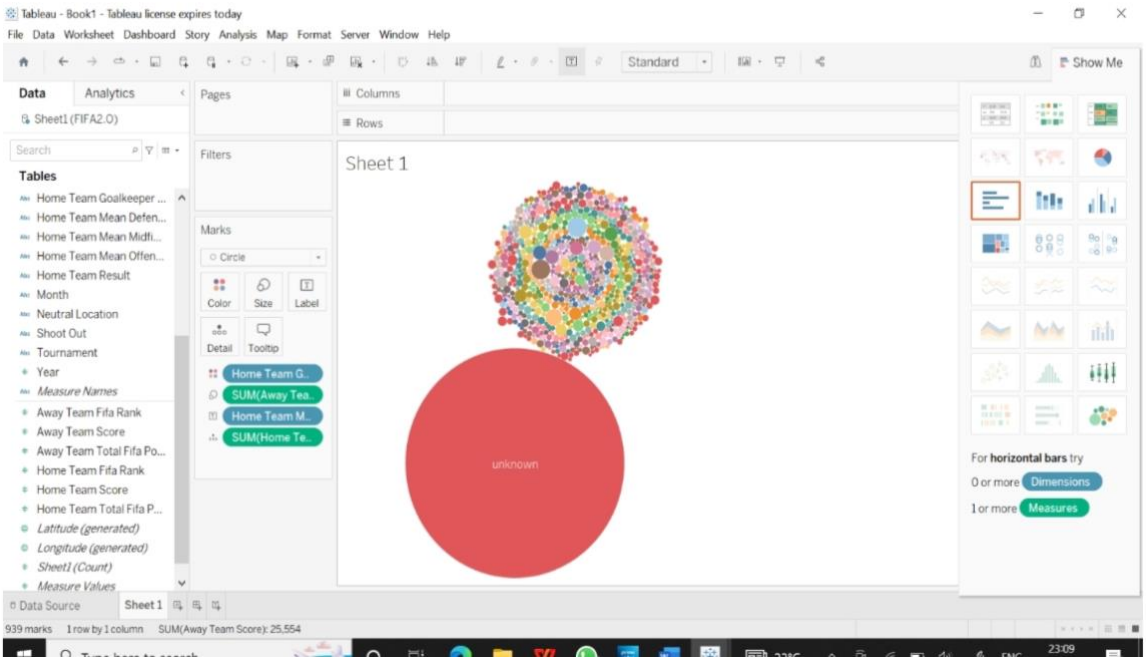
Visualize: Visualize the relationship between attributes.





3.2 VISUALIZATION IN TABLEAU





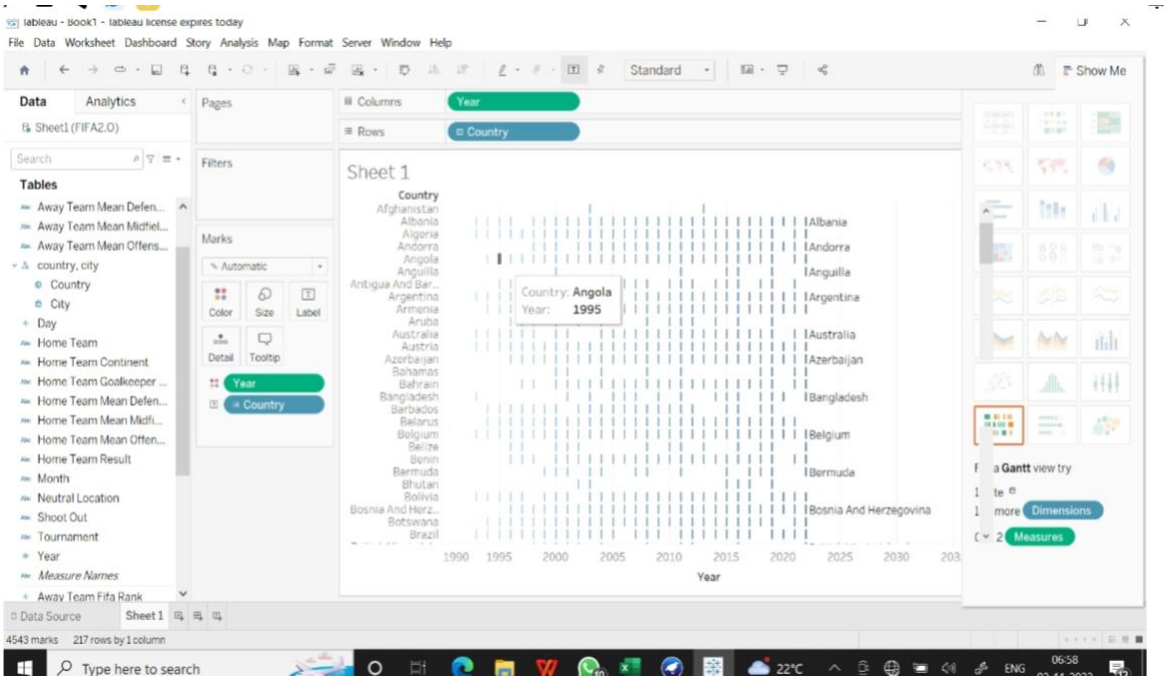


Tableau - Book1 - Tableau license expires today
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Data Analytics
Sheet1 (FIFA2.0)
Tables
Pages
Filters
Marks
Columns Year
Rows Country
Sheet 1
Country: Angola
Year: 1995
Show Me
Data Source Sheet 1
4543 marks 217 rows by 1 column

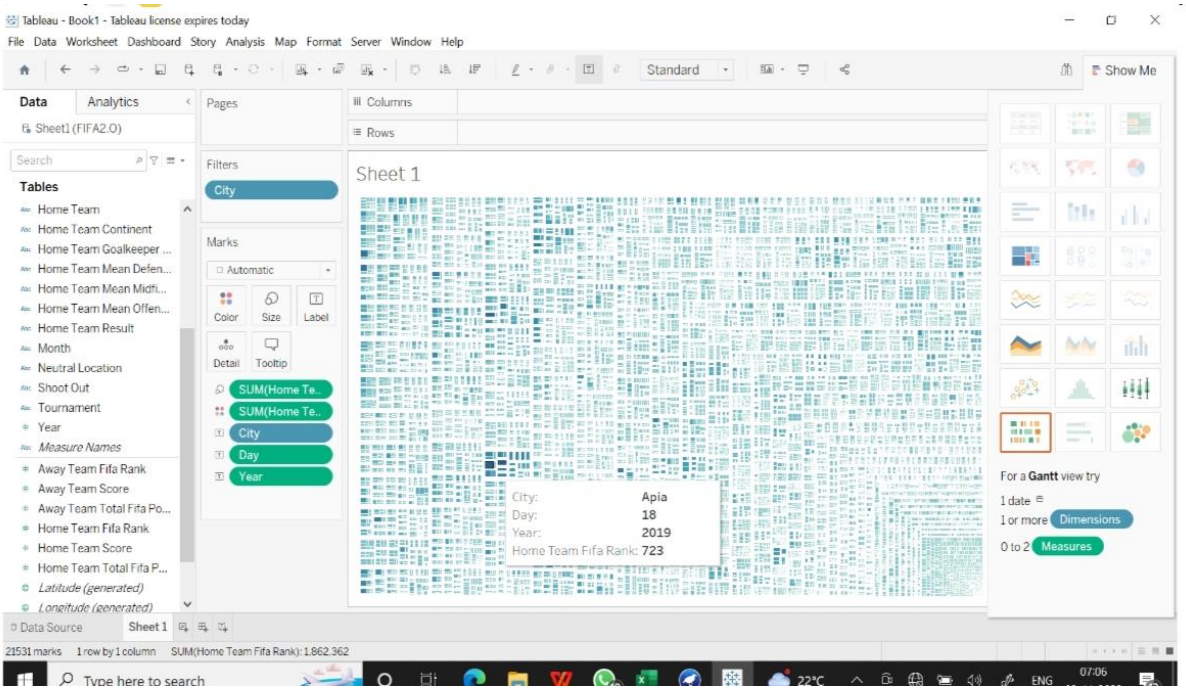


Tableau - Book1 - Tableau license expires today
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Data Analytics
Sheet1 (FIFA2.0)
Tables
Pages
Filters
City
Marks
Columns
Rows
Sheet 1
City: Apia
Day: 18
Year: 2019
Home Team Fifa Rank: 723
Show Me
Data Source Sheet 1
21531 marks 1 row by 1 column SUM(Home Team Fifa Rank): 1,862,362

CONCLUSION

Finally we would like to conclude that through this data analysis , we can analyze team strategy, team that has won more number of matches , highest score count , team strength and even we can predict chances of winning in next tournament(according to our analysis France /Argentina/Croatia having high chances of winning upcoming tournament .) etc. Similarly we can analyze business in profits, loss, reason for loss , improvement in sales, we can also analyze recruitment for clients/customers etc. Data analysis plays a significant role in each and every sector. This work aims to serve as motivation for further research concerning Data analysis and analytics for the growth of career.

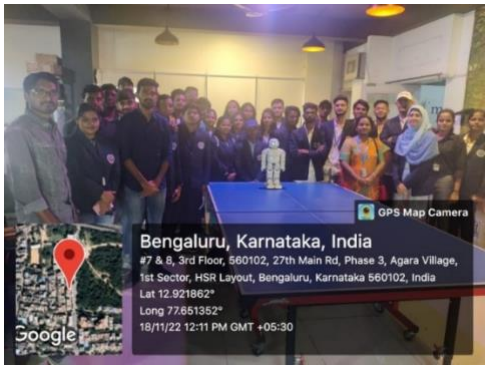
Chapter-4

REPORT ON INDUSTRIAL VISIT

Mr. Hariharan Bojan (Founder & CEO) is an Entrepreneur, a true engineer who wants to steer the education industry by enabling technology accessible at the ripe age of tenderness! A true leader whose definition of success is measured not by leading but by strengthening everyone to become so! A cool boss who never bosses or tells how, driven by compassion and not just aggression! A simple being whose vision is to make the world slightly better with goodness and good will! He has founded/co-founded/setup several successful ventures/teams since 2003.

Sirena Technologies **Founded in** 2013 Provider of robots for education, entertainment, and smart homes. Products include humanoid robots, toy robots, educational robots, and wireless speakers. Its homegrown robotic Platform brings a unique value proposition for the technical colleges providing them an opportunity to experiment with it. Plans to introduce humanoid robots as teaching assistants in the education space in India.

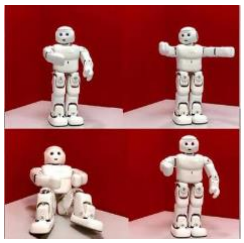
Sirena Technologies was started with an idea of making an advanced humanoid in India. A handful of engineers came up with a prototype of a humanoid within 2 months. Exposure to latest technologies in Wi-Fi audio streaming, artificial intelligence, manufacturing processes has enhanced our products to globally competitive standards.



The company has built the first Indian Humanoid Robot **‘Nino’, Nino-T** a life size humanoid and introduces SKIP (Sirena Knowledge and Information Program) for schools which is crafted to enable students to learn cutting-edge trending technologies which include Humanoids, Artificial Intelligence, Internet of Things, Voice Recognition, Computer Vision, Mechatronics, 3D printing, Android programming and more.

We interrogate with and it was a wonderful experience ,we learnt a lot about sensors , which are programming language they use , what are the challenges they face while designing and programming , what are their future robotic plans , we have a demo of **speaker** they built and “**nino**”. We have a great time with nino .

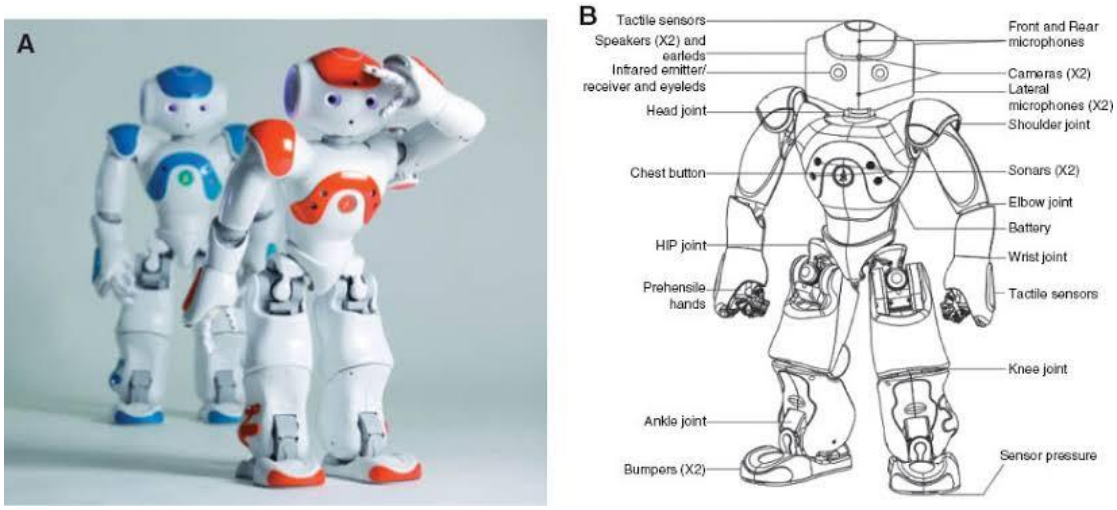








NINO – Humanoid Lad



Inspired from an idea to be with every kid out there, NINO which means a Boy in Spanish is being an educational companion to kids in schools providing them a platform to learn and experiment with a humanoid robot.

Inspired from an idea to be with every kid out there, NINO which means a Boy in Spanish is being an educational companion to kids in schools providing them a platform to learn and experiment with a humanoid robot.

Nino is a humanoid that can talk, walk, dance, sing, play and work wonders with its inherently built intelligence. For each concept that a teacher covers in her classroom, Nino has a set of innovative, developmentally appropriate, hands-on, experiential and interactive activities, stories, poems and projects that can be used to introduce, strengthen, summarize and assess concepts. On one hand, Nino acts as an assistant-teacher and on the other, it works with each child as a buddy.

Nino is also a part of SKIP, Sirena Knowledge and information Program that primarily focuses upon teaching all you kids about technology by implementing robotics lab setup in your schools acquired with the most creative robotics projects to bring curiosity for further learning.

With Neural Network, Artificial Intelligence, ASR, text to speech solution… Nino is evolving everyday

REFERENCE

[1] WEKA material by Dr.S.Vagdevi(Prof. & Head, Dept. of AI&ML).

[2] Machine Learning Mastery With Weka (Analyze Data, Develop Models and Work Through Projects) ,Jason Brownlee.

[3] TABLEAU FOR DATA VISUALISATION material by Vindhya R (Prof., Dept. of AI&ML)

[4] Wikipedia(https://www.wikipedia.org/)

[5]Data cleaning(https://trumpexcel.com/clean-data-in-excel/)

[6] WEKA manual (https://www.google.com/url?sa=t&source=web&rct=j&url=https://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf&ved=2ahUKEwjn5tWT1Iz7AhUOSWwGHZbgDhoQFnoECBQQAQ&usg=AOvVaw17iUjTzz0WNZNWN3EBNAVo)

[7] Tableau manual (https://www.google.com/url?sa=t&source=web&rct=j&url=https://community.tableau.com/s/question/0D54T00000C5qIhSAJ/tableau-desktop-manual-download&ved=2ahUKEwji763p1Iz7AhVcSGwGHcJBDRQQFnoECBsQAQ&usg=AOvVaw0LWVGT9SBRKknn-6A7J3ux)

[8] Refer the link for detail data Cleaning process:

https://youtu.be/e0TfIbZXPeA

Query

a) Which tool is best weka or Tableau?
*Both are similar but each one has it's own specialty like Tableau: It's just drag and drop , It requires less skill compared to weka , easy to learn , it can Visualize data in verity types of graphical representations.
WEKA :It requires some skill , we can analyze data by knowing percentage of efficiency etc.

b) What is role of visualization in data analysis?
*Visualization plays a vital role in data analysis , it make easy understand and analysis. The colorful verities of graphical representations attracts customers / clients / data analyzer and make them understand and analyze easily.

c) Why do we need to clean data?
*Data cleaning helps us to rearrange data in proper format , remove special characters , remove duplicates/multiple entries , filling blank space etc. Which helps us to analyze the with more accurate and increases its efficiency.

Check list of items for final mini project report (with Yes or No marked, as applicable)

a) Is the Cover page in proper format? Y / N

b) Is the Title page in proper format? Y / N

c) Is the Certificate from the Supervisor in proper format? Has it been signed? Y / N

d) Is Abstract included in the Report? Is it properly written? Y / N

e) Does the Table of Contents' page include chapter page numbers? Y / N

f) Is Introduction included in the report? Is it properly written? Y / N

g) Are the Pages numbered properly? Y / N

h) Are the Figures numbered properly? Y / N

i) Are the Tables numbered properly? Y / N

j) Are the Captions for the Figures and Tables proper? Y / N

k) Are the Appendices numbered? Y / N

l) Does the Report have Conclusions/ Recommendations of the work? Y / N

m) Are References/ Bibliography given in the Report? Y / N

n) Have the References been cited in the Report? Y / N

o) Is the citation of References/ Bibliography in proper format? Y / N