SRI VENKATESWARA INTERNSHIP PROGRAM FOR RESEARCH IN ACADEMICS (SRI-VIPRA)

Project Report of 2024: SVP 2453

SRI-VIPRA

"Emerging Trends in Data Analytics: A literature review"

s

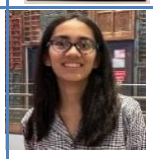
IQAC
Sri Venkateswara College
University of Delhi
Benito Juarez Road, Dhaula Kuan
New Delhi -110021

SRI-VIPRA PROJECT 2024

Title: Emerging Trends in Data Analytics: A literature review

Details of Mentors	
Name of Mentor: Dr. Tarakeswara Rao K **Name of Department:** Electronics **Designation:** Assistant Professor	**Name of Mentor:** Dr. Ravindra Varma P **Name of Department:** Biochemistry **Designation:** Assistant Professor
	

List of students under the SRIVIPRA Project

S. No	Photo	Name of the student	Roll number	Course	Signature
1		Adittee Joshi	1223037	B.Sc. (H) Biochemistry	
2		Prachi Kumar	1222005	B.Sc. (H) Biochemistry	
3		Atiya Ahmad	1222002	B.Sc. (H) Biochemistry	
4		Manasvi Mishra	1722034	B.Sc. (H) Mathematics	
5		Medha Bhatnagar	1223058	B.Sc. (H) Biochemistry	

P. Ravindra Varma

K. Tarakeswara Rao

Signature of Mentor **Signature of Mentor**

CERTIFICATE OF ORIGINALITY

This is to certify that the aforementioned students from Sri Venkateswara College have participated in the summer project SVP-2453 titled "**Emerging trends in data analytics: A literature review"**. The participants have carried out the research project work under my guidance and supervision from **1st July, 2024** to **30th September 2024**. The work carried out is original and carried out in an online/offline/hybrid mode.

Signature of Mentor

Signature of Mentor

ACKNOWLEDGMENT

We would like to express our sincere thanks to the Principal of Sri Venkateswara College, University of Delhi for organizing the SRI VIPRA internship Program, 2024. This program has helped us to put our scientific and computational understanding to application.

We would like to express our gratitude and thanks to our professor and mentor, Dr. Tarakeswara Rao Kaviti and Dr. Ravindra Varma Polisetty for guiding us in this project. They provided us with their valuable suggestions throughout the process. They encouraged us to undertake this project from time to time.

We would also like to thank the coordinators of SRI VIPRA program 2024 for providing us the opportunity to be a part of this SRI VIPRA PROJECT 2024.

We would also like to thank our family and friends for supporting us throughout this project.

We are a team of five students Adittee, Atiya, Manasvi, Medha and Prachi who have worked together with the support and encouragement from our professors and peers towards the successful completion of their project. Our consistent efforts combined with our dedication and team work helped us to achieve this.

TABLE OF CONTENTS

LIST OF FIGURES

15	Comparative time series of NO concentrations in Delhi from March 23rd to 15th April (2019 vs 2020)
16	Comparative time series of PM10 concentrations in Delhi from March 3rd to 28th April (2019 vs 2020)
17	Comparative time series of PM2.5 concentrations in Delhi from March 3rd to 28th April (2019 vs 2020)
18	Comparative time series of NO2 concentrations in Delhi from March 23rd to 15th April (2019 vs 2020)
19	Univariate Analysis of continuous variables using histograms
20	Univariate Analysis of categorical variable using horizontal bar plots
21	Bivariate Analysis of continuous features
22	Bivariate Analysis of categorical Features using stacked bar plots
23	Histogram, Scatter plot, Correlation matrix, bar graph
24	Pair Plot of key features in relation to target
25	Confusion Matrix Models
26	Accuracy of different models

ABSTRACT

Data, raw facts collected from various sources, is transformed into insights through data analytics, a systemic process involving cleaning, transforming, and modelling data to uncover trends. Data analytics is, thus, heavily used in various sectors, from public policy to commercial business ventures to healthcare industry, and, essential for future advancements.

'Data analytics' has undergone changes with time. It evolved from Pre-Digital Era, with early descriptive statistics, to Early-Digital Era, whence computers, structured data, DBMS and SQL were developed, to Business Intelligence Era, which furthered data use and saw rise of tools like PowerBI, EDW and OLAP, to Big Data Era, which saw use of data characterised by its volume, velocity, and variety and utilised tools like Hadoop and Spark, to the modern, AI-ML Era, which is seeing the use of artificial intelligence and machine learning to make predictive analyses and figure hidden trends. With data being so important, it is necessary to explore what more can be done with it. Advent of IoT, cloud computing, LLMs and Generative AI have pushed the boundaries of our sphere of knowledge, possibilities and abilities of what we can do. Data is utilised in many fields, including but not limited to healthcare, finance, environment, retail and e-commerce, public policy and manufacturing, yet efficient extraction of information is complicated by its own set of challenges, such as huge volume of data, poor data quality, data integration and management and data storage. We studied the past trends and did a study on emerging trends that are expected in the future.

Machine learning algorithms (Supervised, Unsupervised, Semi-supervised, Reinforcement and Deep learning) were employed. Techniques used include Decision trees, Random Forest, Support vector machines (SVM) and k-Nearest Neighbour (k-NN).

Two case studies (Air Pollution dataset of Indian cities from 2015-2020 and heart disease dataset) were performed to understand data analysis and to develop accurate models and suitable visualisation techniques. The datasets were downloaded from Kaggle and analysed stepwise.

I. INTRODUCTION

1.1 Introduction

Data is raw, unprocessed, and unorganised facts and figures extracted from various reliable sources. It ranges from databases and spreadsheets to text, images, and videos and can come in practically any form. The process of converting raw data to a more usable form is known as data analytics.

Data analytics is the science of examining raw data to draw inferences by applying various algorithms to identify trends and patterns statistically. It is a multifaceted and comprehensive process that includes cleaning, transforming, and modelling data. It helps to discover information that supports decision-making.

Data is indispensable in nearly every sector. The role of data in today's society is profound with numerous future implications. In the business sector, it serves as a tool to understand consumer behaviour and predict future trends to stay ahead of potential competitors. In healthcare, personalised medicine and predictive analytics have helped with anticipating and managing disease outbreaks. In the public sector, data has transformed the process of policy-making with increased transparency, and improved governance.

With the rise of artificial intelligence and machine learning, dependence on large datasets to train models has also been rising. The main aim is to automate tasks, identify patterns, and make decisions that can open up innovative avenues for the future. Thus, studying emerging trends in data analytics has become an essential pursuit.

1.2. Evolution In Data Analytics

Data science as a concept preceded big data chronologically. In the 1960s, only one paper discussed BD, but 52 papers dealt with DS, based on a search of Clarivate Analytics, part of the Web of Science database. DS tends to be more disciplinarily dispersed, with some focus on computer science and environmental sciences. It can be traced back to the early applications of statistics and mathematics, which have now steadily transformed into complex AI-driven tools that harness huge amounts of data for predictive insights.

Data analytics is a continuously evolving field whose history can be demarcated into several key phases, driven by technological advancements and shifts in focus, from descriptive to predictive to prescriptive analytics.

1.2.1 Pre-Digital Era

The foundation of data analytics was laid with descriptive statistics and simple data summarization techniques, which involved summarising historical data to understand patterns and trends. Early methods such as mean, median, and frequency distribution were used. This approach was predominantly retrospective. The methods used were rudimentary and manual, usually slow and labour-intensive. Data was collected using surveys, forms, and human observation. The primary goal was to maintain accurate records rather than derive any insights. Decision-making was based more on intuition and experience than data-driven insights.
Tools: Hand calculations and early mechanical calculators

1.2.2 Early Digital Era

This era was marked by the introduction of computers, which changed how we collect and process data. This period saw more structured approaches to data management and analytics. Companies began using mainframe computers and relational databases to store and analyse data more efficiently. Relational database management systems and EIS were developed. Structured data, primarily as transactional records, was stored in databases, and SQL became the primary language for data management. Limited data storage and computing power constrained the scale at which data could be analysed. Additionally, descriptive analytics could not offer foresight, as it only analysed past events. The primary focus in this era was the automation of data storage and retrieval.

This era marked the beginning of significant advancements in the field of data analytics. It was driven by the advent of computers and basic software tools.

1.2.3 Business Intelligence Era

The business intelligence (BI) era saw the rise of tools designed to help businesses make more informed decisions by analysing historical data. During this period, data was increasingly used to create reports, dashboards, and visualizations. Tools like SAP BusinessObjects and Microsoft Power BI emerged to offer business leaders easy access to data insights. The focus during this period was on descriptive analytics, which helped organizations understand "what happened" based on historical data.

Technologies: The growth of enterprise data warehouses (EDW) and OLAP (Online Analytical Processing) systems played a critical role in aggregating and visualizing data.

Business Application: Organizations began integrating BI systems to monitor performance metrics and key performance indicators (KPIs), leading to more data-driven decisions.

Limitations: Despite advances, these systems relied on structured data observed in the past and did not output predictive insights.

1.2.4 Big Data Era

Big data refers to a huge and complex collection of data from various sources that continues to grow over time and cannot be stored or analysed by using data processing softwares.

The **3 Vs of big data** are:

1. Volume (amount of data) – involves huge volumes of data in petabytes or exabytes.
2. Velocity (rate at which data is received) – Data streams directly into memory and hence the rate of data transmission is very fast. As a result, the volume of data grows exponentially over time and requires real-time evaluation and action.
3. Variety – Big data involves structured (data having a standardised format like dates /emails /latitudes and longitudes), unstructured (PDFs, Images, Audio files, Video files) and semi-structured data (JSON, XML, HTML code).

Big data originated in the 1960s and '70s with the development of the first **data centres** and the development of the relational database. It involves the storage, extraction, transformation and optimization of data stored in **RDBMS** (Relational database management systems) systems.

From the early 2000s, the internet, online services and web applications started to generate tremendous amounts of web-based unstructured data which provided organisations with a new form of knowledge: insights into the needs and behaviour of customers. Furthermore, the development of open-source frameworks, such as **Hadoop and Spark** made it easier to work with big data and real-time data evaluation became more prevalent.

1.2.5 Data Science and Machine Learning Era

Big data eventually led to the emergence of the field of data science and prevalence of machine learning and artificial intelligence in the field of analytics.

Data Science: It is the study of large amounts of complex data in a company or organisation. It includes where the data has originated from, the actual study of its content matter, and how this data can be useful for predicting the future growth of the company. This helps to get valuable information about the business and market patterns which helps the business have an edge over the other competitors by increasing their efficiency.

Machine Learning: Machine Learning is the field of study which gives computers the ability to learn without being programmed. It uses algorithms to process the data and helps to train the computer for making future predictions without human intervention.

With breakthroughs in deep learning, natural language processing and the convergence of AI with other technologies, data science and machine learning promise greater advancements.

1.2.6 Emerging Trends and Future Directions

With data being an integral part of making data-driven business decisions and getting insights it has become very important to explore trends about what we can achieve with it. Data keeps growing and as a result looking at new techniques to analyse it becomes necessary.

With the advent of the **Internet of Things (IoT)**, more objects and devices are connected to the internet and they can communicate and exchange data with each other. It is widely used for real time communication with users through smart homes/appliances, smart watches, RFID enabled ID cards, etc. This helps to gather data on customer usage patterns and product performance.

Cloud computing has expanded big data possibilities even further where developers work collaboratively on datasets. It helps to offer computer services over the internet and helps in faster work. Example- Google Cloud, AWS, Microsoft Azure, etc. It is also widely used in **SaaS (Software as a Service)** for delivering services to users. And **relational databases** are becoming increasingly important as well, with their ability to display massive amounts of data in a way that makes analytics fast and comprehensive.

Large language models are trained on large datasets and they help to comprehend complex language and generate human-like text responses. **Generative AI** which also involves the use of LLM's is becoming increasingly popular with the advent of platforms like **ChatGPT and Gemini**. They are increasingly being used to create new content which further pushes the boundaries of what we can possibly achieve with technology and AI. All these tools can be used to look at the future scope of analytics in the following ways:

a) **Augmented Analytics**: This will majorly focus on increasing accessibility of various AI and ML tools to non-experts. It would also help to shift towards real-time data processing at all the levels in an organisation because analytics would become more simplified and no technical expertise would be required. It would also promote faster decision making and would increase automation.

b) **Edge Analytics**: Edge computing is a distributed IT architecture in which client data is processed at the periphery of the network. It offers a path to collect data from devices via low-latency connectivity, high-performance processing, and

secure platforms. Performing data analysis at the source (e.g., IoT devices) to reduce latency and bandwidth use, which is crucial for industries like manufacturing and transportation.

c) **Predictive and Prescriptive Analytics**: It involves the use of ML to make models and build algorithms for solving real-life business problems. These approaches provide foresight into future trends and help organisations in better decision making, optimise processes in the long term and strategize effectively. Even the system of recommendations on various social media platforms (like YouTube, Instagram) makes use of this to recommend videos based on the past viewing history of the person.

d) **Natural Language Processing (NLP)**: It helps computers to interpret human speech and language. It involves interaction through an interface, allowing users to query data and gain insights using everyday language. It has a wide range of applications including chatbots, smart assistant, text analysis, translation and sentiment analysis among others.

e) **Data Privacy and Ethics**: These norms are used to ensure that data is used in a responsible way while respecting the privacy of users. Increased focus on responsible data usage and compliance with regulations like GDPR (General Data Protection Regulation), drives organisations to adopt ethical data practices and ensure transparency. In India, the Digital Personal Data Protection Act (DPDP) is a comprehensive data protection law that aims to ensure cybersecurity over online platforms.

f) **Data Visualization Evolution**: With the advent of AI, data visualization has become more advanced as the software can automatically detect relationships between various numerical and categorical variables and it tries to suggest various representations which would be well-suited to the dataset. The visualization tools provide more interactive and intuitive ways to present complex data, thus, enhancing understanding and storytelling. Moreover, tools like Tableau and Power BI also help to develop interactive dashboards that help to look at the trends in various metrics simultaneously.

1.3 Applications and Impact

Data is being produced in all sectors of the world today. The data available to us needs to be used optimally and intelligently to produce the desired results. It helps us understand what happened in the past and what needs to be done in the future to create informed decisions and strategies to help run the sectors efficiently.

1.3.1 Healthcare Sector

The Healthcare sector involves great usage of data analytics throughout the world. Common data sources in healthcare include medical research / journals, biometric data, Internet of Medical Things (IoMT), social media, payer records, omics research, data banks, Electronic Health Records (EHR), electronic medical records, personal health records, and public health records. The analysis of data helps us in identifying the root cause of a particular disease. It can further prevent its occurrence or an outbreak that might affect larger populations.

The processing of healthcare data serves several key objectives:

1. Understanding historical events
2. Determining Causes
3. Predicting future events
4. Prescribing future actions

Data analysis in the healthcare sector includes areas like **epidemiology** and **genomics**.

- The field of *epidemiology* has long played a pivotal role in monitoring, understanding, and mitigating the impact of infectious diseases on human populations. Recent advances in data collection and machine learning techniques present an unprecedented opportunity to revolutionize epidemiological research.
- *Genomics* also includes storage of the genetic data, analysing it and drawing different conclusions from it. This can be helpful in giving solutions for diseases related to genes.

1.3.2 Finance Sector

Another sector which is greatly impacted by the use of data analytics is the financial sector. Financial analytics helps answer business questions and also forecast possible future financial scenarios or shape business strategy. All of this is done to boost a company's value.

Financial analytics help the companies understand the risks they might face. They also help them understand the better business processes that will help them run the businesses effectively without the wastage of resources and also how they can optimise the organisation's investments in all the areas. Detecting unusual patterns can help in fraud detections.

Beyond aiding decision-making, financial analytics enhances transparency by allowing analysts to examine the reasons behind decisions and factors affecting the decision-making process. The future of data analysis in the financial sector includes leveraging

machine learning models for market trend predictions, fraud detection, real-time financial monitoring, and investment predictive analysis.

1.3.3 Environment Sector

Environmental analytics is a field where data collection and analysis are extensively applied. Data is typically gathered through methods such as satellite imagery, sensors, and environmental monitoring stations. Key tools and techniques in this field include remote sensing, predictive modelling, and geographic information systems (GIS), which help visualise data and make predictions based on both current and historical trends. This analysis can monitor changes in climate patterns, air and water quality, biodiversity, and land use. It is used to develop strategies for mitigating climate change effects, managing natural disasters, and creating sustainable future plans.

1.3.4 Other Fields

Data analytics is increasingly expanding across numerous fields such as retail and e-commerce, telecommunications, manufacturing and supply chain management, and insurance, among others.

In each of these sectors, the application of data analytics provides numerous benefits. These benefits include acquiring valuable insights into market trends and consumer behaviour, which can lead to more informed decision-making and strategic planning. This in turn helps in enhancing customer experience by personalising interactions and services based on individual preferences and behaviours. It also plays a crucial role in optimising resource allocation, ensuring that resources are used efficiently and effectively. Furthermore, it aids in

streamlining operations by identifying inefficiencies and improving processes, leading to overall operational improvements and cost savings. This widespread application of data analytics helps organisations across these fields to remain competitive and adapt to changing market conditions.

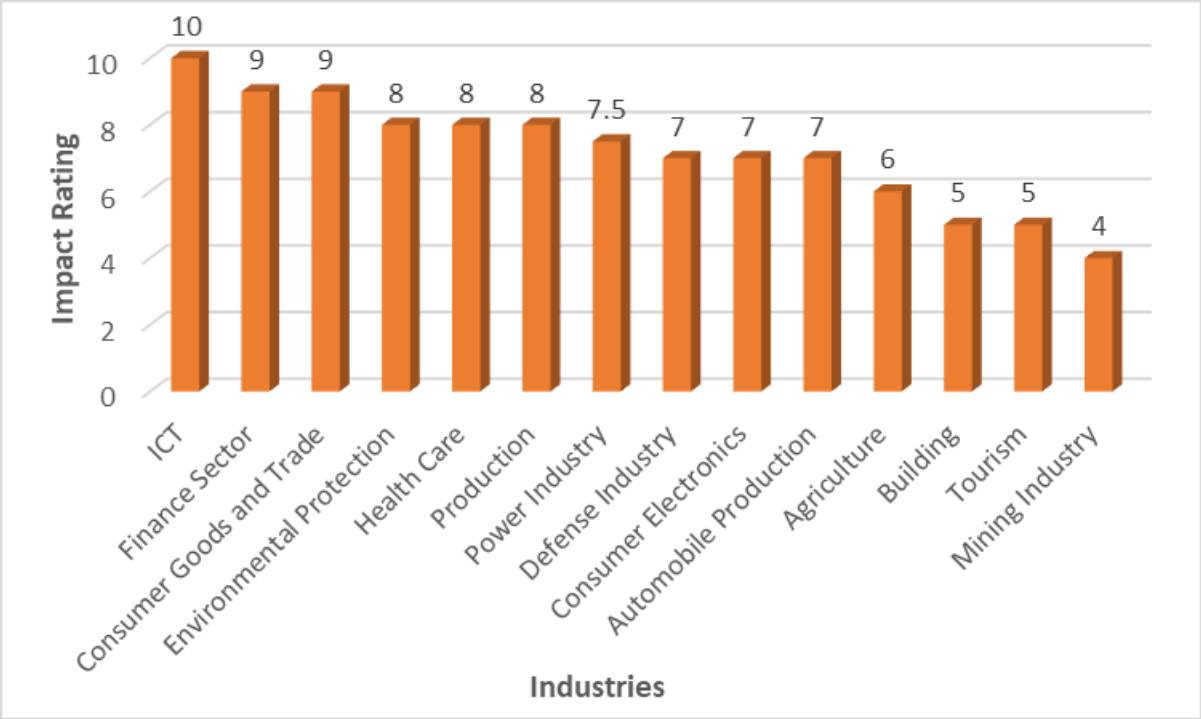


Figure 1: The impact of various industries on the Big Data analytics market (Assessed by Frost & Sullivan Company)

1.4 Challenges Faced in Data Analytics

With the evolution of data analytics ever transforming the way copious amounts of data are stored and processed, ever-increasing the dependence of organisations on its techniques, various challenges also arise. Different industries are dependent on data analytics for the extraction of valuable information efficiently and effectively but are complicated by their own set of issues.

1.4.1 Huge Volume of Data

The amount of data being produced has seen an exponential rise, where in the 2010s, 2 zettabytes (ZB) of data was being generated annually, the amount of data generated sky rocketed to 64.2 ZB by 2020. It is expected that nearly 180 ZB of data will be produced annually by 2025. With the expansion of social media, networked systems, and the Internet of Things (IoT), this surge in data production will continue.

Existing systems may not be equipped to process the massive influx of data and can lead to inconsistencies, errors, and duplicates, thereby decreasing data quality. More data requires more storage capacity, rendering traditional storage systems inadequate and increasing the need for cloud storage or distributed databases. Expenses also increase for organisations to maintain backup storage systems as well as invest in technologies focused on real-time analytics to quickly extract the required information.

1.4.2 Poor Data Quality

Data analytics can only be efficient if the data is of high quality, that is, precise, comprehensive, consistent, up-to-date, and pertinent to its intended use.

Incomplete Data: It arises when either the data is not freely available to those who require it, making it difficult for organisations to extract valuable information and arrive at meaningful conclusions. Incomplete data can also arise owing to technical glitches, data entry errors, or omissions, making the dataset incomplete, which can hinder, slow down, and distort the results that could have been of value.

Unstructured Data: Traditional databases are trained to deal with structured data (organised information that fits into a predefined format), while unstructured data (lacking a predefined model) comes in variable and inconsistent formats (text files, videos, social media posts), posing difficulty in its integration requiring significant preprocessing, is dependent on specialised storage systems (NoSQL databases, data lakes, or cloud storage solutions), and advanced algorithms and machine learning models for its interpretation.

Inaccurate Data: Errors or mistakes, which could be due to incorrect data collection methods, mis inputs due to human or technical reasons, or multiple inputs causing duplicate records can skew the data, affecting the integrity of outcomes, leading to false and incorrect conclusions.

1.4.3 Data Integration and Management

The data to be stored and analysed is received from various sources with their own structure, quality, and style. The integration of such data into a unified dataset becomes complicated owing to the standardisation and normalisation required by the variety of data. The merging of diverse data can lead to inconsistencies, errors, and discrepancies that degrade the data quality. Moreover, data silos, where data becomes isolated and fragmented relative to their departments, lead to time and money being wasted in navigating between different data resources and cross-referencing for extracting valuable information. This puts focus on the need for centralised platforms, tools, and techniques whose adoption is crucial for a unified dataset to overcome the difficulty posed.

1.4.4 Real Time Analytics

In various industries, such as marketing, healthcare, and banking, examining data to obtain real-time insights is extremely important for adapting and making quick decisions. Manual processing of data requires more time, and by the time the conclusion is obtained, the dataset on which it is based becomes redundant, leading to losses and harming the decision-making process. Machine Learning (ML) and Artificial Intelligence (AI) algorithms play a key role in real time analytics for identification of

patterns and processing key information from data as soon as it is created allowing for well-informed decisions to be made and enable immediate action.

I.5 Machine Learning Algorithms

Machine Learning Algorithms employs computational methods to analyse datasets, patterns and trends to enhance the process of automation and decision-making process. The main types of ML algorithms include:

A. Supervised Learning Algorithms: This algorithm is trained on a labelled dataset, where the input data and corresponding output labels are known. The aim is to learn a mapping from inputs to outputs and use this mapping to make predictions on new, unseen data. Examples include: **Linear Regression** which predicts a continuous value based on the input variables, **Logistic Regression** which is used for binary classification problems, **Decision Trees** which utilises a tree-like model of decisions based on features, **Support Vector Machines (SVM)** which finds the best boundary between classes, **k-Nearest Neighbours (k-NN)** which classifies data points based on their nearest neighbours, **Neural Networks** that includes a set of algorithms inspired by the human brain to recognize patterns.

B. Unsupervised Learning Algorithm: This algorithm deals with finding hidden patterns or structures within the given data without explicit labels. Examples include: **K-Means Clustering** which partitions data into clusters based on feature similarity.

C. Semi-Supervised Learning Algorithms: These algorithms fall between supervised and unsupervised learning. They use a small amount of labelled data and a large amount of unlabelled data for training. Examples include: **Self-Training** which uses its predictions on unlabelled data to retrain itself.

D. Reinforcement Learning Algorithms: This algorithm is based on a system of rewards and penalties and deals with interacting with an environment and learns to perform a task by receiving feedback based on its actions. Examples include: **Q-Learning** which is a value-based approach where an agent learns the value of actions in states.

E. Deep Learning Algorithms: This is a subset of machine learning that uses neural networks with multiple layers (deep networks) to model complex patterns in data. It is especially useful for tasks like image recognition, natural language processing, and speech recognition.

1.5.1 Decision Tree

Decision trees are a popular and intuitive method for making decisions based on data. Imagine the decision process as a branching structure where each split breaks down a large question into smaller, more manageable ones. At each branching point, the algorithm asks a question based on a feature (like whether an animal has feathers or not) and follows a path based on the answer. The process continues until the data is organized into clear, distinct groups, or until a final prediction is made, like categorizing an object or estimating a value.

The key idea behind decision trees is to simplify complex decisions by finding the features that give the most valuable information at each step. For example, when classifying animals, the tree will first ask the most informative question, such as "Does it have wings?" and based on the answer, will ask more specific questions until the animal is correctly classified.

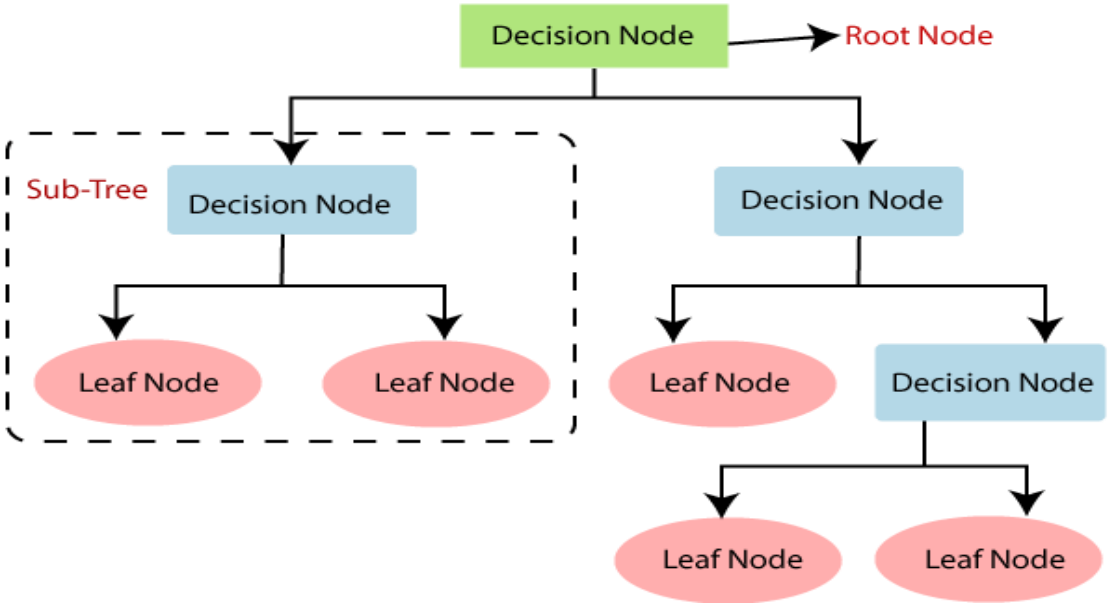


Figure 2: Decision Tree Algorithm in Machine Learning

This method is widely used because it's easy to understand and interpret. However, decision trees can sometimes be too good at fitting the data, which can lead to overfitting—where the tree performs well on the training data but struggles with new, unseen data. To overcome this, more advanced techniques like Random Forests and Gradient Boosting have been developed to reduce these weaknesses while keeping the strengths of decision trees.

This process is recursive, with each internal node of the tree representing a feature, each branch representing a decision rule, and each leaf node representing a class label (in classification) or a numerical value (in regression). The driving principle behind decision trees is to maximise the purity of the nodes. For classification tasks, this involves minimising **impurity** (e.g., Gini impurity or entropy), while for regression

tasks, it involves minimising **variance**. The goal is to partition the dataset in such a way that the data within each subset (or node) becomes as homogeneous as possible.

1. **Splitting Criterion**:
    - **Gini Impurity**: A common measure used for classification tasks; it quantifies the likelihood of misclassifying an instance if it were randomly classified according to the distribution of classes in the node.
    - **Entropy (Information Gain)**: This measure is based on the concept of entropy from information theory, where splits are made to maximize information gain by reducing entropy.
    - **Variance Reduction**: In regression tasks, the algorithm selects splits that minimize the variance within the resulting nodes.
2. **Recursive Binary Splitting**: The decision tree algorithm starts at the root node and splits the data into two groups based on the chosen feature and splitting criterion. This process is repeated at each node, creating a binary tree structure, until a stopping criterion is met (such as a minimum number of samples in a node or a maximum tree depth).
3. **Pruning**: To avoid overfitting, decision trees often use a technique called pruning, which involves removing branches that add little predictive power. Pruning can be either **pre-pruning** (halting the growth of the tree early based on certain conditions) or **post-pruning** (allowing the tree to grow fully, then cutting back some branches).

Key advantages include:

1. **Interpretability**: The tree structure clearly lays out the decision-making process, making it easy to understand the logic behind predictions. This is particularly important in fields like healthcare and finance, where interpretability is crucial for decision-making.
2. **No Need for Feature Scaling**: Unlike algorithms like Support Vector Machines or k-Nearest Neighbours, decision trees do not require feature scaling (e.g., normalization or standardization). This is because the algorithm makes decisions based on the ordering of features rather than their magnitude.
3. **Handling of Categorical and Numerical Data**: Decision trees can handle both categorical and numerical data effectively. They can split categorical data by considering each category as a separate decision and numerical data by finding optimal thresholds for splits.
4. **Non-Parametric Nature**: Decision trees are non-parametric models, meaning they do not assume a specific distribution for the data. This makes them highly

flexible and capable of modelling complex relationships between features and outputs.

5. **Robust to Outliers**: Since decision trees divide the data into regions based on feature thresholds, they are relatively unaffected by outliers, which only impact specific branches of the tree.

Key limitations include:

1. **Overfitting**: Decision trees are highly susceptible to overfitting, especially when they are allowed to grow too deep. This leads to models that perform well on training data but generalize poorly to new data. Techniques like pruning, setting a maximum depth, or using ensemble methods (e.g., Random Forests) are needed to mitigate this issue.
2. **High Variance**: Decision trees can exhibit high variance, meaning that small changes in the data can lead to significantly different trees. This instability can be problematic when building predictive models for real-world applications.

1.5.2 Random Forest

Random Forests are a highly effective machine learning technique used for both classification and regression tasks, designed to address common challenges like overfitting and high variance. Introduced by Leo Breiman in 2001, this method builds on decision trees, which split data based on features to make predictions. Instead of relying on just one decision tree, Random Forests combine the power of many trees to create a more reliable model.

The strength of Random Forests comes from ensemble learning—specifically through techniques called Bagging (Bootstrap Aggregating) and random feature selection. In Bagging, multiple subsets of the training data are created by sampling with replacement, meaning some data points can be selected more than once. Each subset is then used to train a separate decision tree. This creates a forest of decision trees, each learning from slightly different data. To further increase variety, Random Forests add randomness by allowing each tree to pick a random subset of features when making decisions. This reduces the chance that any two trees will produce similar results, ensuring the model stays diverse and avoids overfitting.

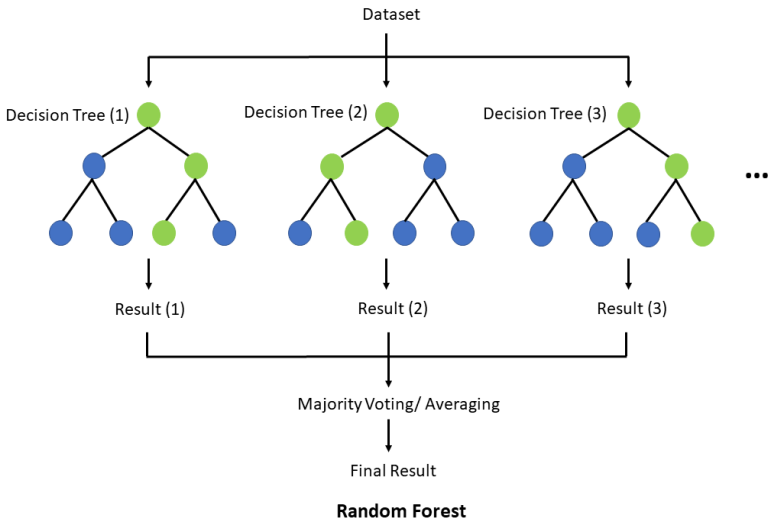


Figure 3: Random Forest Algorithm in Machine Learning

Once all the trees have made their predictions, the results are combined to form a final output. For classification tasks, this is done by majority voting (whichever class most trees predict wins), and for regression tasks, by averaging the predictions. By blending the predictions of many "weak" decision trees, Random Forests create a strong, accurate model that performs well even with complex or noisy data. This makes it a go-to choice for a wide range of machine learning applications.

Key advantages include:

1. **Reduced Overfitting**: Traditional decision trees tend to overfit on training data, capturing noise and specific patterns that do not generalise well to unseen data. Random Forests mitigate this by averaging multiple trees, which leads to a more generalizable model.
2. **High Accuracy**: By utilising a large number of decision trees and selecting random subsets of features, Random Forests often outperform other algorithms in classification and regression tasks.
3. **Handling Missing Data**: Random Forests are robust to missing data. If a feature is missing during training or prediction, the algorithm can simply choose a substitute feature for decision-making.
4. **Feature Importance**: Random Forests provide insights into the importance of each feature. This is helpful in feature selection processes, where less important features can be discarded to simplify the model without a significant drop in performance.
5. **Scalability**: Random Forests scale well to large datasets with high dimensionality, making them suitable for a variety of practical applications.

Key limitations include:

1. **Interpretability**: One of the main drawbacks of Random Forests is their lack of interpretability. While decision trees are relatively easy to interpret, the aggregation of hundreds or thousands of trees makes it difficult to understand the reasoning behind specific predictions.
2. **Resource-Intensive**: Random Forests can be computationally expensive, both in terms of time and memory, especially when dealing with very large datasets or a high number of trees. For real-time applications, where speed is critical, this can be a significant drawback.

1.5.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) are a powerful supervised learning algorithm widely used for both classification and regression tasks. SVMs were first introduced by Vladimir Vapnik and Alexey Chervonenkis in 1963 and gained popularity in the 1990s with advancements in computational power and kernel methods. The key idea behind SVMs is to find a hyperplane that best separates data points belonging to different classes in a high-dimensional space. By doing so, SVMs create a decision boundary that maximizes the margin between classes, ensuring better generalization and reducing classification error.

At the core of SVMs lies the concept of **support vectors**, which are the data points closest to the decision boundary. These support vectors are critical because they define the hyperplane and the margin. In simple terms, SVMs aim to find the hyperplane that maximizes the margin between the support vectors of different classes. In the case of linearly separable data, this hyperplane is a straight line (in 2D) or a flat surface (in higher dimensions). However, not all data is linearly separable, which is where kernel methods come into play.

SVMs use **kernel functions** to transform data into higher dimensions, allowing them to handle non-linear relationships between features. Common kernel functions include the **linear kernel**, **polynomial kernel**, and **Radial Basis Function (RBF)** kernel. By mapping data into higher dimensions, SVMs can identify a hyperplane that separates classes in this transformed space, even if the data is not linearly separable in its original form.

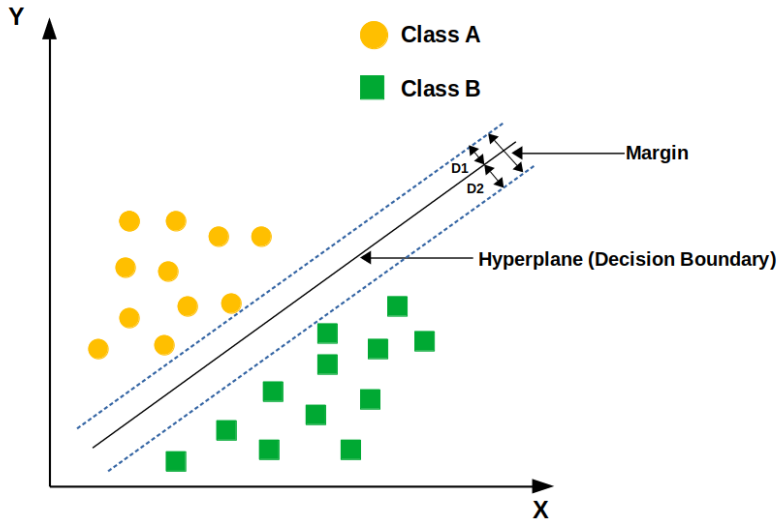


Figure 4: Support Vector Machine (SVM) Classification with Optimal Hyperplane and Margin: SVM model separating Class A (yellow circles) and Class B (green squares). The solid black line represents the hyperplane, and the dashed blue lines define the margin, which is maximized for better classification. Support vectors are the closest points to the hyperplane, and D1, D2 indicate the distances from the hyperplane to the nearest points in each class.

Key advantages of SVM:

1. **Effective in High-Dimensional Spaces:** SVMs work effectively with datasets that have a large number of features, even when there are more features than data points.
2. **Robust against Overfitting**: Support vector machines (SVMs) are less likely to overfit when they concentrate on optimizing the margin between support vectors, particularly in high-dimensional feature spaces.
3. **Non-Linear Decision Boundaries:** Compared to linear models, SVMs offer greater flexibility in modelling non-linear connections between features by utilizing kernels.
4. **Versatility:** Support vector machines (SVMs) can be used to solve regression (also known as Support Vector Regression, or SVR) and classification issues. With some modifications, they can also handle binary and multi-class categorization.

Limitations of SVM:

1. **Expensive Computation:** Because sophisticated quadratic programming issues must be solved, training SVMs can be slow, especially when dealing with huge datasets. Because of this, SVM is less useful for complex issues.

2. **Sensitive to Kernel and Parameter Selection:** The SVM's performance greatly depends on selecting the kernel and regularization parameters correctly. Poor performance or overfitting may result from improper tuning.
3. **Memory-Heavy:** SVM can become memory-heavy since it uses support vectors to make predictions. This is particularly true for large datasets that produce a high number of support vectors.
4. **Lack of Probabilistic Interpretation:** SVM does not inherently produce probabilistic results, in contrast to techniques like Random Forests or logistic regression. To get probability estimates, though, techniques like Platt scaling can be applied.

1.5.4 k-Nearest Neighbours (k-NN)

k-Nearest Neighbours (k-NN) is a simple yet effective machine learning algorithm used for both classification and regression tasks. Developed as a non-parametric method, k-NN relies on the principle of similarity, where the classification or prediction of an unknown data point is based on the characteristics of its closest neighbours in the feature space.

At the core of k-NN is the idea that similar data points exist in close proximity to one another. In the context of classification, k-NN works by assigning a label to a new data point based on the majority class of its "k" nearest neighbours, where "k" is a user-defined parameter. In regression tasks, the algorithm predicts the value of a new data point by averaging the values of its nearest neighbours. The choice of "k" plays a critical role in determining the algorithm's behaviour—small values of "k" can make the model sensitive to noise, while large values can lead to overgeneralization.

k-NN calculates the distance between data points to identify the closest neighbours. The most commonly used distance metric is Euclidean distance, but others, like Manhattan or Minkowski distances, can be applied depending on the nature of the problem. For each new data point, k-NN searches through the entire training dataset to find the closest "k" neighbours and then bases its prediction on these neighbours’ characteristics.

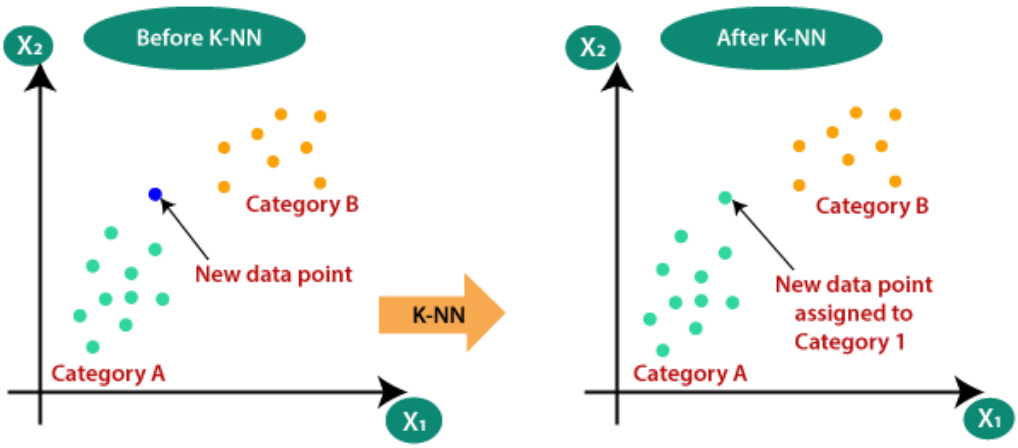


Figure 5: k-Nearest Neighbours (k-NN) Classification, Left Panel: Before k-NN (**Blue point** represents a new data point that is to be classified based on its nearest neighbours), Right Panel: After k-NN (new data point (red) is classified as belonging to **Category A**, indicating that its nearest neighbours were primarily from this category)

Advantages of k-NN:

- **Simplicity:** k-NN is simple to comprehend and use. Since predictions are made immediately from the stored training data, there is no need for an explicit training process.
- **Adaptability:** k-NN is a flexible algorithm that may be used for both regression and classification tasks. It functions effectively with non-linear decision boundaries without requiring intricate data modifications.
- **No Data Presumptions:** Since k-NN is a non-parametric technique, it doesn't make any assumptions on the distribution of the underlying data. This enables it to function effectively in situations where more complex or data sets with unknown distributions prove problematic for older approaches.
- **Takes Care of Multi-class Classification:** k-NN is useful for a variety of jobs since it can effortlessly handle multi-class classification problems by only taking the majority class among neighbours.

Limitations of k-NN:

- **Computational Complexity:** One of k-NN's primary shortcomings is that, particularly for large datasets, it is computationally expensive to calculate the distances between each data point in the training set. This may cause the prediction process to lag noticeably.

- **Sensitivity to Irrelevant Features:** Since all features are evaluated equally in the distance calculation, k-NN's performance may suffer if the dataset has a large number of irrelevant or noisy features.
- **Curse of dimensionality:** As the dimensionality of the feature space increases, the distance between data points becomes less meaningful, leading to poor performance. This is a significant issue for high-dimensional datasets.

I.6 Literature Review

The literature review delves into the landscape of dana analytics discussing key research, methodologies and insights obtained from existing studies. It identifies emerging trends, particularly the influence of technologies like artificial intelligence, deep learning, and machine learning on data analysis practices. By illustrating how these advancements are reshaping the field, it also highlights the importance of interdisciplinary coherence, showcasing how data analysis techniques are applied across different domains.

It also lays the framework for future research by outlining the current state and suggesting directions for further studies, which is vital for advancing data analysis in response to new challenges. This contribution is precious for academics, professionals, and decision-makers who rely on data analysis to inform their choices and strategies and make valuable conclusions. Thus, it is not only establishing a comprehensive understanding of the current landscape of data analysis but also identifying critical gaps shaping up critical areas which can be focused upon, thus making it a vital component of the paper's overall contribution to the field of data analysis.

While significant advancements have been achieved in the field of data analytics, there are still some notable gaps in the existing literature. For instance, the scalability of algorithms continues to pose challenges, particularly when dealing with real-time or high-dimensional datasets. Additionally, the interpretability of complex models, such as deep learning, often falls short, which can restrict their use in fields that require transparency, like healthcare. Moreover, current studies do not sufficiently address ethical concerns related to data privacy and algorithmic bias. These gaps highlight the importance of pursuing further interdisciplinary and ethical research to enhance the development of data analytics in a more impactful manner.

For this project, research papers and articles were thoroughly examined to understand the efficacy and accuracy of numerous ML algorithms. Using the findings of the paper, four ML algorithms were the most accurate and effective in performing prediction analysis and were chosen for our dataset.

II. GOAL OF THE STUDY

The study aims to explore the potential of how data analytics can be used to tackle social issues by analyzing datasets related to air pollution and heart disease. By applying machine learning models, the study seeks to gain insights into prediction accuracy and identify trends that can help address these critical problems.

III. OBJECTIVES OF STUDY

The objective of this project is to visualise data and assess the accuracy of machine learning models applied to two datasets: one focusing on air pollution and the other on heart disease prediction. The project aims to evaluate the performance of various models in predicting outcomes related to each dataset, while also providing meaningful data visualisations to better understand the trends and patterns within the data.

IV. METHODOLOGY

In this paper we performed two case studies to understand the whole process of data analysis as well as develop accurate models and the best ways to visualise a given data. We took the Air pollution dataset of cities of India from 2015-2020 and the heart diseases data from Kaggle. The process of analysis was as follows:

1. Loading the dataset

It is the first step in any data analysis or machine learning project. The dataset can be loaded from various sources such as CSV files, XML files, JSON, databases, APIs, or other formats. We imported the dataset as .csv file *Steps:*

a) Import libraries: Libraries including pandas for data manipulation, NumPy for numerical operations and matplotlib and seaborn for data visualization were imported.

b) File location and reading: Data stored in CSV file was then read using pandas (pd.read_csv()). You could also load data from other formats such as Excel, JSON, or SQL databases.

c) Data exploration: After loading, the data was explored using functions like .head(), .info(), and .describe() to get a summary statistic, to understand the data and its variables and thus identify the target variable.

2. Data Transformation/Cleaning

Once the dataset is loaded, it needs cleaning. Raw data is messy and it needs some sorting and cleaning. Firstly, for the air pollution dataset, we narrowed down our focus of study to Delhi. Then the steps were applied as follows.

a) Handling missing data: In our air pollution dataset there were a lot of missing values that needed to be treated. We dealt with missing data by removing rows/columns with many missing values, filling missing values by interpolation. Heart disease data did not contain any missing data thus there was no need to perform this step.

b) Removing duplicates: Duplicates were identified and removed rows in the data to avoid bias or incorrect conclusions in the analysis.

c) Feature engineering: Scaling and encoding was done for the heart disease data. Scaling ensures that all the features contribute to the building of model equally than being dominated by features with larger magnitude.

d) Outlier detection and treatment: Outliers can skew the analysis and models. We identified outliers via IQR method and removed them. In both the datasets.

3. Data Visualization

Data visualization is a crucial part of exploring the dataset and gaining insights. It helps understand the underlying patterns, distributions, and relationships between variables. Visualization tools were used to explore the dataset. Python libraries like matplotlib, seaborn, and numpy were used in both the cases. For the air pollution data, line charts and scatter plots were used to understand the relationship between the air pollutants in air and the time span. Similarly, during the univariate analysis in the heart disease data, histograms were used to view central tendencies and shapes of the distribution of numerical data while horizontal bar charts were used for the categorical data. In the bivariate analysis, Line charts were used to understand the continuous features of the data variables and stacked bar plots were used to view categorical data.

4. Model Building

Once the dataset is pre-processed and explored, the next step is to build a model to analyse the data or make predictions. Model building typically involves selecting an appropriate algorithm, splitting the data, and training the model.

- Choosing the model: For classification problems, we selected Logistic Regression model, Decision Trees, Random Forest and Gaussian NB model, kNN and SVM.
- Defined the model
- Hyperparameter tuning of the model was done

- Model was finally evaluated

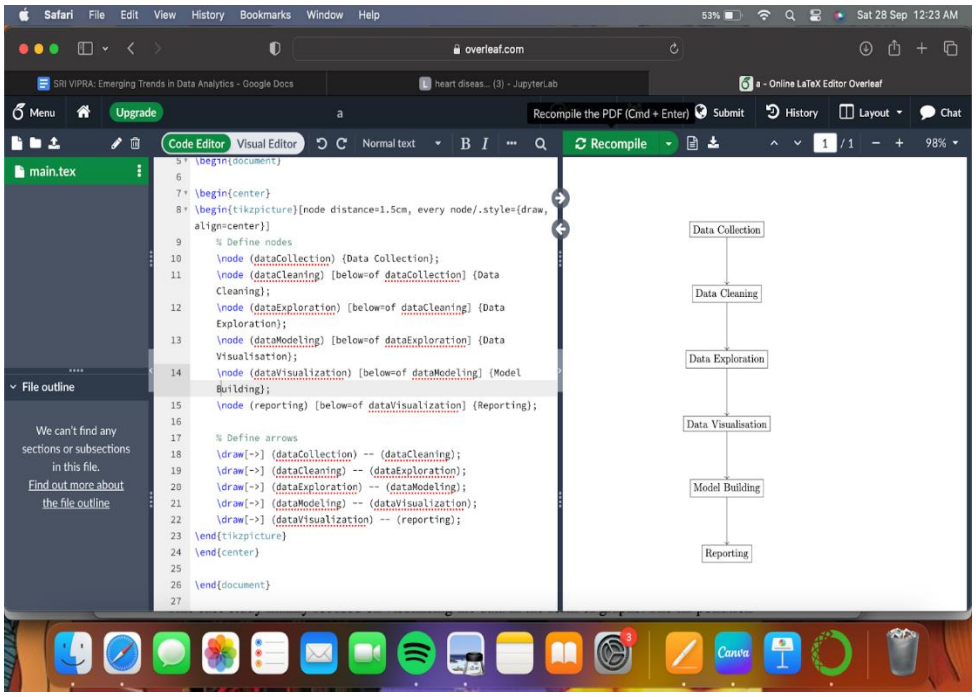


Figure 6: Process of Data Exploration through various steps

V. OBSERVATIONS

PART A- Air Pollution Dataset

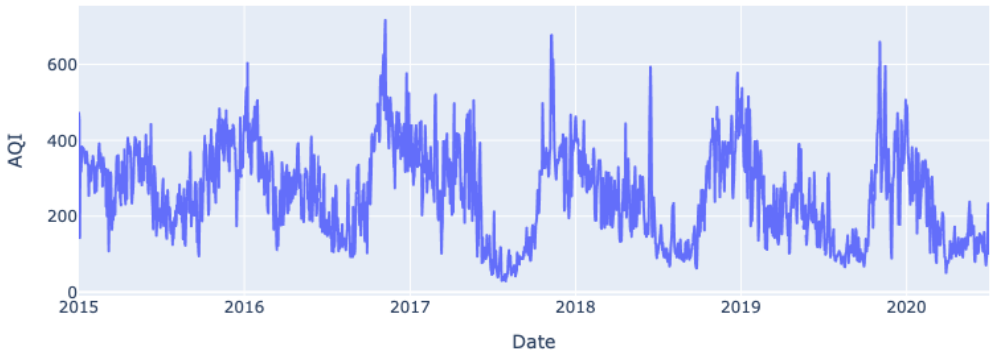


Figure 7: Time Series of AQI (Air Quality Index) in Delhi from 2015 to 2020

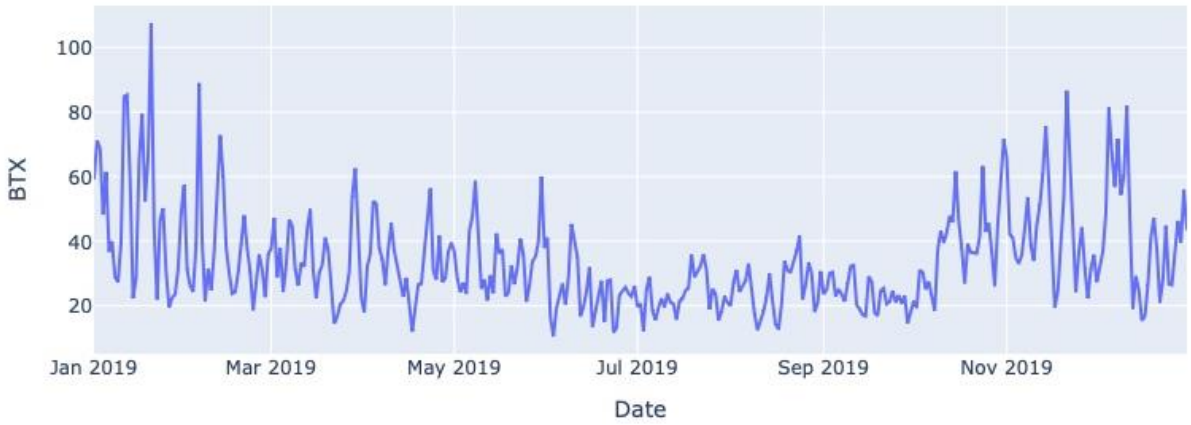


Figure 8: Time Series of BTX (Benzene, Toluene, Xylene) Concentrations in Delhi for the Year 2019 from January to November

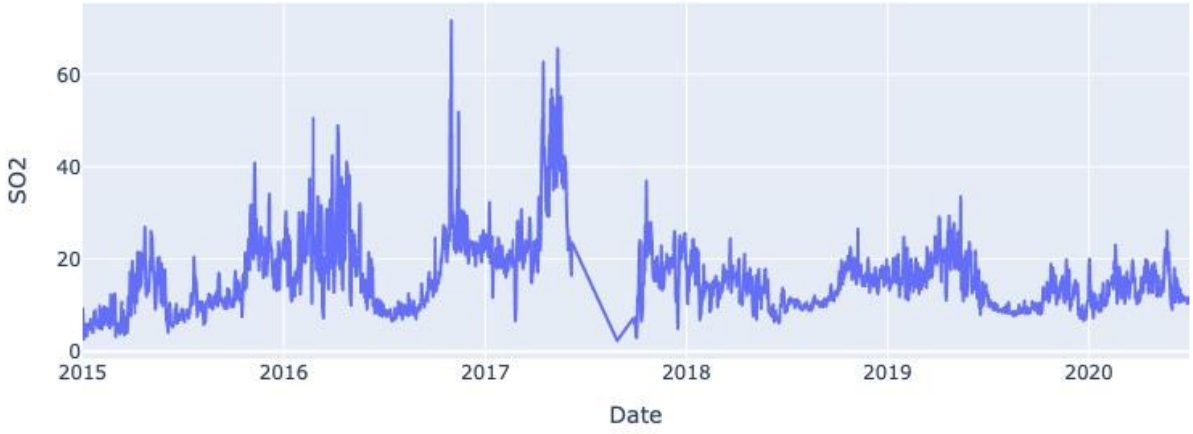


Figure 9: Time Series of SO2 (Sulphur Dioxide) Concentrations in Delhi from 2015 to 2020

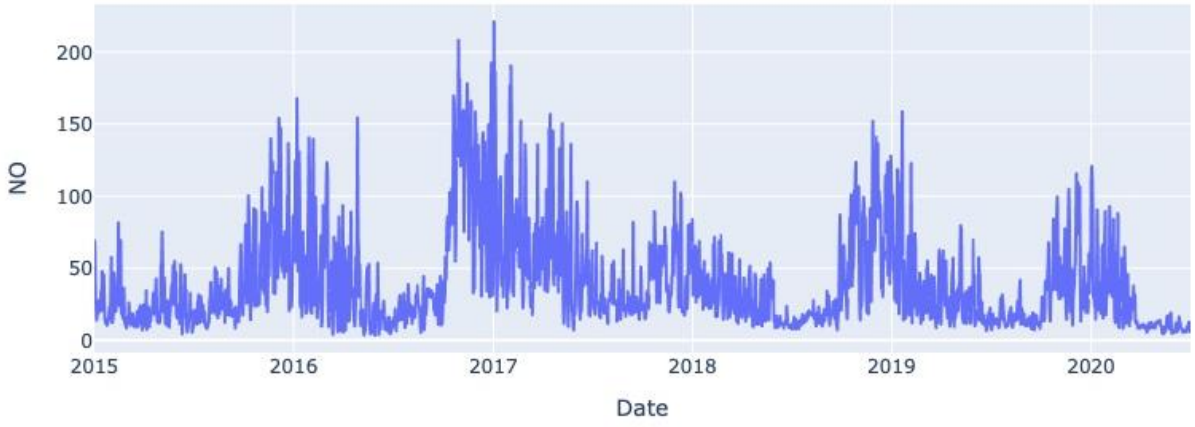


Figure 10: Time Series of NO (Nitric Oxide) Concentrations in Delhi from 2015 to 2020

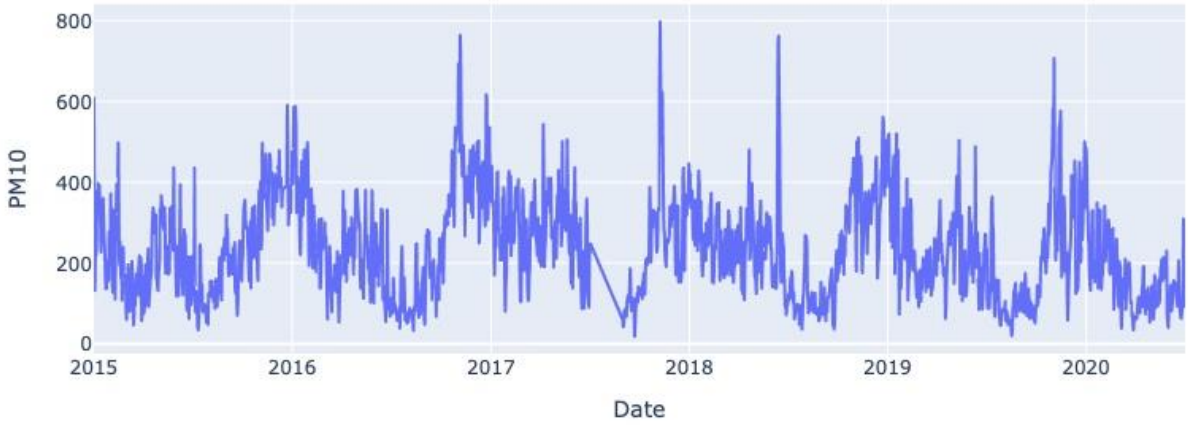


Figure 11: Time Series of PM10 (Particulate Matter 10 micrometres) concentrations in Delhi from 2015 to 2020

Figure 12: Comparative time series of AQI levels in Delhi from March 23rd to 15th April (2019 vs 2020)

Figure 13: Comparative time series of SO2 levels in Delhi from March 23rd to 15th April (2019 vs 2020)

Figure 14: Comparative time Series of O3 concentrations in Delhi from March 23rd to 15th April (2019 vs 2020)

Figure 15: Comparative time series of NO concentrations in Delhi from March 23rd to 15th April (2019 vs 2020)

Figure 16: Comparative time series of PM10 concentrations in Delhi from March 3rd to 28th April (2019 vs 2020)

Figure 17: Comparative time series of PM2.5 concentrations in Delhi from March 3rd to 28th April (2019 vs 2020)

Figure 18: Comparative time series of NO2 concentrations in Delhi from March 23rd to 15th April (2019 vs 2020)

PART B- Heart disease Dataset

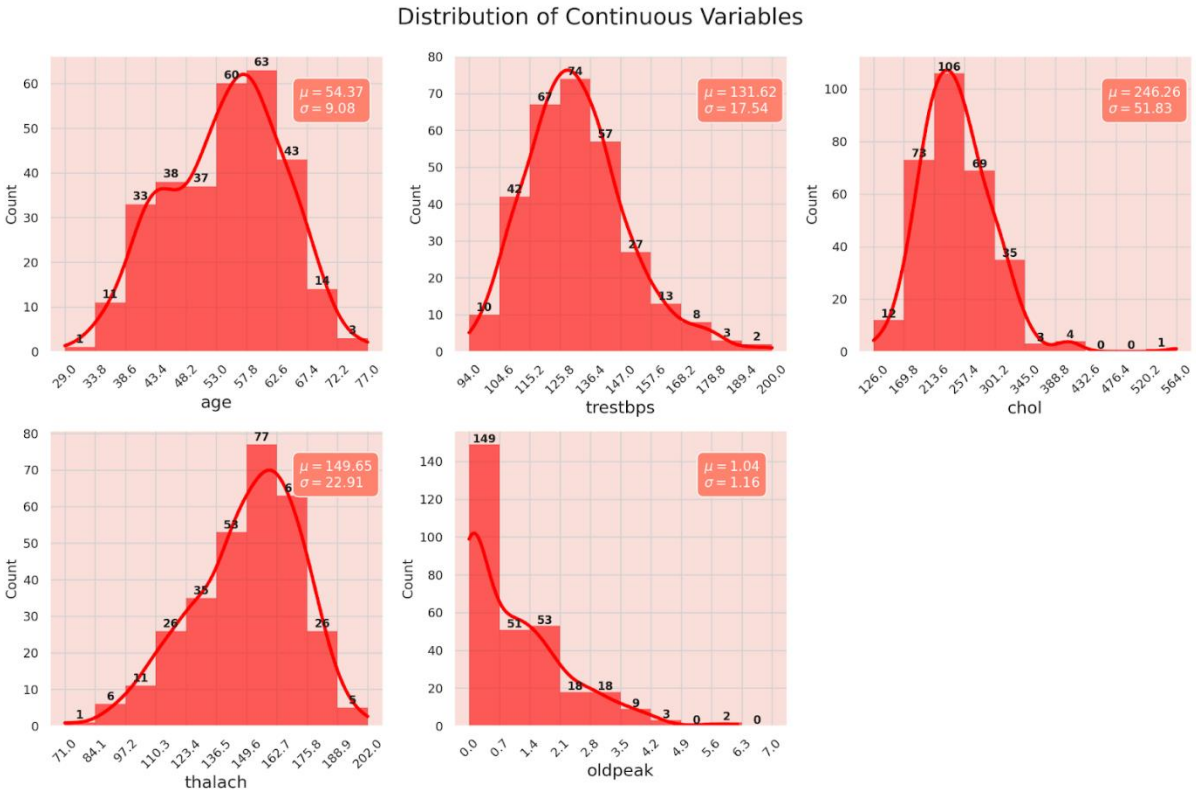


Figure 19: Univariate Analysis of continuous variables using histograms

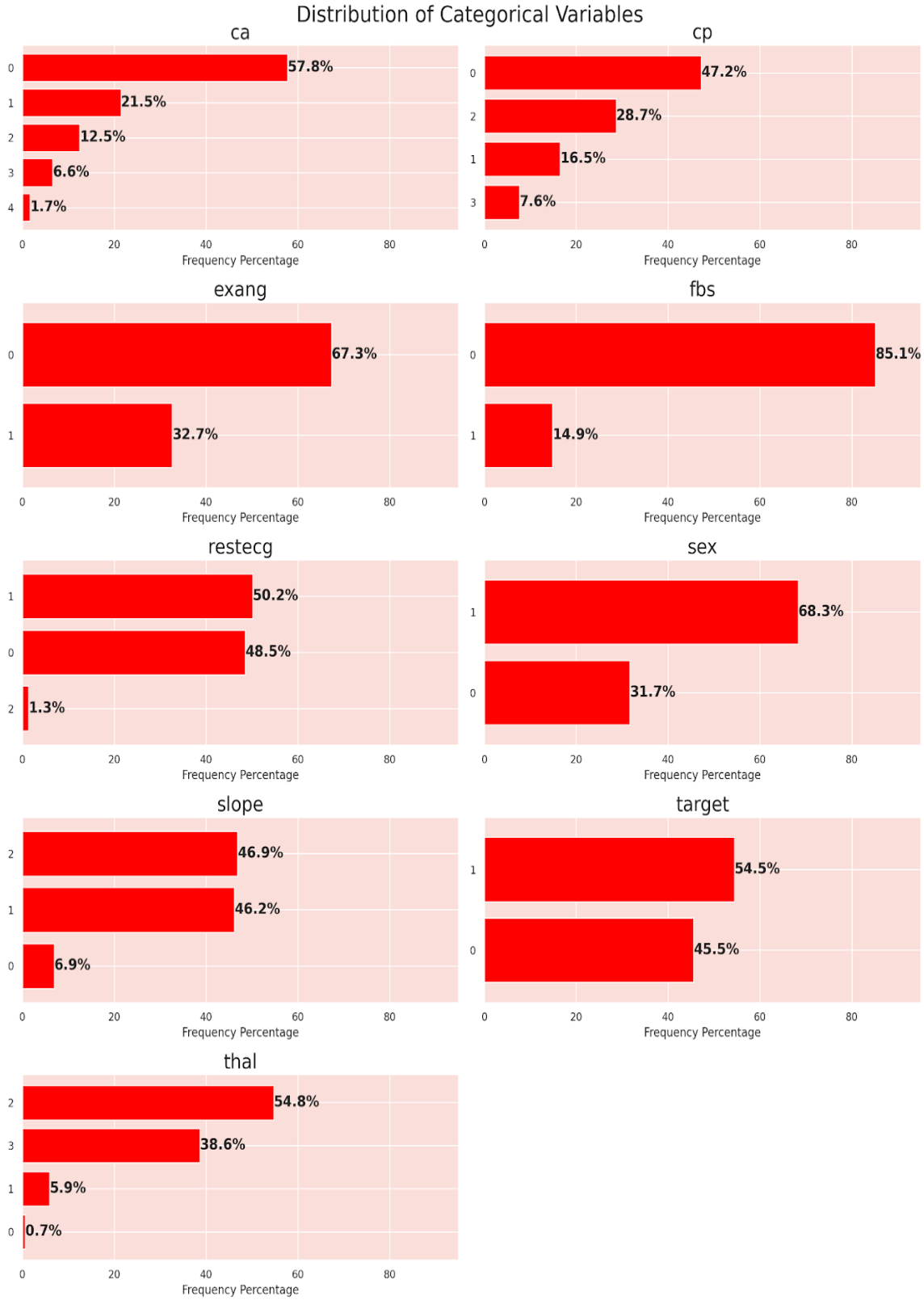


Figure 20: Univariate Analysis of categorical variable using horizontal bar plots

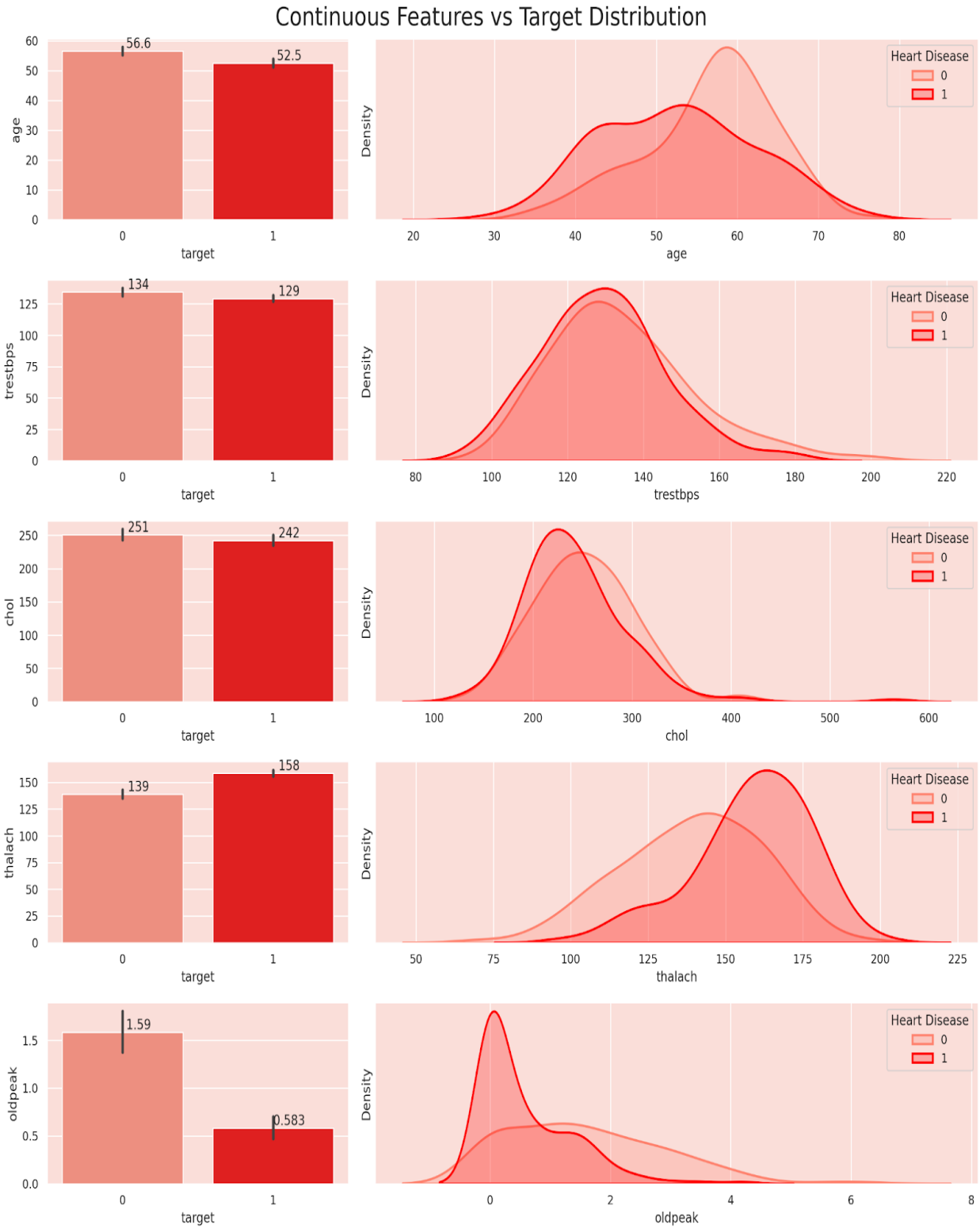


Figure 21: Bivariate Analysis of continuous features

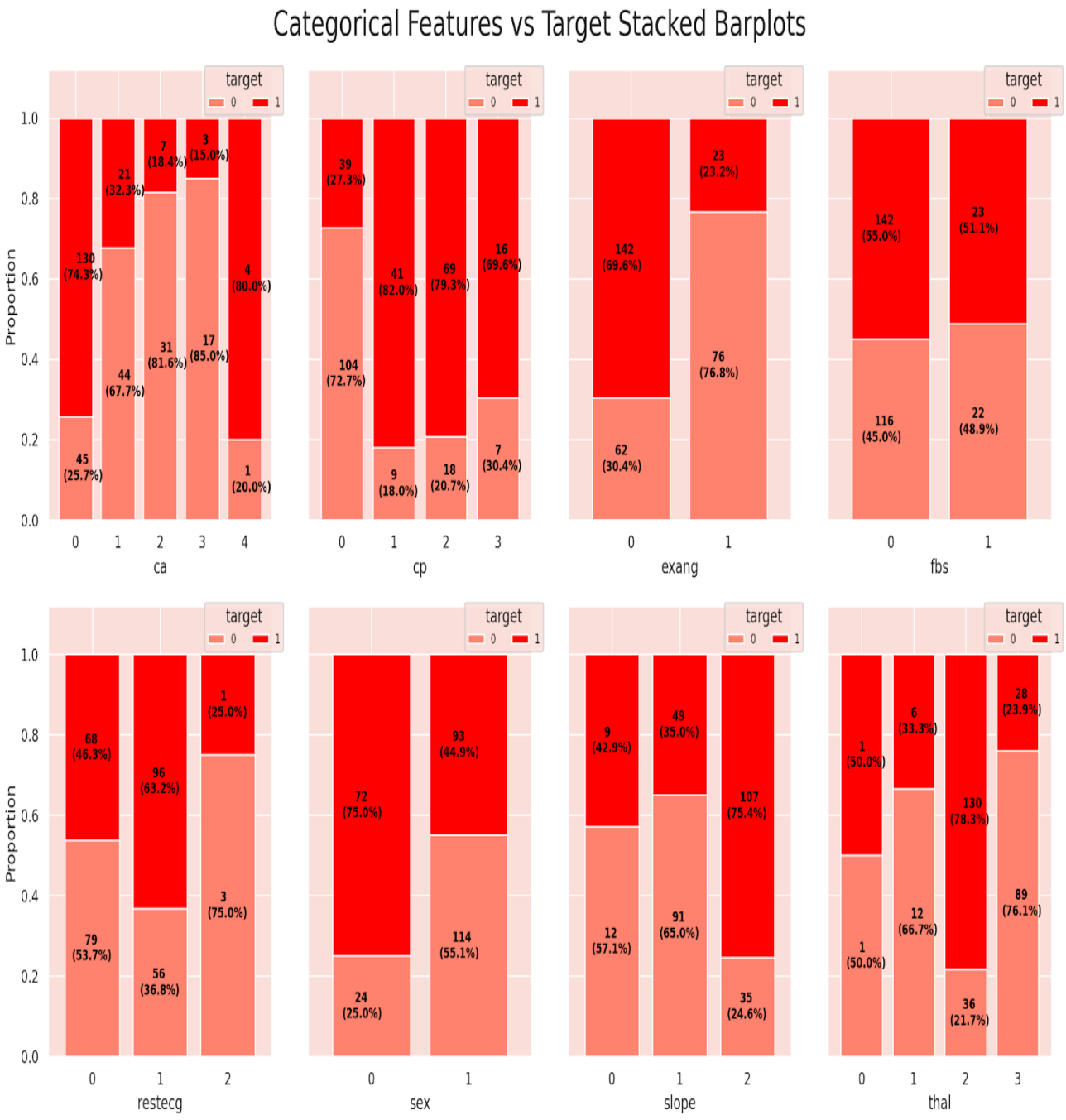


Figure 22: Bivariate Analysis of categorical Features using stacked bar plots

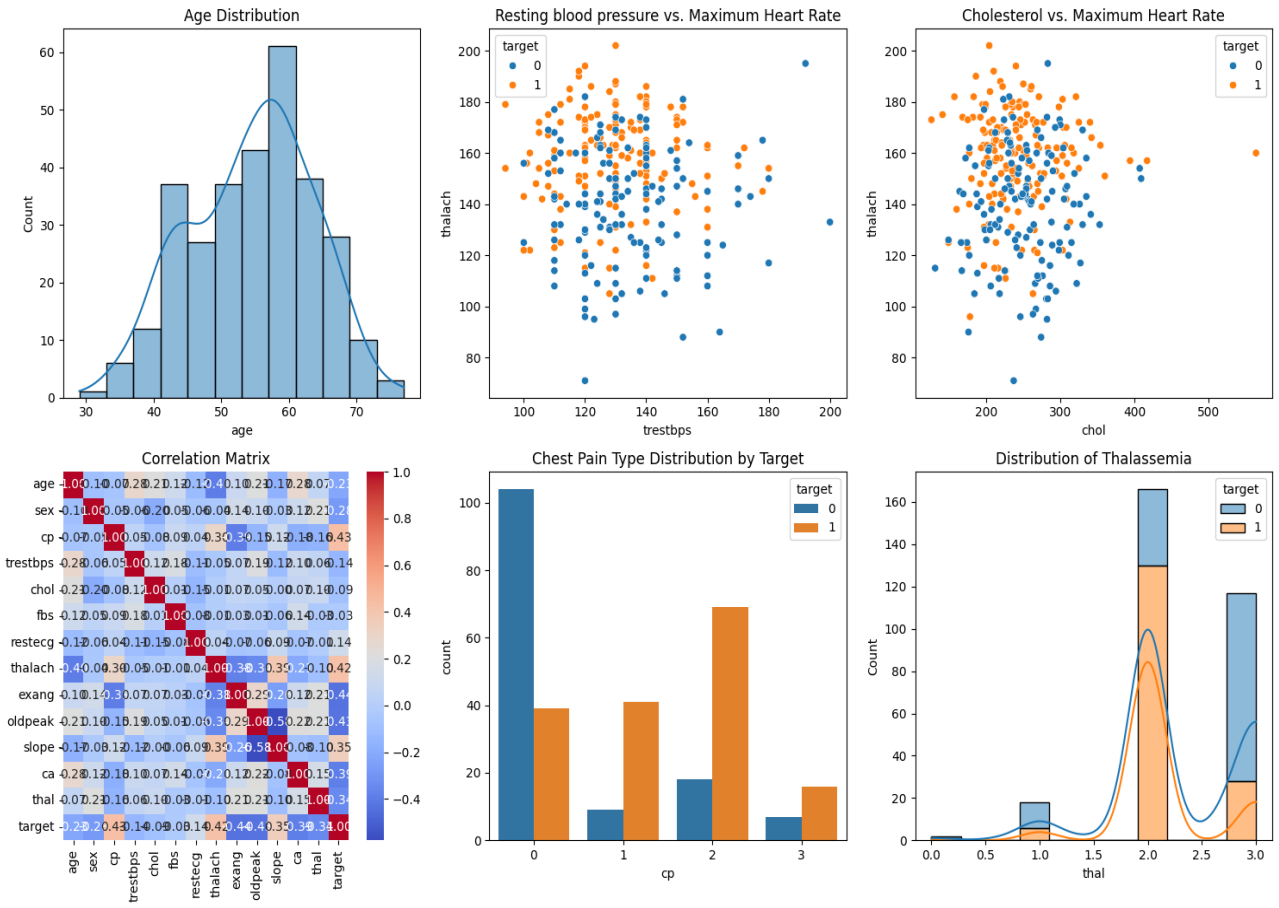


Figure 23: *(from left to right)*

(a) Age distribution: histogram showing the distribution of patient ages with a density curve

(b) Resting Blood Pressure vs Maximum Heart Rate: Scatter Plot showing the relationship between these two variables, color-coded by target (0 = no heart disease, 1 = heart disease)

(c) Cholesterol vs Maximum Heart Rate: Scatter Plot showing relationship between the two, color coded by target

(d) Correlation Matrix: heatmap showing correlation coefficients between key features in the dataset

(e) Chest Pain type distribution by target: Bar plot showing the frequency of different chest pain types based on target class

(f) Distribution of thalassemia by target: Histogram showing distribution of thalassemia values in the dataset for each target class

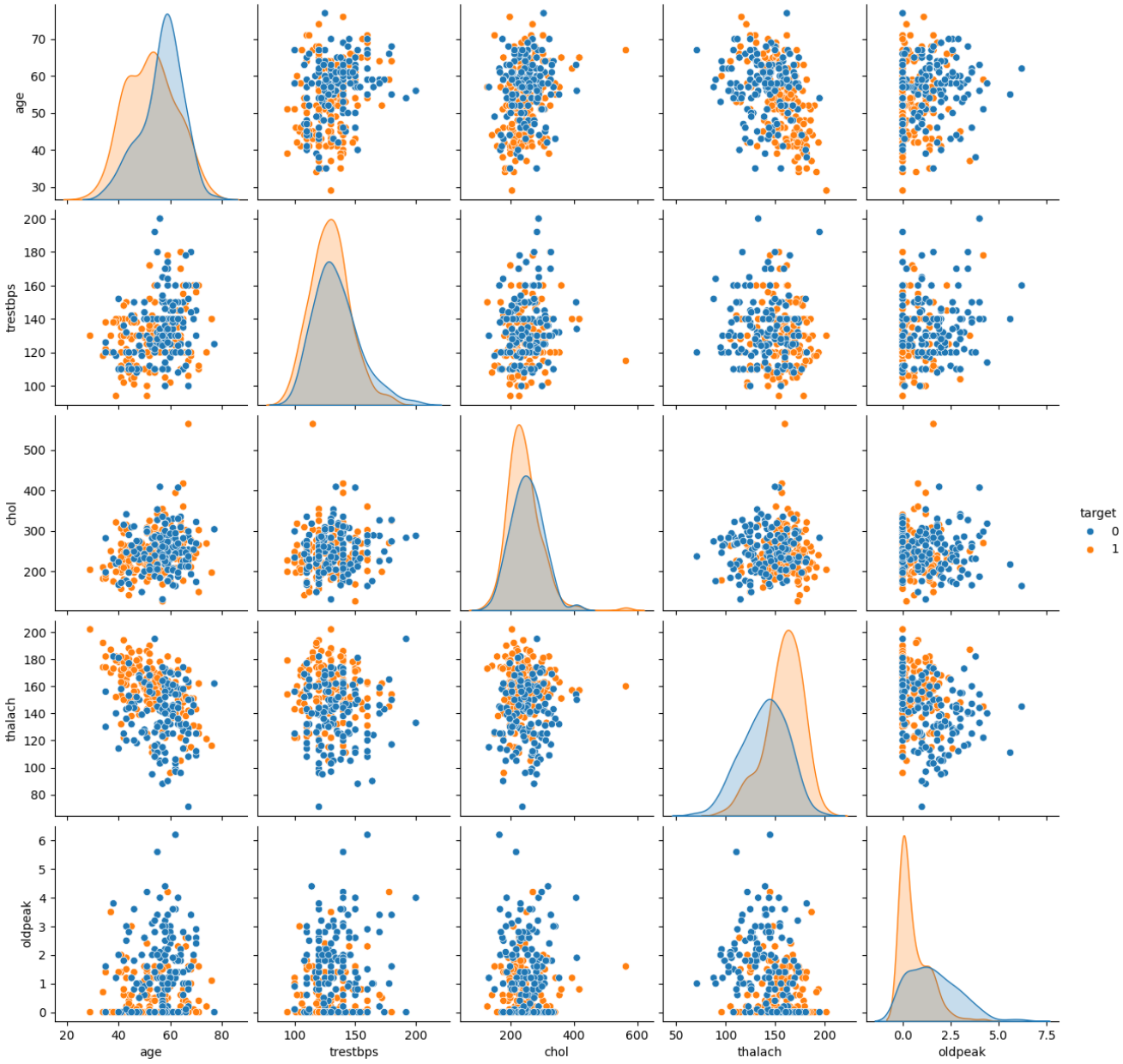
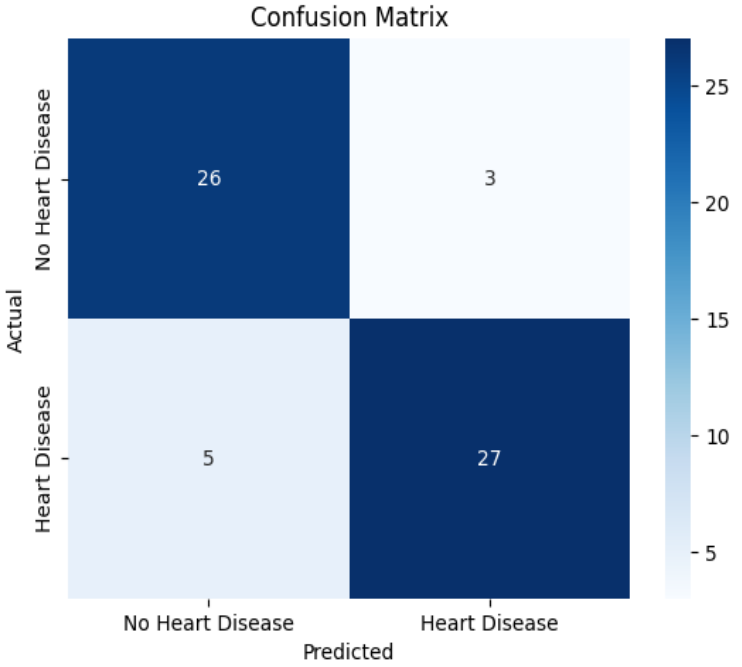
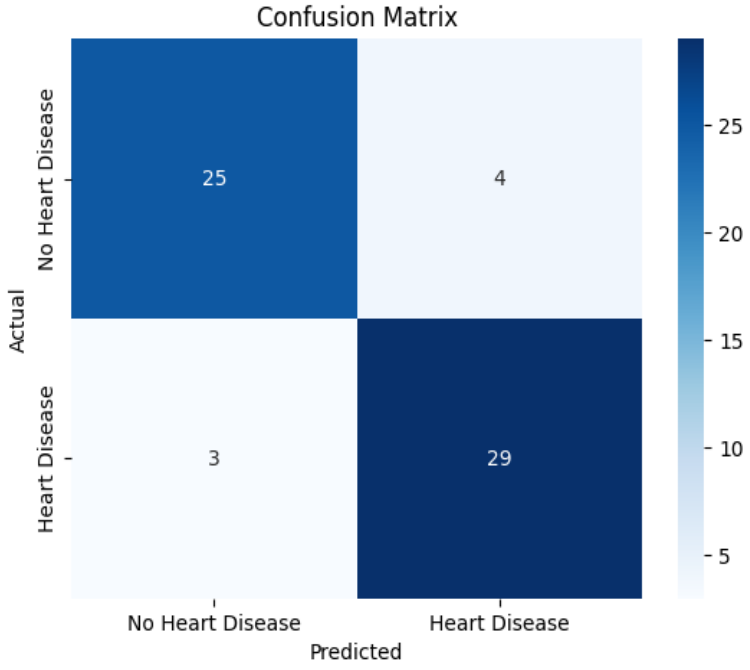


Figure 24: Pair Plot of key features in relation to target

- Diagonal containing density plots (kernel density plots) showing the distribution of individual features for each target class (0 and 1)
- Off-Diagonal containing scatter plots displaying pairwise relationships between different features with data points color coded by target class (0 = No heart disease, 1 = Heart disease)
- Features in plot: Age, trestbps - resting blood pressure, chol - cholesterol level, thalach - Maximum heart rate achieved, oldpeak - ST depression induced by exercise relative to rest



a)



b)

Figure 25: Confusion Matrix Models

(a) Confusion Matrix Model 1: Visualisation showing the distribution of true positives, true negatives, false positives, and false negatives for the first model, predicting heart disease occurrence

(b) Confusion Matrix for Model 2: Similar to Model 1, shows accuracy and misclassification rates for the second model in predicting heart disease.

VI. RESULTS AND DISCUSSION

Data visualisation was performed for the two datasets.

In case study 1, we mainly dealt with numerical variables thus we visualised the data using various line graphs while in case study 2, we dealt with both numerical and categorical variables thus we visualised the data using line graphs, histograms, bar graph, scatter plot, correlation matrix, pair plots and confusion matrix.

a) For case study 1 (air pollution dataset), we plotted various line graphs to depict the relationship between different numerical variables. We used time series analysis to show how different metrics like AQI, PM10, PM2.5, SO2, BTX, O3, NO, NO2 have varied over a period of time in Delhi.

b) For case study 2 (heart disease dataset), we used the following charts to show the relationship between different metrics. We used :-

- Line graphs and histograms to depict the relationship between different numerical variables.
- Bar graph and scatter plot to show the relationship between 2 different numerical variables. In addition to that it is also used to show the distribution of the target variables (categorical) for the 2 variables plotted on the axes.
- Correlation matrix to identify patterns between different sets of variables in the data set which helps in analysing relationships and also improves model performance.
- Pair plots to visualise the patterns between each pair of numerical variables in the dataset. It helps to combine scatter plots to show correlation between data points. In addition to this, the plots also show the distribution of the target variable over the data.
- Confusion matrix to evaluate the ML model's performance. It also helps to provide insights into a model's accuracy, recall, precision and F1 score.

Accuracy of models was also assessed.

a) We used different models like SVM, Decision trees, Random Forest, Gaussian NB (Naive Bayes) KNN and logistic regression.
b) Using the above models, we calculated the accuracy of different models and we found that in the heart disease dataset, accuracy of SVM was the highest.
c) Below we have tabularised the accuracy results obtained for different models for the two case studies.

Model used	Random Forest	Decision Tree	Support Vector Machine	k-Nearest Neighbour	Gaussian Naive Bayes	Logistic Regression
Accuracy	88%	85%	97%	85%	86%	88%

d) Below we have graphically represented the accuracy results of different models.

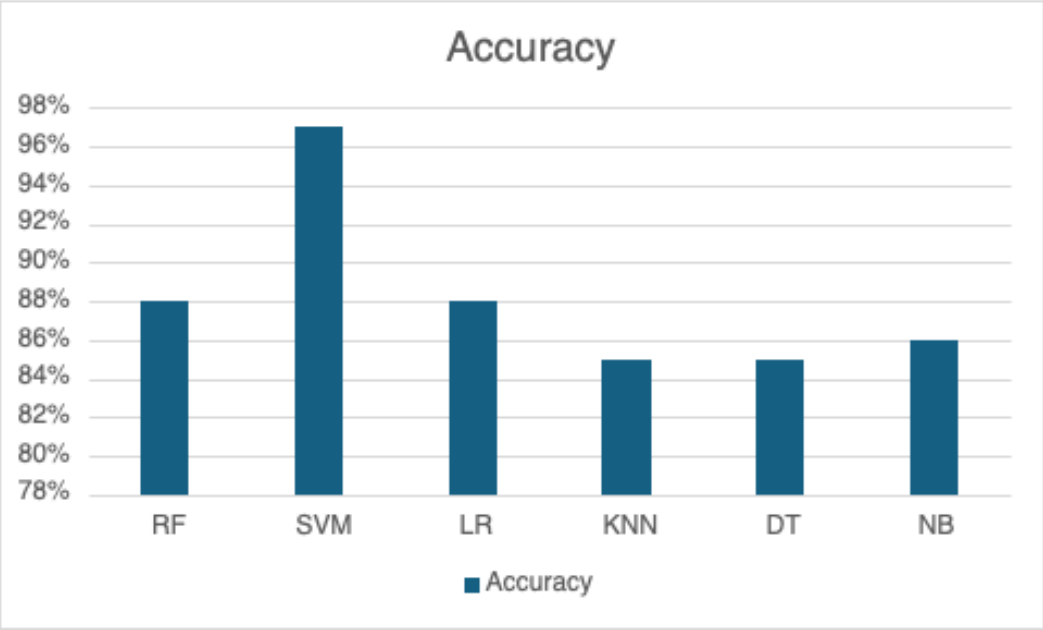


Figure 26: Accuracy of different models

VII. CONCLUSION

Artificial Intelligence and Machine Learning algorithms are revolutionising the science of data processing. Through our study, we were able to assess the efficacy and accuracy of four ML algorithms for the given datasets. The findings of the result helped us to understand how the scope of analytics has expanded through diverse fields such as healthcare and environmental studies. Moreover, the role of literature review served as a foundation for identifying the emerging trends and understanding the field of data analytics.

In conclusion, data analysis serves as a critical tool for extracting information from vast and complex datasets. The scope of analytics is expected to grow even further with integration of technologies such as quantum computing, edge computing and real-time data processing. These advancements will quickly enable more precise analysis and allow for new levels of predictive and prescriptive analytics along with the rise of automation and augmented analytics, overall shaping the future of this ever-evolving field.

VIII. REFERENCES

1. Staff, C. (2024, April 19). *Data Analytics: Definition, Uses, Examples, and More*. Coursera. https://www.coursera.org/articles/data-analytics
2. Yasar, K., & Stedman, C. (2024, April 2). *data analytics (DA)*. Data Management. https://www.techtarget.com/searchdatamanagement/definition/data-analytics
3. GeeksforGeeks. (2024, September 3). *Machine Learning Algorithms*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning-algorithms/
4. *Machine Learning Algorithms - Javatpoint*. (n.d.). www.javatpoint.com. https://www.javatpoint.com/machine-learning-algorithms
5. Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR), 9*(1), 381–386. https://doi.org/10.21275/art20203995
6. Simplilearn. (2023, November 7). *Random Forest Algorithm*. Simplilearn.com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm
7. GeeksforGeeks. (2024b, September 4). *Random Forest Regression in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-regression-in-python/
8. Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees *International Journal of Computer Science Issues (IJCSI). 9.*
9. Cohen, S. (2020). The basics of machine learning: strategies and techniques. In *Elsevier eBooks* (pp. 13–40). https://doi.org/10.1016/b978-0-323-67538-3.00002-6
10. *Decision Tree Algorithm in Machine Learning - Javatpoint*. (n.d.) www.javatpoint.com. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
11. Kartik, Asooja., Georgeta, Bordea., Gabriela, Vulcu., Paul, Buitelaar. (2016). *Forecasting emerging trends from scientific literature. 417-420.*
12. Kalyan, Nagaraj., G., S., Sharvani., Amulyashree, Sridhar. (2018). Emerging trend of big data analytics in bioinformatics: a literature review. *International Journal of Bioinformatics Research and Applications, 14:144-205*. doi: 10.1504/IJBRA.2018.10009206
13. Taylor, P. Volume of Data/Information Created, Captured, Copied, and Consumed Worldwide from 2010 to 2020, with Forecasts from 2021 to 2025. Available online: https://www.statista.com/statistics/871513/worldwide-data-created/
14. Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(0), 2. https://doi.org/10.5334/dsj-2015-002

15. Medida, Lakshmi & Kumar, G.L.N.V.s. (2024). Addressing Challenges in Data Analytics: A Comprehensive Review and Proposed Solutions. 10.4018/979-8-3693-2260-4.ch002.

16. Brownlee, J. (2020). *A gentle introduction to k-nearest neighbors algorithm (k-NN)*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com

17. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

18. Brownlee, J. (2019). *Support vector machines (SVM) for machine learning*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com

19. Zinke, A. (2024, July 29). *23 Big Data Trends - Hexagon's Manufacturing Intelligence Blog*. Hexagon's Manufacturing Intelligence Blog. https://blog.hexagonmi.com/23-big-data-trends/

20. Bernstein, C. (2022, December 29). *financial analytics*. ERP. https://www.techtarget.com/searcherp/definition/financial-analytics

21. Appsierra Digital Engineering Services. (n.d.). *Role and importance of data analytics in telecom industry*. https://www.appsierra.com/blog/data-analytics-in-telecom-industry

22. University of San Diego Online Degrees. (2023, November 8). *What is Health Care Analytics?* https://onlinedegrees.sandiego.edu/what-is-health-care-analytics/#1

23. Ilin, I., Klimin, A., Shaban, A., & Peter the Great St.Petersburg Polytechnic University. (2019). Features of big data approach and new opportunities of BI-systems in marketing activities. In *E3S Web of Conferences* (Vols. 110–110, pp. 1100–2018). EDP Sciences. https://doi.org/10.1051/e3sconf/201911002054

24. Yashvi. (2020, September 16). Air quality analysis of delhi. Kaggle. https://www.kaggle.com/code/yashvi/air-quality-analysis-of-delhi

25. Farzadnekouei. (2023, August 11). Heart Disease Prediction. Kaggle. https://www.kaggle.com/code/farzadnekouei/heart-disease-prediction

26. Kothari, S. (2024, June 4). Applications of Data Analytics: Real-world Applications and Impact. Simplilearn.com. https://www.simplilearn.com/tutorials/data-analytics-tutorial/applications-of-data-analytics

27. What is predictive analytics and how does it work? | Google Cloud. (n.d.). Google Cloud. https://cloud.google.com/learn/what-is-predictive-analytics

28. Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-54807-1

29. Srinivasa Gupta, N., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. In Hindawi, Journal of Environmental and

Public Health (Vol. 2023, pp. 1–26) [Journal-article]. https://doi.org/10.1155/2023/4916267

30. Aram, S. A., Nketiah, E. A., Saalidong, B. M., Wang, H., Afitiri, A., Akoto, A. B., & Lartey, P. O. (2023). Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. International Journal of Environmental Science and Technology, 21(2), 1345–1360. https://doi.org/10.1007/s13762-023-05016-2

31. Edwards, T. O. a. J. (2024b, April 13). What is predictive analytics? Transforming data into future insights. CIO. https://www.cio.com/article/228901/what-is-predictive-analytics-transforming-data-into-future-insights.html

32. Segner, M. (2024, February 12). The Future Of Big Data Analytics & Data Science: 6 Trends Of Tomorrow. Monte Carlo Data. https://www.montecarlodata.com/blog-the-future-of-big-data-analytics-and-data-science/

33. The Future of Data Analytics: Trends of Tomorrow. (2024, January 18). https://www.knowledgehut.com/blog/data-science/data-analytics-future