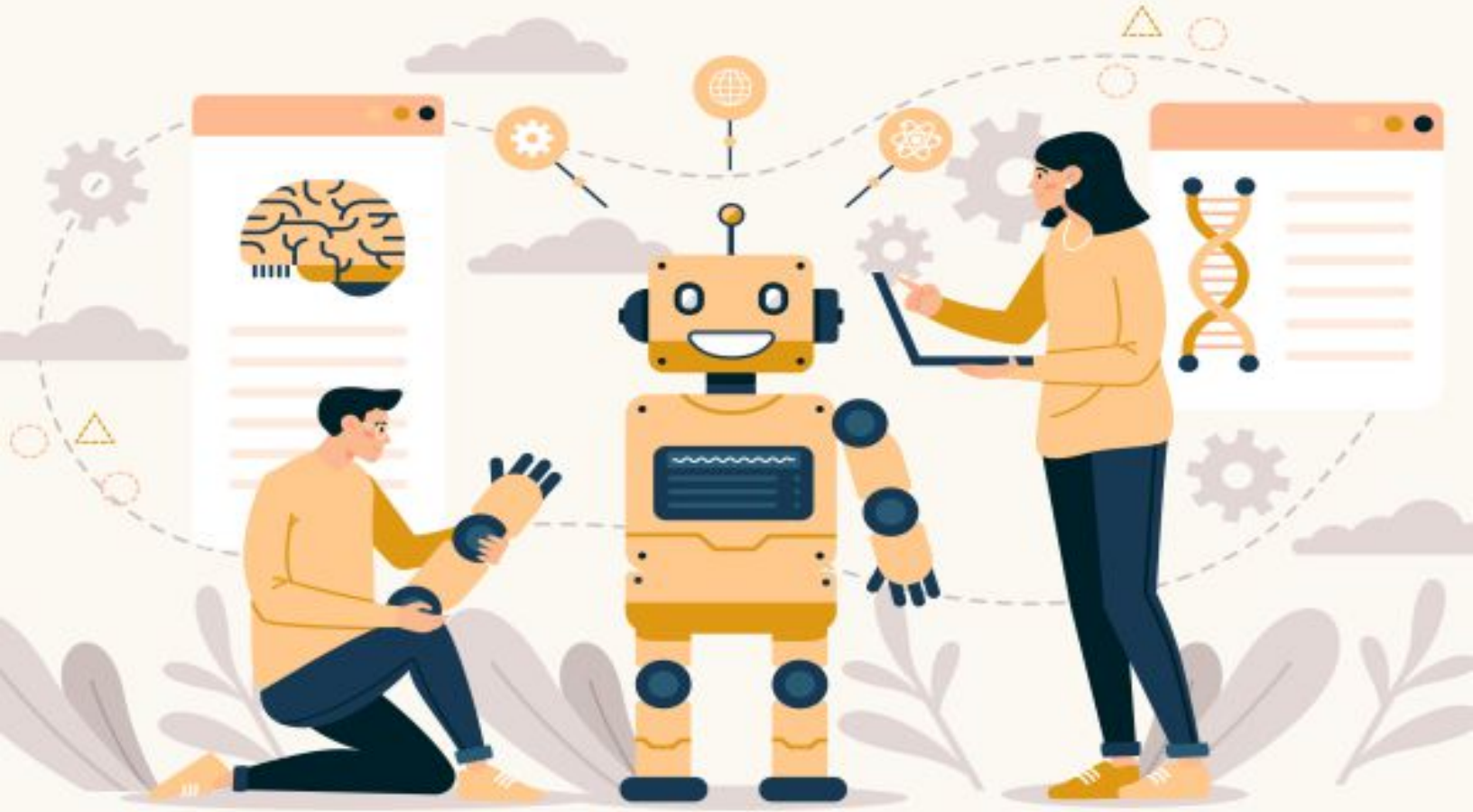
Fun with Machine Learning

Simplify the Data Science process by automating repetitive and complex tasks using AutoML



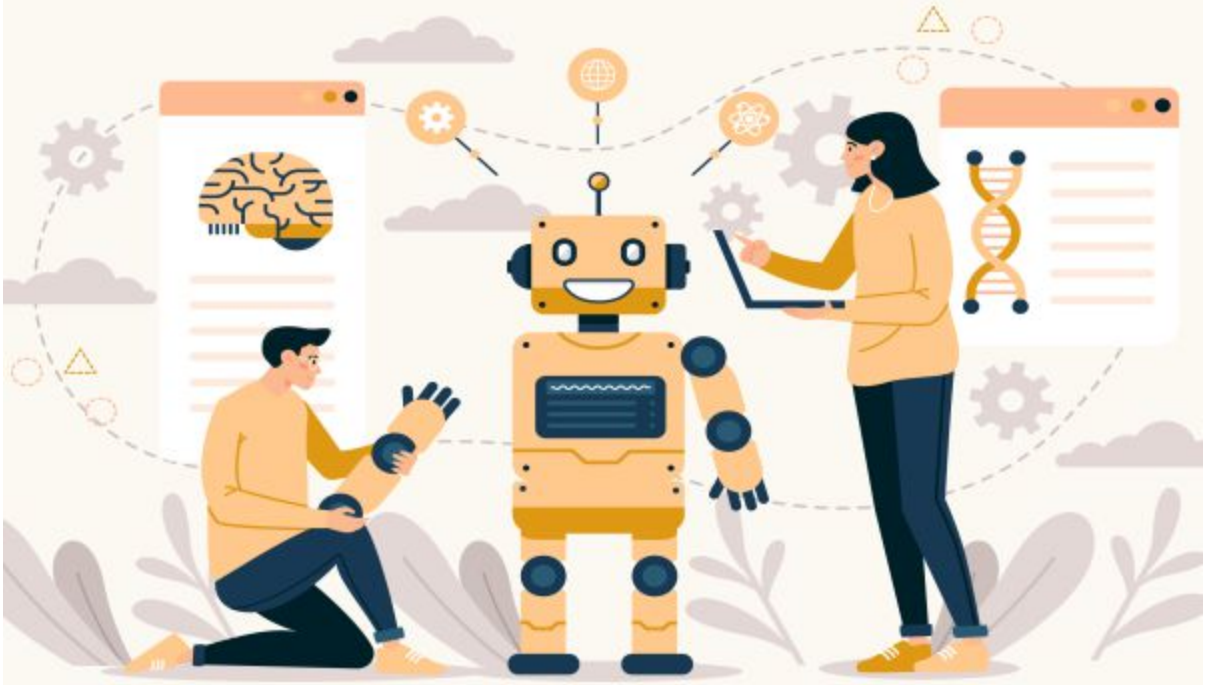
Arockia Liborious

Dr. Rik Das

bpb

Fun with Machine Learning

Simplify the Data Science process by automating repetitive and complex tasks using AutoML



Arockia Liborious
Dr. Rik Das

bpb

Fun with Machine Learning

Simplify the Data Science process by automating repetitive and complex tasks using AutoML

Arockia Liborious

Dr. Rik Das

www.bpbonline.com

First published: 2023

Published by BPB Online

WeWork

119 Marylebone Road

London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-5551-784-5

www.bpbonline.com

Dedicated to

Arockia Liborious

Dedicated to the light of my life, my Wife: Gargi
My pillars of support: Mom: Mary, and Brother: Camillus

Dr. Rik Das

Dedicated to my parents: Mr. Kamal Kumar Das and
Mrs. Malabika Das
My wife: Simi
My kids: Sohan and Dikshan

About the Authors

- **Arockia Liborious** stands out as a celebrated leader in analytics, having amassed over a dozen accolades for his work in machine learning over the past 12 years. He is a seasoned professional who has conceptualized and implemented machine learning solutions for businesses across different domains. His expertise includes computer vision and natural language processing, which enables him to provide insights into emerging technologies and their impact on businesses. Arockia has a unique blend of technical and business expertise, which allows him to evaluate the long-term return on investment (eROI) of machine learning solutions, as well as their strategic positioning within the market. He has successfully led several initiatives to develop innovative service and product strategies that leverage machine learning to solve complex business problems.

  Throughout his career, Arockia has worked with clients in diverse industries, including Manufacturing, Fintech, Banking, Food and Beverage, and large Energy clients. His experience has equipped him with the ability to understand and cater to the unique needs of each industry, enabling him to deliver custom-tailored solutions that drive business results. Apart from his work in machine learning, Arockia is also an expert in data analytics, predictive modeling, feature engineering, and data visualization. He is passionate about using data to drive informed decision-making and believes that businesses that leverage the power of data will be the ones that succeed in today's competitive market. Arockia holds a master's degree in data science and is a sought-after speaker at industry events and conferences. He has received numerous awards throughout his career, recognizing his contributions to the field of analytics.
- **Dr. Rik Das** is a Lead Software Engineer (Artificial Intelligence Unit) at Siemens Technology and Services Pvt. Ltd. (Siemens Advanta), Bengaluru. He is a thought leader with over 18 years of experience in Industrial and Academic Research. Rik has been a seasoned

academician before joining the corporate and has served as a Professor in multiple prestigious Universities, Institutions and EduTechs of the country.

Dr. Das is conferred with the title of ACM Distinguished Speaker by the Association for Computing Machinery (ACM), New York, USA. He is also a Member of International Advisory Committee of AI-Forum, UK. Dr. Das is a Ph.D. (Tech.) in Information Technology from University of Calcutta. He has also received his M.Tech. (Information Technology) from University of Calcutta after his B.E. (Information Technology) from University of Burdwan. As an innovation leader, Rik has carried out collaborative research in multiple domains. He has filed and published two Indian patents consecutively during the year 2018 and 2019 and has over 80 International publications till date with reputed publishers He has published 10 books on applications of AI/ML and Computer Vision. He has also chaired multiple sessions in International Conferences on Machine Learning and has acted as a resource person / invited speaker in multiple events and refresher courses on Information Technology. Dr. Rik Das is recipient of multiple prestigious awards due to his significant contributions in the domain of Technology Consulting, Research Innovations and Academic Excellence. He is awarded with Warner von Siemens Country Award 2021 followed by Divergent category award 2022 consecutively during his current employment at Siemens Advanta. He has received “Best Innovation Award” in Computer Science category at UILA Awards 2021. He was featured in uLektz Wall of Fame as one of the “Top 50 Tech Savvy Academicians in Higher Education across India” for the year 2019. His keen interest towards application of machine learning and deep learning techniques for real life use case solutions has resulted in joint publications of research articles with Professors and Researchers from various reputed Multinationals brands namely Philips-Canada, Cognizant Technology Solutions, TCS, etc. and International Universities including College of Medicine, University of Saskatchewan, Canada, Faculty of Electrical Engineering and Computer Science, VSB Technical University of Ostrava, Ostrava, Czechia, Cairo University, Giza,

Egypt and so on. Dr. Rik Das is always open to discuss new research project ideas for collaborative work and for techno-managerial consultancies.

About the Reviewers

Peter Henstock is the Machine Learning & AI Lead at Pfizer. His work has focused on the intersection of AI, visualization, statistics, and software engineering. At Pfizer, he has been mostly developing solutions for the drug discovery area but has more recently focused on clinical trials. Prior to Pfizer, he worked at MIT Lincoln Laboratory in computational linguistics and image analysis. Peter holds a Ph.D. in Artificial Intelligence from Purdue University and 7 Master's degrees including an MBA. The Deep Knowledge Analytics group recognized him as among the top 12 leaders in AI and Pharma globally. He also currently teaches two graduate-level courses at Harvard: "Advanced Machine Learning, Data Mining and AI" and the Software Engineering capstone course.

Sheena Siddiqui is a machine learning engineer, working with one of the leading global organizations. She has 5 years of work experience in diverse AI and ML technologies, where she has made her mark in the top 1% of all employees. She has authored several introductory and advanced-level online courses and conducted webinars for learners in her field. She is a postgraduate in Electrical Engineering from Jamia Millia Islamia. In addition to AI and ML, her area of interest includes Quantum Computing and Sustainability Research. She also works as a video editor, graphic designer, and technical writer.

Acknowledgements

We would like to express our deepest gratitude to the many people who have supported us throughout the process of writing this book. First and foremost, we want to thank our family, who have been unwavering in their love and encouragement. Their belief in us and our abilities has been the foundation of our success.

We would also like to thank our friends, who have been a constant source of inspiration and support. Their feedback and encouragement have been invaluable in shaping this book.

We also want to thank BPB Publications for their consideration in publishing this book. Their guidance and suggestions have been instrumental in shaping this work, and we are deeply grateful for their support.

In writing this book, we have been privileged to work with some truly amazing people, and we are grateful to every one of them for their contributions. Thank you all for your support, encouragement, and belief in us.

Preface

Welcome to “Fun with Data Science” - a book that aims to make data-driven decision making accessible and easy for everyone. In today’s world, data is everywhere and plays a critical role in every aspect of business. However, not everyone has the skills and expertise to use it effectively. That’s where this book comes in.

With the help of auto ML tools, we can make data-driven decision making easier for everyone, even those who are not data scientists. This book is designed to help organizations move from intuition-driven decision making to a more data-driven approach. It provides a step-by-step guide to using auto ML tools to solve business problems and make data-backed decisions.

This book is for anyone who wants to take data-backed decisions but does not know where to start. Whether you are a business leader, a manager, or a professional from any field, this book will help you understand the basics of data science and how you can use it to drive your organization’s success.

We hope that you find this book informative, engaging, and most importantly, fun. Let’s dive in and explore the world of data science together!

Chapter 1: Significance of Machine Learning in Today’s Business- The first chapter of the book emphasizes the growing importance of machine learning in modern business. It explains how machine learning is used to automate business operations, make data-driven decisions, and improve customer experience. The chapter also provides an overview of the different types of machine learning algorithms and their applications, such as predictive analytics, natural language processing, and computer vision. It highlights the potential benefits of machine learning, including increased efficiency, reduced costs, and improved accuracy.

Chapter 2: Know your Data- The second chapter of the book focuses on the importance of data in machine learning. It explains the different types of data, including structured, unstructured, and semi-structured data. The chapter also introduces the concept of “dark data,” which refers to the vast

amounts of unutilized data that organizations possess. It highlights the importance of collecting and analyzing data to derive valuable insights that can drive business decisions. The chapter emphasizes the need for high-quality, trusted data to ensure accurate results in machine learning applications. It concludes by emphasizing the significance of data governance, which ensures that data is collected, managed, and used in an ethical and compliant manner.

Chapter 3: Up and Running with Analytical Tools- The third chapter of the book focuses on how to get started with analytical tools for addressing business issues quickly. It explains the different data analytics approaches that can be used and emphasizes the importance of predictive modeling. The chapter provides information on how to perform predictive modeling without any prior coding expertise and highlights how data cleansing and visualization can be automated using open-source and commercial solutions. The chapter also discusses various analytical tools, including Excel, Tableau, KNIME, Weka, Rapid Miner, Orange, and many others. It highlights how these tools can be used to analyze data and gain valuable insights, helping businesses to make informed decisions.

Chapter 4: Machine Learning in a nutshell- The fourth chapter of the book focuses on how machine learning can be used to solve business problems and anticipate future problems. The chapter emphasizes the importance of understanding the business problem well enough to select the appropriate data and machine learning algorithm that can help arrive at the right decision-making steps. It also highlights how anticipating problems before they occur can help businesses take corrective steps to eliminate or reduce their impact. The chapter emphasizes the interrelated nature of everything, from the business problem to the solution implementation process.

Chapter 5: Regression Analysis- The fifth chapter of the book focuses on different types of machine learning algorithms, with a specific emphasis on regression analysis. The chapter provides an overview of the various types of regression analysis, such as linear regression, logistic regression, and polynomial regression. It explains the nuances of regression analysis and provides insights into when to use regression analysis and what kind of business problems can be solved using it. The chapter also provides

practical examples and use cases to illustrate how regression analysis can be used to solve real-world business problems.

Chapter 6: Classification- The sixth chapter of the book focuses on classification models in machine learning. The chapter provides insights into how to specify the input and output of a classification model and how to solve both binary and multiclass classification problems. It explains how a logistic regression model and a non-linear decision tree model can be implemented to solve classification problems. The chapter also covers several assessment criteria that can be used to evaluate the performance of classification models.

Chapter 7: Clustering and Association- Clustering and Association: The seventh chapter of the book focuses on clustering and association, which are widely used techniques for discovering unknown relationships in data. The chapter provides insights into how clustering and association can serve as a starting point for individuals with minimal or no understanding of the data. Clustering can help identify similar patterns and correlations among data points, like product recommendations based on purchase history on e-commerce websites. The chapter covers different clustering algorithms and provides examples of how they can be used to identify patterns and correlations in data.

Chapter 8:Time series Forecasting- The eighth chapter of the book focuses on time series forecasting, which is a commonly used technique for making scientifically backed predictions on a time stamp basis. The chapter explains that a time series is simply a list of events in chronological order, collected over a fixed interval of time. The dimension of time adds structure and constraint to the data, making it easier to analyze and predict future trends. The chapter covers different time series forecasting techniques, including moving average, exponential smoothing, and ARIMA models, and provides examples of how these techniques can be used to make accurate predictions.

Chapter 9: Image Analysis- In the ninth chapter of the book, the focus is on image analysis and how it enables us to extract useful data from photos through image processing and computer vision. The chapter explains how recent developments in machine learning and deep learning have made it

possible to provide imaging data in near-real-time. The chapter also highlights the potential benefits of information extraction, which are far greater than most people realize. For example, image analysis has applications in healthcare, manufacturing, safety, and more, in addition to improving video surveillance.

Chapter 10: Tips and Tricks- The tenth chapter emphasizes the importance of understanding the details of data and data storytelling. Analytics can be challenging for some people, and presenting data in a narrative form can help make it more accessible to everyone. By using storytelling techniques to explain the key aspects of your analytics, you can engage your audience and help them better understand your findings. This can be especially important when presenting data to others who may not be as familiar with analytics.

Coloured Images

Please follow the link to download the
Coloured Images of the book:

https://rebrand.ly/sunkzyn

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com.** We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

Index

CHAPTER 1

Significance of Machine Learning in Today's Business

Let us be purpose-driven and empowered with data and ethics.

Everyone in today's fast-moving digital world wishes to be data-driven and wishes to create value. Do we really need to be data-driven, or can data even drive things? A million-dollar question, right? Yes indeed. Humans are emotional beings and always need a "just cause" or purpose, as often cited by leadership and management guru *Simon Sinek*. Data can unravel the mystery of whether we are progressing toward our cause but not where to go. As leaders, students, managers, and decision makers, it is imperative to use data to power our purpose. The issue of poor data analysis has plagued humanity for some time, but it has become increasingly apparent in the current era of widespread digital transformation and interconnectedness.. Everyone must know their way around data and be comfortable talking about it. This chapter will explain that the need for insights from data is stronger than ever before.

Structure

In this chapter, we will cover the following topics:

- Hype behind machine learning and data science
- Benefits of machine learning in business introducing data
- Types of data in business context
- Challenges with data
- Citizen data science
- Data science for leaders

Objectives

After studying this chapter, you should be able to relate how data plays a critical role in the business decisions we make. The chapter will also help you understand

how machine learning is helping improve the decision-making process and how to utilize data to our advantage and make decisions based on data insights.

Hype behind machine learning and data science

Imagine the time you first learnt how to add two numbers in your Mathematics class in school. For some years, you would manually add up numbers till you had to use a calculator to perform the same task. Did it mean you could not do it manually any longer? No, it was because the process of calculation had to be quickened so that you could focus on the other critical steps of the problem.

In business too, we can take decisions based on experience and suggestions from experts we trust. However, is this the right step that would help us take faster and accurate decisions in the new normal world that is heavily data-driven? So, the question is, ‘Do you want to spend your valuable time in tasks that can be automated, or do you want to spend time utilizing the data insights to make critical decisions? If you want to do the latter, then you made the right choice by getting this book.

We will help you make the best use of your time and effort to utilize data, wrangle it fast and then derive relevant insights from the data. Now let us walk through the history of machine learning and look at how it has evolved over the years. Its origin can be traced back to the 17th century, when people were trying to make sense of data and process to make quick decisions. A simple evolution chart depicting the machine learning journey is shown in *Figure 1.1*:

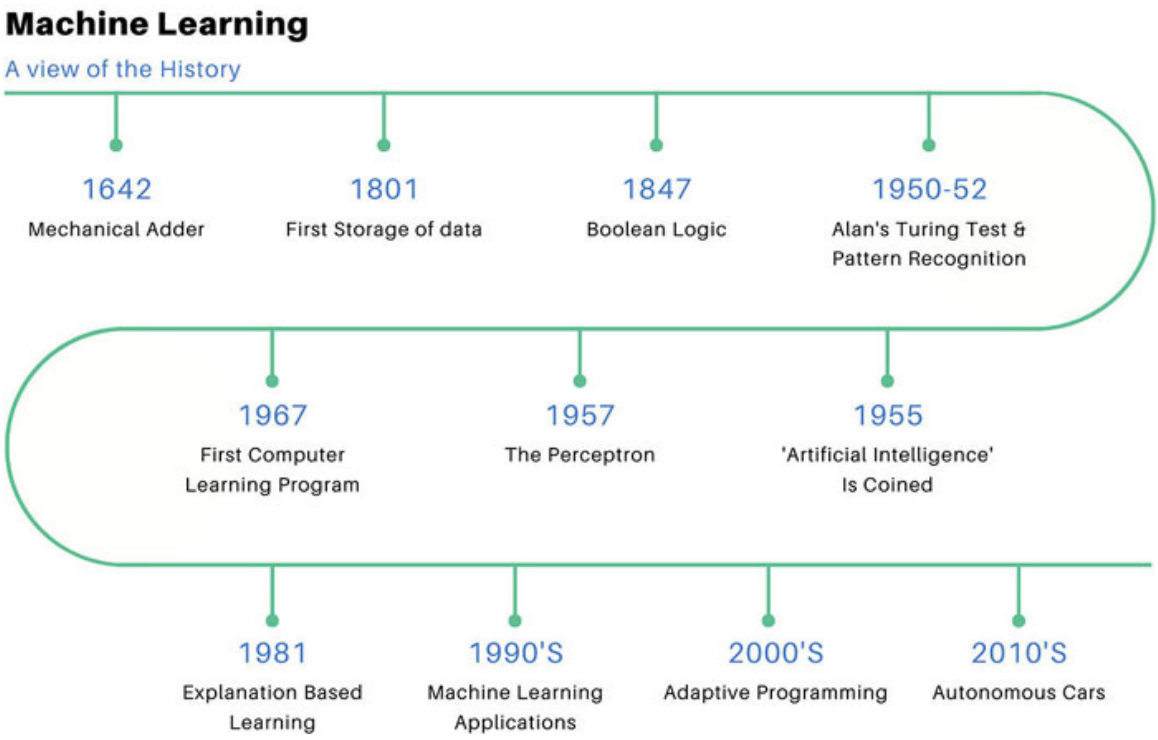


Figure 1.1: *History of Machine Learning*

Blaise Pascal created the very first mechanical adding machine in 1642. Next, the data storage challenge was overcome using a weaving loom to store data by Joseph Marie Jacquard. Over time, we developed concepts like Boolean logic, statistical calculations, and the Turing test to evaluate whether a computer had intelligence, and eventually, the phrase **artificial intelligence** (**AI**). After a series of other inventions, recent years have seen advancements in machine learning algorithms. Today, three major innovations, as listed here, have fuelled the recent buzz and helped companies and individuals use and experiment machine learning technologies. In a nutshell, they have democratized machine learning to all:

- **Higher volume of data and cheap storage**: Business-critical applications are producing and storing more data than ever before, thanks to cloud-based tools and the decreasing cost of storing data through services like Google Cloud Storage, Amazon Redshift, Microsoft Azure Services, and others. Most of these tools are highly intuitive and user friendly, with easy-to-use click and move features that simplify your work process incredibly.
- **Open-source:** Open-source machine learning libraries, such as scikit-learn, Google’s TensorFlow and Orange, make cutting-edge algorithms more usable and accessible to a larger community of data scientists and engineers.

- **Greater computing power:** With the advent of cloud-based technologies and custom hardware designed for machine learning, these systems can now run faster and at a lower cost, making them more suitable for a wide range of business needs.

Consider machine learning in this light. You, as a person and as a user of technology, carry out such actions, which allow you to make a decisive judgement and classify something. Machine learning has advanced to the point that it can mimic the pattern-matching ability of human brains. Algorithms are now used to teach machines how to recognise features of an object.

To provide just one example, a computer may be shown a cricket ball and instructed to treat it as such. The programme then uses the data to identify the different characteristics of a cricket ball, each time adding new data to the mix. Initially, a machine could identify a cricket ball as round and construct a model that states that everything round is a cricket ball. The programme then discovers that if anything is round and red, it is a cricket ball, when a red colour ball is added later. Then, a reddish-brown colour ball is introduced, and so on.

The machine must update its model as new knowledge becomes available and assign a predictive value to each model, indicating the degree of certainty that an entity is one item over another. Here, predictive value refers to the probability of identifying the ball correctly as cricket or tennis ball. As you can see in *Figure 1.2*, a machine learns the information provided to it by the user, executes certain action and receives feedback for it, and the learning continues as a feedback loop. The learning step is called as “model training”, and the feedback loop is called “model retraining”.

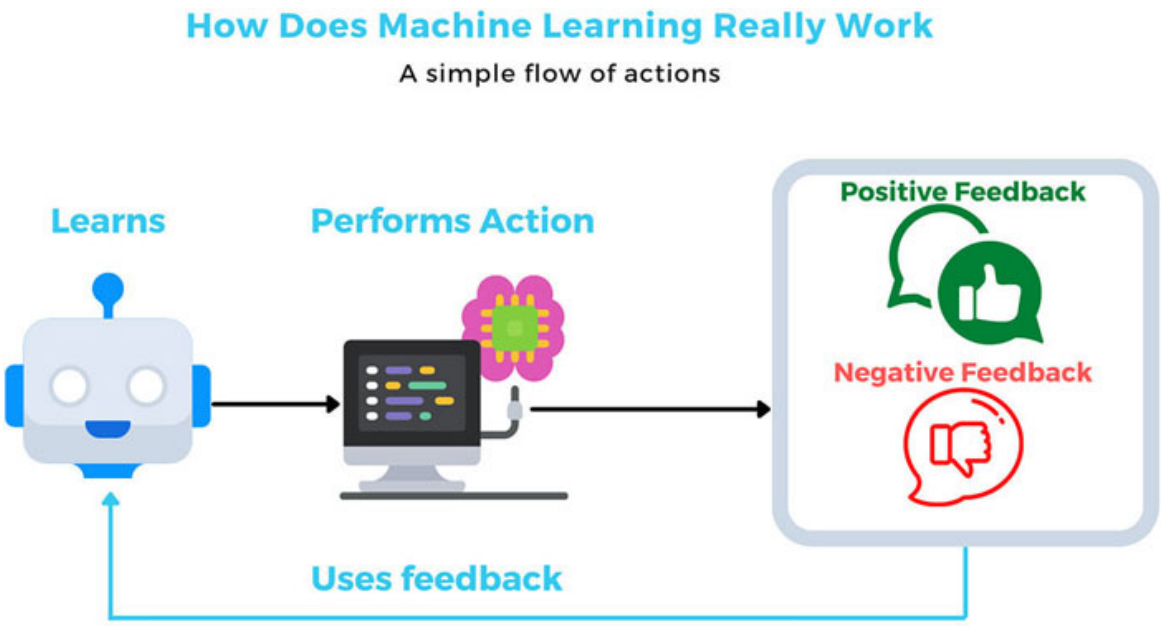


Figure 1.2: *How does machine learning work?*

The following figure gives an overview of the most common types of machine learning techniques available today: supervised learning, unsupervised learning, and reinforcement learning:

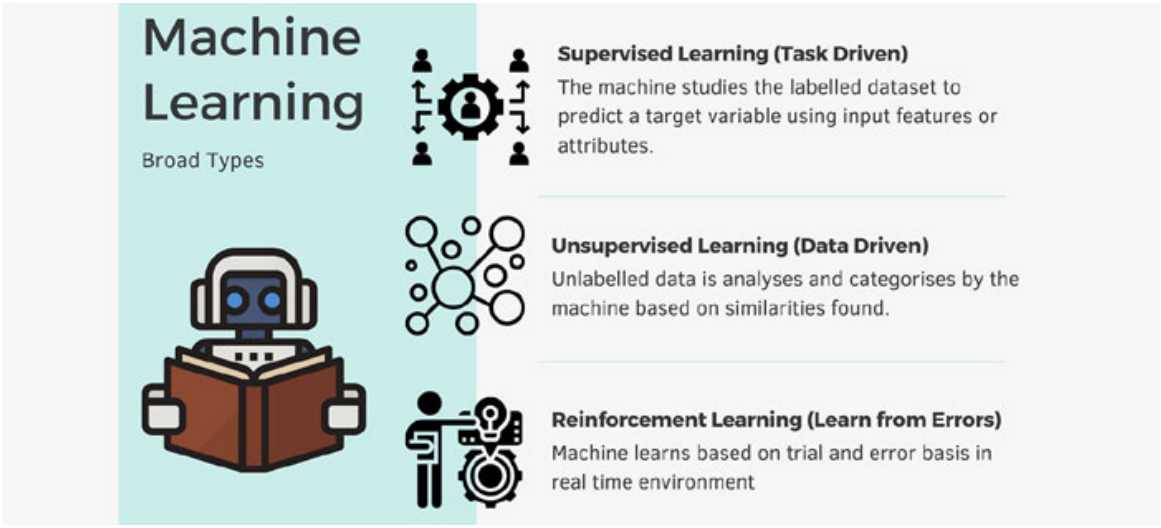


Figure 1.3: *Types of Machine Learning*

Supervised Learning

When an algorithm is trained to predict an output (also called a label or target) from one or more inputs (also called features or predictors), we say that the algorithm is engaging in supervised learning. As the name implies, “supervised”

learning occurs when the algorithm is given labelled training instances that consist of input-output pairs. The algorithm's objective is to generalise from the training examples and provide reliable predictions on novel, unseen data. For example, to design a face detection algorithm, you can feed or train the machine with images of people, animals, structures etc., along with their labels, to a point where the machine can accurately recognize a face in an unlabelled image. *Figure 1.4* depicts this learning process where the machine is provided with an image of mangoes labelled as mango, learns various hidden features and then predicts in real time whether an image contains mango when an unlabelled image is given.

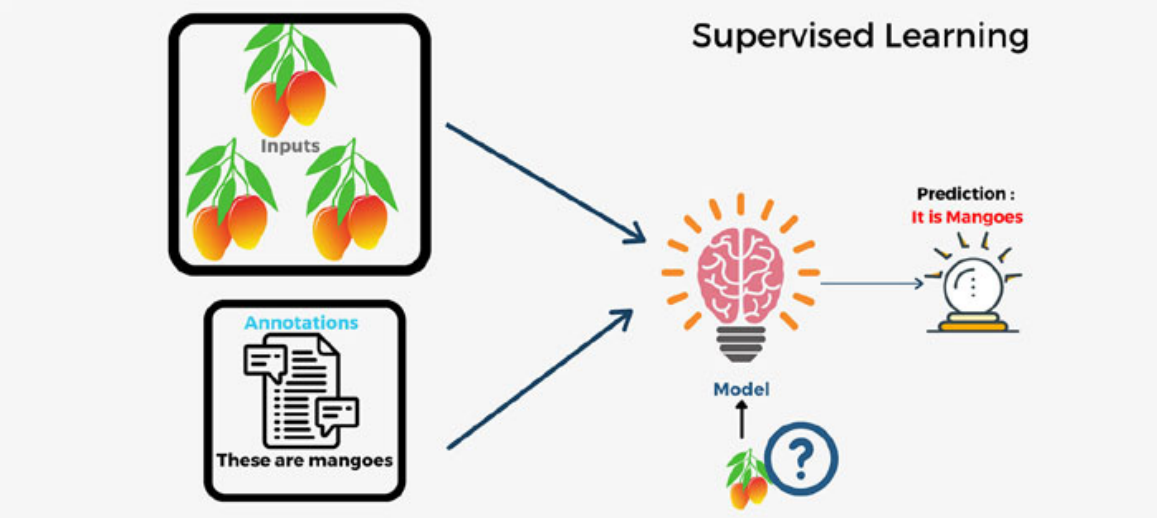


Figure 1.4: *Supervised Learning*

Unsupervised learning

Unlabelled data is analysed and categorised by the computer based on similarities found. Unlike the previous example, you provide the image but without labels. Still, the machine will be able to group the images based on certain common features (for example, texture, edges, and colour). The missing piece, however, is whether spherical objects qualify as faces.

The primary purpose of these algorithms is to unearth previously unseen patterns in data. In the following image, you can see how unlabelled data is provided, and the algorithm is left to discover patterns and relationships within the data on its own to cluster the data such that it makes sense.

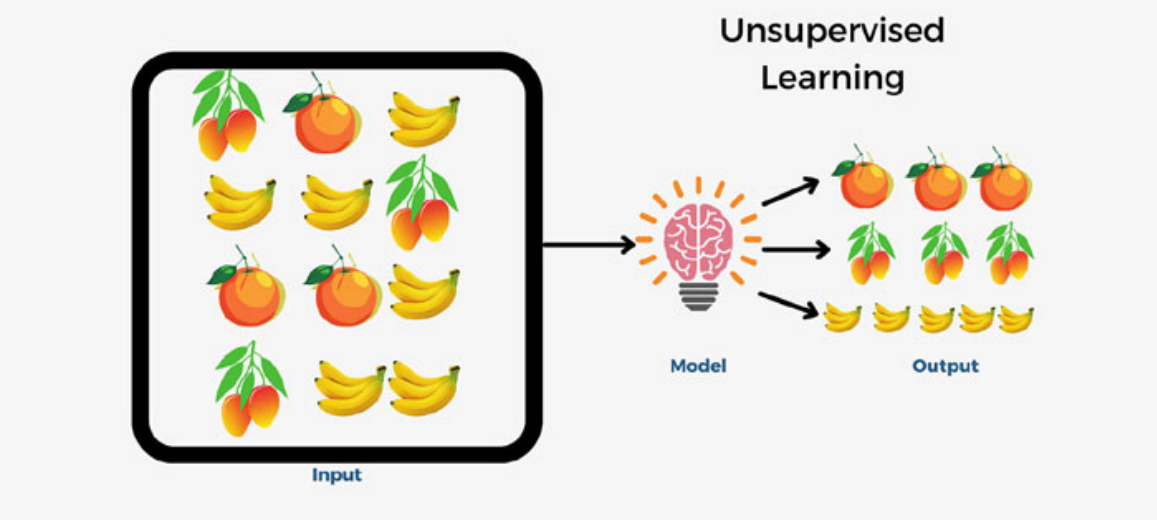


Figure 1.5: *Unsupervised Learning*

Reinforcement learning

Reinforcement learning is a form of machine learning technique that is unique and totally different from supervised and unsupervised learning techniques. We use the principle of giving incentives for any good outcome as the foundation of our algorithm.

Let us use the example of a dog to make it clearer. We may teach our dog to perform specific acts, but it will not be easy. You would tell the dog to perform certain tasks, and with each successful completion, you would reward them with a biscuit. The dog will recall that if it performs a specific action, it will be rewarded with biscuits. This will ensure that it follows the instructions correctly the next time. Here, we impose or attempt to impose a correct action in a specific manner. In a nutshell, **reinforcement learning** is a form of learning technique in which we reward the algorithm with feedback so that it can learn from it and enhance future performance. In the following image, you can see how the machine can learn in a real environment by making mistakes and learning from it through a feedback loop, just like a human learning to ride a bicycle.

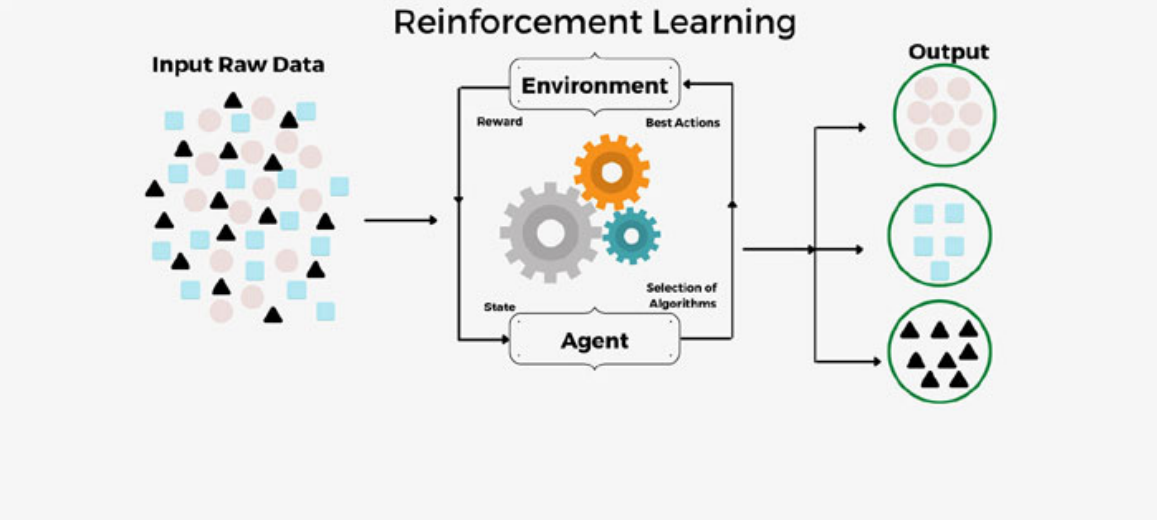


Figure 1.6: *Reinforcement Learning*

Benefits of machine learning in business

Many businesses dealing with large quantities of data have acknowledged the value of machine learning. Starting from the talent hiring process to achieving favorable business results, everything is hugely data-driven. For example, a candidate's resume for a job role passes through an automated Applicant Tracking System that matches the keywords in the resume with the keywords in the job description to select or reject a candidate.

Organizations can operate more effectively or gain an advantage over rivals and competition by carefully extracting information from this data – often in real time. Can you picture how many people who want to purchase Nike shoes online end up on Amazon instead? Amazon, like most other online retailers, leverages customer search data and keywords from other sites to fine-tune its own website's keyword strategy and attract more shoppers.

Try this: Search "Buy Nike Shoes 'online in your city", check which website and what offers show up for you. For each person, depending on their past searches for related objects, it will be different. This is the strength of data, which has the ability to transform businesses. Machine learning's ability to scale through a wide range of fields or domains, such as contract management, customer service, finance, legal, distribution, quote-to-cash, quality and pricing, is due to its ability to learn and evolve over time.

Machine learning has several completely realistic applications that can lead to tangible business outcomes – such as saving time and money – that can have a significant impact on the company's future and resource optimization plans. At

critical interaction points, we are seeing a huge effect in the customer service industry, where machine learning is helping people get things done faster and more effectively.

Machine learning automates tasks, such as updating login credentials, which would otherwise entail support from a live agent. Every website, from “Policy Bazar” to “Pepper fry,” now has a chatbot. This frees up staff time to provide the type of individualized attention to customers that can’t be replicated by a machine.

The most powerful means of engaging trillions of users in social media is machine learning. Machine learning is at the core of all social media sites for their own and consumer benefits, from personalising news feeds to making targeted advertisements.

The most common application of machine learning that uses image recognition is Facebook’s and Apple iPhone’s auto-tagging function. This is something you might have noticed while saving or uploading photos if you are a regular user of Facebook or Apple. Facebook’s face recognition can identify your friend’s face with only a few manually tagged photos (almost equalling human capabilities). The three phases of image recognition are detection, classification, and identification. Detection is the process of analysing a picture to spot certain objects within the image. The terms “Image Detection” and “Image Classification” are sometimes used interchangeably. Classification is used to categorise objects within the image or the image itself, while detection can be used if your aim is to just count the number of objects within the image without knowing what the object is. Identification is the process of analysing the likeliness of a face or object match between two or more photos of the same individual or item. Other uses of facial recognition systems are attendance tracking, airport authorities to verify passenger information, and so on. However, one has to be mindful about the bias that goes into these applications as well. We would highly recommend watching the Netflix documentary “Coded Bias” to know more about this.

Combating fraud is one of the most popular applications of machine learning in the banking and finance industry. Machine learning is ideally suited for this use case because it can sift through massive volumes of transactional data and spot anomalies. Any transaction a customer makes is evaluated in real time and assigned a fraud score, which indicates the chances of the transaction being fraudulent. In the event of a fraud transaction, depending on the severity of fraud-like trends, the transaction is either blocked or handed over for manual inspection.

When you swipe your card for a large transaction, you would get a call from your bank requesting for a confirmation on the transaction. This is a real-life example of how financial data is being used by the bank to identify a suspicious transaction.

Manufacturing or developing a new drug is a costly and time-consuming process. It involves multiple iterations of combining various components, testing them at different levels and finally, identifying a potentially effective drug. Nothing could have made us realize this better than the COVID-19 pandemic. With advancements in technology, some of these steps and iterations can be accelerated with machine learning, making the entire drug discovery process much faster.

- In cancer immunotherapy research, IBM Watson is being used by Pfizer. Pfizer also used an advanced ML tool called "*Smart Data Query*" to quicken vaccine clinical trials. What normally might take years can be developed within months with state-of-the-art effectiveness.
- Some of the other common use cases of machine learning in today's business world can be seen in the following infographic:

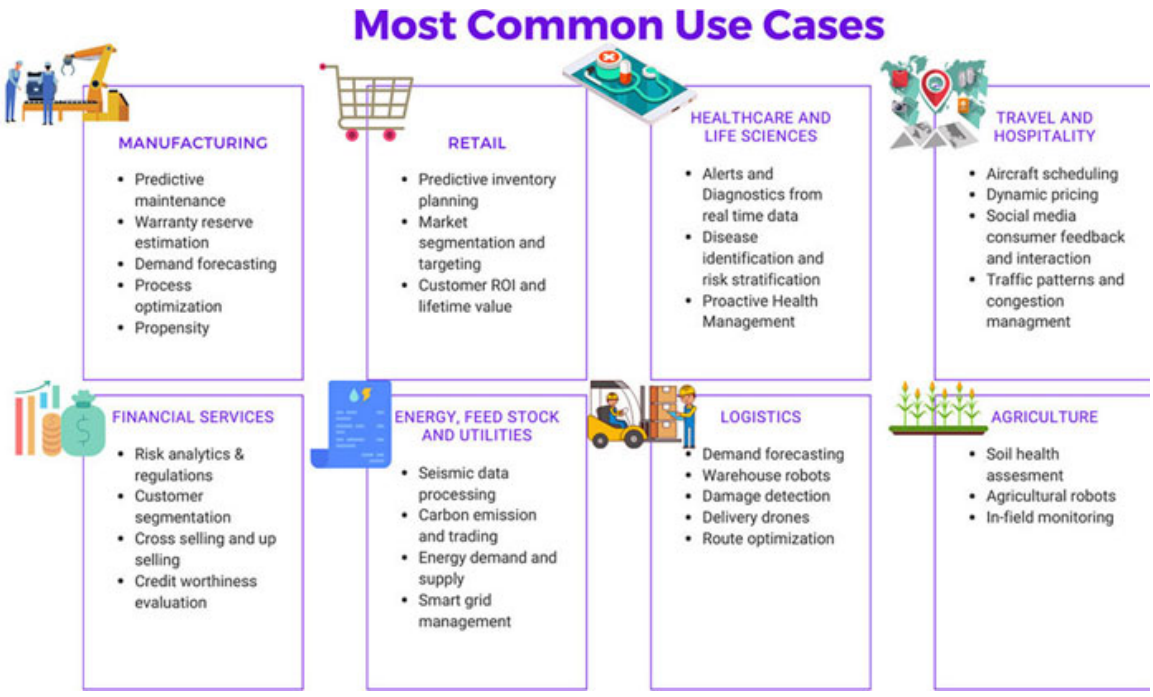


Figure 1.7: *Common Use Cases of Machine Learning*

Introducing data

We have seen the details of machine learning, the most used algorithms and the industrial applications, but all these are dependent on data. What is data? Do we need data in a specific format to make sense out of it? Well, some of your questions will be answered in this segment:

Data is a collection of info (numbers, terms, measurements, observations, and so on) that has been converted into a computer-processable format. Whatever industry you work in, you have almost certainly come across a story about how "information" is changing the way we live. It may be used in a study to help cure a disease, increase a company's income, improve the efficiency of a house, or be the source of those targeted advertisements you keep seeing. Data can be numbers, text, visual, fact, graph and so on, as shown in the following image:

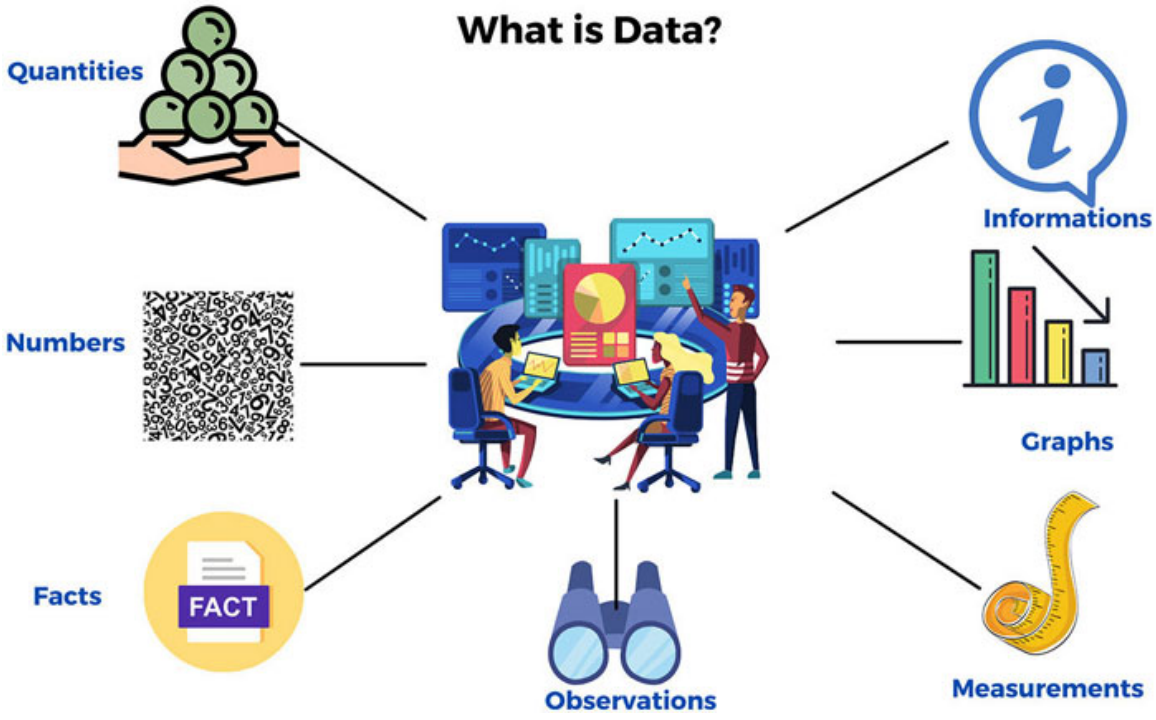


Figure 1.8: *What is Data?*

Data is the bedrock of data science; it is the raw material that all analyses are built on. There are two types of data in data science: conventional data and big data.

Conventional or Traditional data is organized and stored in databases that analysts can access from a single computer; it is in table format and contains numeric or text values. Let us use the word "traditional" for clarity's sake. It assists in emphasizing the difference between big data and other forms of data.

- **Personal information:** Personal data refers to information about you that is exclusive to you. It includes information about your demographics, location,

email address, and other distinguishing characteristics.

- **Data from transactions:** Anything that involves an operation to obtain is considered transactional data, for example, clicking on an advertisement, making a purchase, or visiting a specific web page. Businesses need transactional data because it allows them to reveal variability and optimize their activities for the best possible performance.
- **Data from the internet:** Web data is a broad term that refers to any form of data you might obtain from the internet, whether for research or otherwise. This may include information about what your competitors are selling, publicly available government statistics, football scores, and so on. Web data can be used to keep track of competitors, potential clients, and channel partners, to generate leads and create applications, and much more.
- **Sensor Data:** The Internet of Things refers to the collection of sensor data produced by objects. It includes everything from a smartwatch that monitors the heart rate to a building equipped with weather sensors.

Big data, on the other hand, is much larger than conventional data, and not just in the sense of volume. Big data is typically spread through a network of computers because of its variety (numbers, text, but also images, audio, mobile data, and so on), velocity (retrieved and computed in real time), and volume (measured in tera-, peta-, and exa-bytes).

Big data differs from "conventional data" in that it requires new approaches to data collecting, storage, and analysis due to its massive scale and complexity. To understand this difference further, refer to the following image:

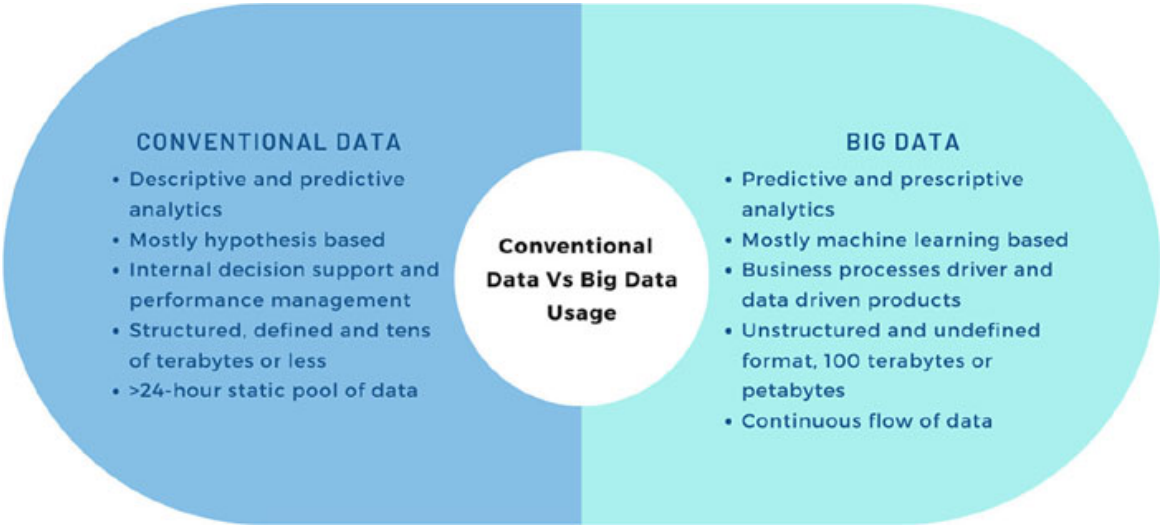


Figure 1.9: *Conventional vs Big Data*

Types of data in business context

When it comes to business applicability, we would need to understand what kind of data would be needed to make profitable business decisions. Most businesses houses, as explained earlier, store billions of data points that are collected from hundreds of processes. Here, we would see how big data, smart data and dark data are changing the decision-making process in organizations:

Big Data: Today, one of the most popular buzz words that you can hear everywhere is “big data”. It is information that is distinguished by its volume, variety, and velocity. This indicates that it is too large and complex to be analyzed using conventional business intelligence methods. Companies spent the previous decade studying how to work with big data and how to use it to make better business decisions. Big data is everywhere, and it is not a new concept. Big data, for example, used to be measured in terabytes, but it is now measured in exabytes as the volume of data increases over time. There are different features of big data, which are popularly known as 6 Vs of big data. It is important to get a clear idea about these; the following image will give you a clear picture. It shows the categorization of these Vs in terms of data properties:

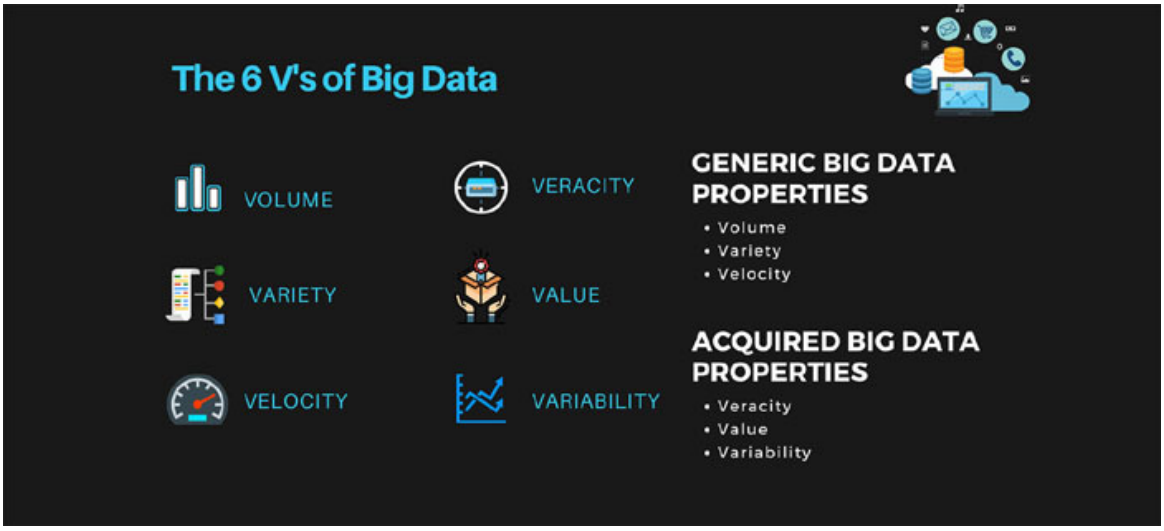


Figure 1.10: *Six Vs of Big Data*

Smart Data: Smart data, in comparison to Big Data, is actionable, makes sense, and serves a specific function. It is not so much about the amount of data you gather as it is about the actions you take in response to it. It is a term that arose in tandem with the advancement of algorithmic technology, like artificial intelligence and machine learning. Edge computing technologies are typically used to produce smart data near the data source. Instead of gathering all data from

a source, we process it in the source and dump only the valuable data into our data lake: the Smart Data. In a nutshell, big data and analytics together constitute Smart Data, as shown in the following image:

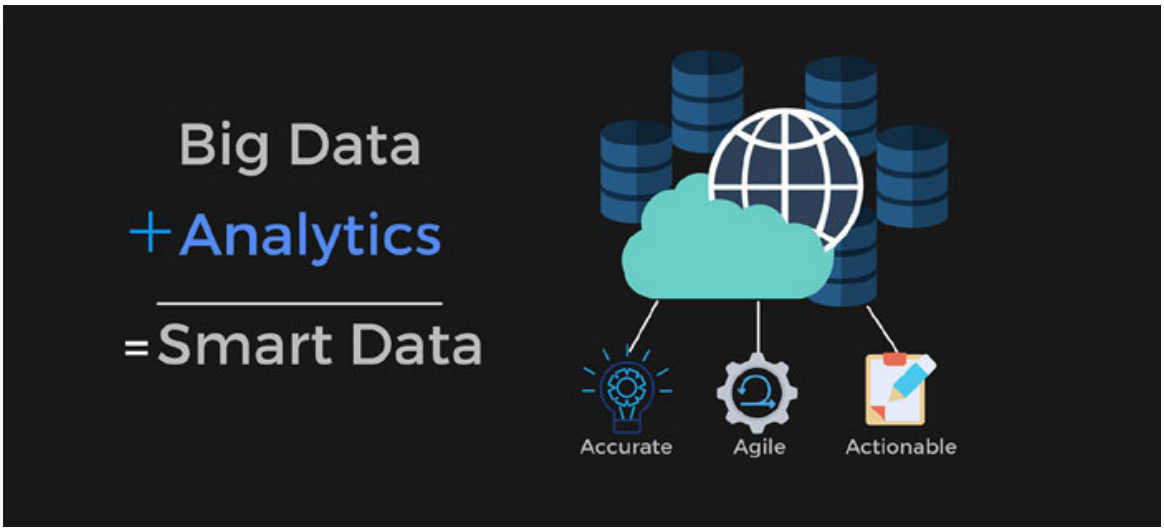


Figure 1.11: *Smart Data*

Dark Data: It's common to find that a considerable amount of data remains unused, such as the log files containing manually entered customer data from various sites. These data points are usually manually collected by a customer and shared with the service provider to be stored in the database. However, this data is mostly collected and stored, and there are limited occasions when this data is retrieved and correlated with other data points to gather insights.

This is nothing but Dark Data, data points which are collected and stored but hardly used; however, when they are used, they can unveil a whole new world of business opportunities. Dark data is data that exists under the surface, hidden within a company's internal networks and containing large amounts of relevant information that can be transferred to a data lake and used to produce critical business and organizational insights. This type of data accounts for most data currently available.

According to KPMG analysis, dark data accounts for 80% of all data. As a result, there are numerous untapped opportunities lurking in the dark corners of the data world. The following image will give you an idea of what constitutes dark data:

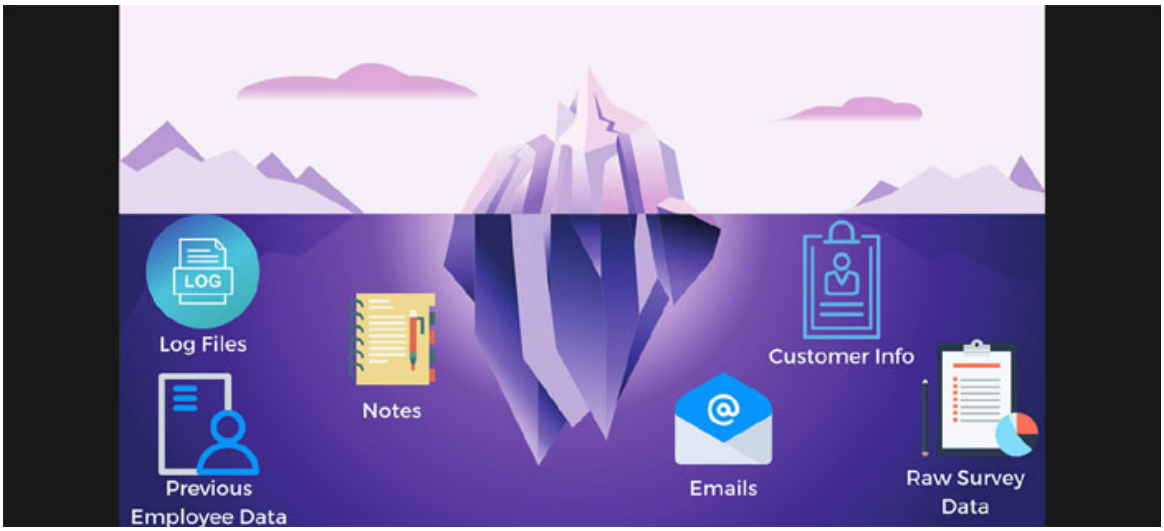


Figure 1.12: *Dark Data*

Transactional data: Do you know that every time you perform a task, there is a data point associated with it. This type of information defines the primary business operations. If you are a trading firm, this may provide information about your buying and selling activities. If you run a manufacturing company, this is where you will keep track of your production activities. Data related to recruiting and firing workers may also be categorized as transactional data. Therefore, in comparison to other types of data, this type has an exceptionally large volume.

Master data: It is made up of key pieces of information that make up transactional data. Let us visualize an example here: a cab company's trip details, for example, can include driver, passenger, path, and fare information. The master data includes the driver details, passenger details, places, and basic fare details. The driver data could include the driver's name and all associated details. The same goes for the passenger details. They make up transactional data when combined. Master data is application-specific, which means its uses are limited to the application and business processes associated with it.

Reference data: A subset of master data is reference data. Most of the time, it is structured data that is regulated by a set of rules. Understanding reference data is not difficult. It is data that determines values for other data fields to use. These values are frequently consistent and do not fluctuate significantly over time. Country codes and units of measurement are examples of reference data.

Reporting Data: It is a set of aggregated data used for analytics and reporting. There are three types of data in this set: transactional, master, and reference. For example, on March 13th, in the Pune Hadapsar area, trip data (transaction +

master) was collected (reference). Data reporting is very strategic, and it is normally done as part of the decision-making process.

Metadata: It is word that refers to data that identifies other data. Metadata, to put it another way, is data that explains the structure and provides context to other data. It describes data concepts that you might not notice at first, such as data use, data producers, data users, and data connections. You might think of it as a small book containing everything you need to know about your data. Simply put, metadata is information about information. In your day-to-day life, every data point has metadata associated with it.

Challenges with data

When we are talking about data, machine learning, information, and insights, we need to accept the fact that the whole process of data-driven decision-making comes with a lot of challenges. Most of the challenges lie with the data itself and range from measurement to storage. Some of the common data-related challenges are discussed here:

- **The volume of collected data:** You would notice that you would mostly get frustrated by the volume of data generated in your organization. An organization can obtain information on every incident and interaction every day, leaving analysts with thousands of interlocking data sets. Hundreds of data points get collected every second, minute, and hour. This calls for a system for collecting and organizing information automatically. In today's world, manually executing this procedure is very time-consuming and wasteful.
- **Gathering appropriate, real-time data**: You could wonder whether all the data being gathered is really helpful in making decisions. With so much data available, it is difficult to find the information that is most important and locate the details that are most relevant. In addition to this remote analysis, bias in data, various sources of data, and so on add complexity and can negatively impact decision-making. This problem can be solved with the help of a data system that gathers, organizes, and automatically informs users of trends.
- **Information from various sources**: The next issue is acquiring and analyzing data from various sources. These sources also contain different types of data. Sometimes it is possible to be unaware of this fact, resulting in partial or incomplete analysis. Manually integrating various data

consumes a lot of time and forces one to analyze only the information that is readily and easily available. Employees would have access to all forms of information in one place through a robust and centralized system. This not only saves time spent accessing different databases but also allows for cross-comparisons and guarantees that the data is accurate.

- **Data that is inaccessible:** Data gathered and packaged into an integrated system (often called business warehouse) has no impact until it is accessible by people who need it the most. Hence, all organizational data should be available for front-line decision makers and analysts who operate remotely. Any problems with usability would be eliminated by a well-designed database. Authorized employees will be able to securely access or edit data from any place, highlighting organizational improvements and facilitating quick decision-making.
- **Data of poor quality**: Inaccurate data is the greatest enemy of data analytics. If you feed useless data in your machine learning models, you would get useless results out of it. It is like human mind, if you think of negatively about things, you will end up having only negative outcomes of your actions. Simply put, it is garbage in, garbage out. Imagine a situation when you were analyzing a data sheet or an Excel file with data points that consisted of city names. How many different spellings of the same city did you find when the users had to enter the input manually? Typos and manual errors are a major reason for inaccurate data. If the research is used to influence decisions, this can have serious negative effects. These problems are solved by using a centralized structure. With mandatory or drop-down fields, data can be entered automatically, leaving no space for human error.

What you would need to recognize and understand is without organizational help, both at the top and at the bottom, data analytics would be ineffective. As a millennial, you would understand the different aspects of data-driven decision-making and how it can be beneficial to the whole organization, and also how much you can do with the data. However, if the top bosses do not understand that, then it would be nearly impossible for you to utilize the data to your advantage.

Citizen data science

These days, the coolest of job roles in terms of money, fame and reputation is that of a data scientist. Everyone wants to be a data scientist, and some even believe that they are already one if they have dealt with data. Just Google this, and you might find interesting articles. However, the truth is you cannot become a data

scientist in a year; it is a learning process over the years. A citizen data scientist is a person who has a strong business or functional understanding and is inclined to take data-driven decisions with easy-to-use analytics tools.

Gartner coined the term “citizen data science”. The describe it as “An individual who constructs or produces models that use advanced diagnostic analytics, predictive and prescriptive capabilities, but whose primary job role is outside the field of statistics and analytics”. In a nutshell, they are non-technical workers who can solve business challenges using data science methods. What sets them apart from a data scientist is their strong background on the domain knowledge. They can successfully incorporate data science and machine learning into business processes because of their business knowledge and understanding of business goals. The following image summarizes the traits of a citizen data scientist, which combine business understanding and processes with basic data science skills:

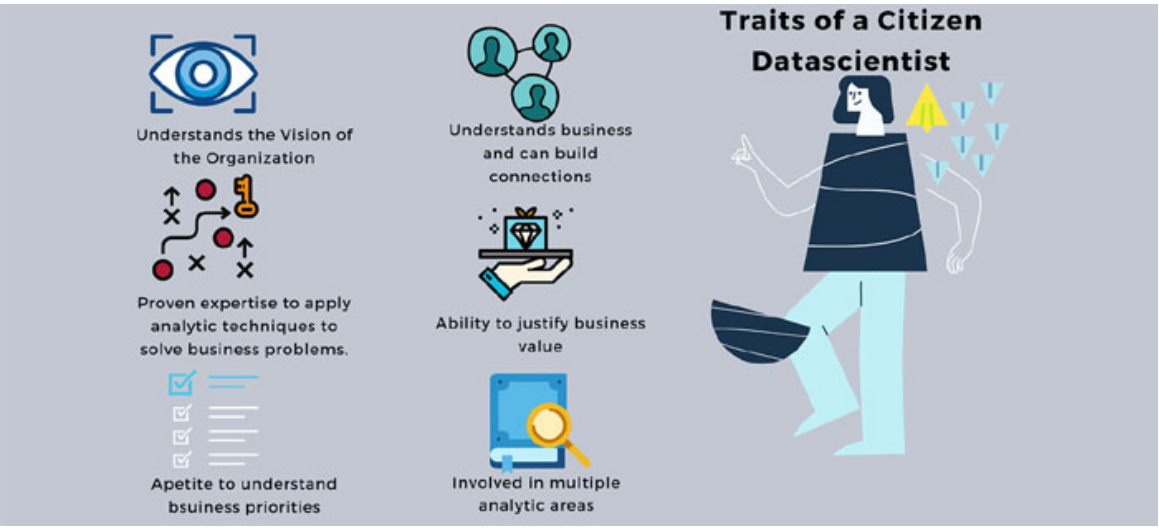


Figure 1.13: *Citizen Data scientist*

There are tools to help citizen data scientists, as mentioned below.

Automated Machine Learning (**AutoML**): AutoML solutions can help citizen data scientists by automating manual and repetitive machine learning tasks. The following are ML tasks that AutoML tools can automate:

- Preparing the data
- Feature extraction
- Selection of algorithms and hyperparameter optimization

Augmented analytics/AI-driven analytics: ML-led analytics, where tools extract insights from data in two forms:

- **Search-driven**: Software responds to citizen data scientists' questions with results in various formats (reports, dashboards, etc.).
- **Auto-generated**: Machine learning algorithms recognize trends and automate the production of insights.

Drag-and-drop interfaces, no/low-code and RPA solutions reduce coding, allowing citizen developers to deploy the models they create in production.

Data science for leaders

It might sound unusual, but data science leaders should focus on influencing others to embrace the data-driven decision-making process. Most of the times, stakeholders expect data science teams to deliver magic, that is, take any data and turn it into a magical solution. Sometimes, when it does not happen, there are a lot of unethical practices adopted to achieve the perceived outcomes. You would come across situations when the stakeholders already have a set of expected results that they would want to achieve, but when the analysis is being done on the data, the expected results might not come out. This is where the conflict starts.

As a Data Science Leader, you need to understand that you should have an unbiased look at the results that the data tells you. It is like when you think that the next car you would want to buy would be blue in color, and you end up seeing only blue cars on the road. In most organizations, key decision makers believe that their intuitive power is far more accurate than data-driven insights, and this might result in their lower confidence level in your and your team's ability to deliver results. You might face certain unique challenges, as follows:

- The competition to hire and retain right data science talent is fierce and highly stressful. There is a lot of demand for data scientist and a severe shortage of the skilled data scientists.
- The ethical dilemmas in data science are especially perplexing.
- There is no universally accepted method for handling data science teams.

If you are leading a team of data driven individuals, the following tips might be helpful:

- Focus on "Why" rather than "What" and "How". The "what" and "how" is something the team can collaboratively find out, but as a leader, you need to guide them to explore the "Why." When starting a new project, dig deeper to ensure that every project the team takes on has a "project why" that is compatible with the team's motivating intent. This takes time and effort.

However, investing in a straightforward “why” pays off for the team in various ways, including increased efficiency, improved employee retention, and, finally, clearer analyses and performance.

- Engage stakeholders, identify the right stakeholders (might not be just the usual project requester) and identify their needs carefully. Sometimes, all the needs are not laid down in a project charter, so dig deeper. Do not keep stakeholders in the dark or misunderstand their “requests” for “needs.” Instead, dig a little deeper.
- Implement effective processes, focus on the deliverables at continuous intervals. One of the successful methodologies would be Scrum. However, do not just implement Scrum because every other organization is doing it; you need to have a better understanding of why you should go for a specific project management methodology. This is closely related to the “Why” of the project itself: what would suit the project nature and its complexities.
- Last but not least, build the right data science team. It is not true that a fully functional data science team has a bunch of data scientists. Rather, a data science team consists of talented people who are focused on delivering a solution. In doing so, some of the team members might be data scientists, some might be data analysts, some might be data engineers, and some might be product managers as well. It depends on your solution and the organizational bandwidth to hire relevant talent.

NOTE:
Myth: Data Science Teams are Software Teams

Truth: Data science teams are not software teams, but they belong to diverse background who are focused on discovery and exploration. Their skill set includes mathematics, statistics, business knowledge, and some amount of coding.

One of the topmost qualities as a Data Science Leader you should possess is the ability to question “Is it ethical?” Adopting ethical practices to let the data do the talking is important, and it is necessary to keep biases out of the way. It would help you to understand when you need to learn more or upgrade your team’s capabilities.

Conclusion

Machine learning, in amazingly simple terms, is learning from data by finding patterns of interest in the data. When you conjoin data with the right mathematical algorithm, you end up building a relevant machine learning model. Data is the backbone of machine learning, so relevant and clean data would be instrumental in driving the accuracy of your machine learning model. Data cleanup also takes the maximum amount of time.

Thumb Rule: Garbage in Garbage Out

Always remember, machines do not take decisions, people do; hence, you need to take the best decision for your business based on the recommendation of the machine learning model. Understanding the business problem is critical and hence, the need for citizen data scientist is growing. These citizen data scientists have strong functional knowledge, which equips them to make sense out of the data much more easily than a data scientist. One of the hidden threats to a machine learning model is of biasness, so as a leader, you should ensure that the machine learning models are less impacted due to biases. Last but not the least, organizational maturity and support is critical to the success of data-driven decision-making process. This is more of a mindset change that everyone in the organization needs to embrace. In the next chapter, we will look at the intricacies of data and the impact of data cleaning on decision-making process.

Points to remember

- Every day, our lives are touched with machine learning algorithms in some or the other form:
    - Google ranking of website
    - Bank’s sorting and deciding whom to provide credit cards and suggesting investment plans
    - E-commerce websites suggesting products that you should buy or offering you discounts to avoid churn
    - Your social media feed based on your preferences and the contents that you prefer
- There are primarily three types of Machine Learning techniques:
    - In Supervised Learning (also called inductive learning), the desired outputs are included in the training data; learning takes place under supervision.

- Unsupervised learning occurs when the training data does not include the desired outcomes; clustering is a good example.
- Reinforcement learning is when a series of acts results in a reward. It is the most ambitious method of learning because it is based on feedback to the response generated.

Multiple choice question

1. What are the most common types of machine learning techniques?
    a. Supervised, Unsupervised and Reinforcement Learning
    b. Classification, Regression and Clustering
    c. Artificial Intelligence, Data Science and Data Analytics
    d. None
2. Which of the following is data collected and stored but often not used?
    a. Big data
    b. Transactional data
    c. Dark data
    d. Master data
3. A citizen data scientist is a person who has a strong business or functional understanding and is inclined to take data-driven decisions utilizing analytics tools?
    a. True
    b. False
4. What is the collection of sensor data produced by objects known as?
    a. Sensor data
    b. Internet of Things
    c. Personal data
    d. Database

Answers

Question Number	Answer

1	a
2	c
3	a
4	b

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 2

Know Your Data

Understand your data: where it came from, what it contains, and what it implies. Developing a strong learning model requires a thorough understanding of the data. The following are a few questions worth answering before building a machine learning model:

- Is the data reliable?
- Is the dataset balanced in terms of different attributes?
- Is there any inconsistency in the data?

This chapter will start off by helping you understand the data better and look at the steps to prepare the data, the journey of data science, the importance of math and some of the key mathematical concepts, like normal distribution, central limit theorem and hypothesis testing.

Structure

In this chapter, the following topics will be discussed:

- The most common types of data you will encounter in your school, college, and business landscape
- Prepare your data and understand its importance
- Understand the data science journey and the impact of clean data on this journey
- Mathematical concepts that you should remember

Objectives

After studying this chapter, you will understand how data impacts the complete data science process and why clean data is instrumental for generating meaningful insights. This chapter will also help you focus on key

mathematical concepts that you would need to refer to frequently as you embark on this journey.

Most common data types you will encounter

Remember the day you wanted to join the volleyball team and your physical trainer teacher asked you to get your height measured, or when you were asked to record your body mass for an entry into the kabaddi team. Imagine the day you had to take the time bound pH value during your chemistry practical class and identify the pH and time at which the colour of the solution would change.

Imagine your first day at work when you were asked to fill in the onboarding form, with information starting from your name to your emergency contact details.

All these are examples of different forms of data point we encounter through our educational and professional career. However, what is important to remember is the process of data analysis does not immediately start once data is collected. It is imperative that we understand the different categories of data points that we have collected. The pH values will not have a correlation with the height measurement even if the same person is in consideration. Data categorization and understanding what data we are dealing with is important.

Before you can begin to analyse your data, you must first become acquainted with it. The first step is to examine each variable and determine the data form to which it belongs. From here, you will be able to see which mathematical operations you can perform on these data and which ones you cannot. This may not sound exciting, but most researchers skip this phase in their research – and almost every one of them regrets it later.

Did you know that data may be classified into four types? You may have seen them before, and you may even know what they are called. If you do, congratulations; you are well ahead of most people! However, only a handful of us are familiar with these four data forms. They are listed as follows:

- **Ordinal**: Ordinal data type is a type of statistical data that can be measured and ordered, is equidistant and has a meaningful zero. Ordinal data types have a natural ordering, but they continue to

maintain their class of values. Let us take an example of clothing: when you go to buy a shirt for yourself, you can see on the tag that the order of the shirt size is small<medium<large<extra-large. This is a particularly good example of ordinal data. Also, the same can be observed in case of the grading system in a class of students in a school, where A+ is better than grade B.

- **Nominal**: Nominal data is the simplest type of data; it is described as data used to label or name variables. One of the distinctive properties of nominal data is its ability to be classified into distinct categories that do not overlap. This data is not measured or evaluated but is assigned to multiple groups. The groups are unique in nature and have no common elements. One must remember that nominal data can never be quantified; there is no order; it is qualitative in nature and a mean cannot be calculated. You might remember filling a form where you would have to select your gender, male, or female; this is a simple example of nominal data.
- **Ratio**: Since you can divide the values of these data, they are referred to as ratio data. Distance and weight measurements are both ratio. Here are a few examples: 20 m is twice as far as 10 m (20/10 = 2). 150K has half the energy of 300K (150/300 = 1/2). Almost every mathematical procedure on ratio data yields accurate results. You can add, subtract, multiply, divide, and compare (less than, equal to, or greater than) this data.

  It is worth noting that ratio data can be combined with other ratio data in a division. The ratio of the weight to the square of the height, for example, is used to calculate BMI, a quite simple measure of body fat. Both the weight and height, and the resulting BMI, are ratio data.The important thing to remember here is that for values to be divisible, the data must have a meaningful zero point. Since a tape measure, a bottle, or a collection of weighing scales cannot make negative measurements, everything weighed with them has an absolute zero and can only take positive values – negative numbers are not permitted.

  Interval: A numerical scale with an identical distance between neighbouring values is called an interval long. These values are referred to as “intervals.” A simple example would be the measurement

of temperature in Celsius and Fahrenheit. Each point on these scales is separated from its neighbours by exactly one-degree intervals. The difference in degrees between 20 and 21 is equivalent to the difference in degrees between 225 and 226. The key point to remember here is that the scales will have arbitrary zero points, which means zero degrees is not the lowest possible temperature.

The following image will give the various characteristics of different types of data-based rank, spacing, category, and so on:

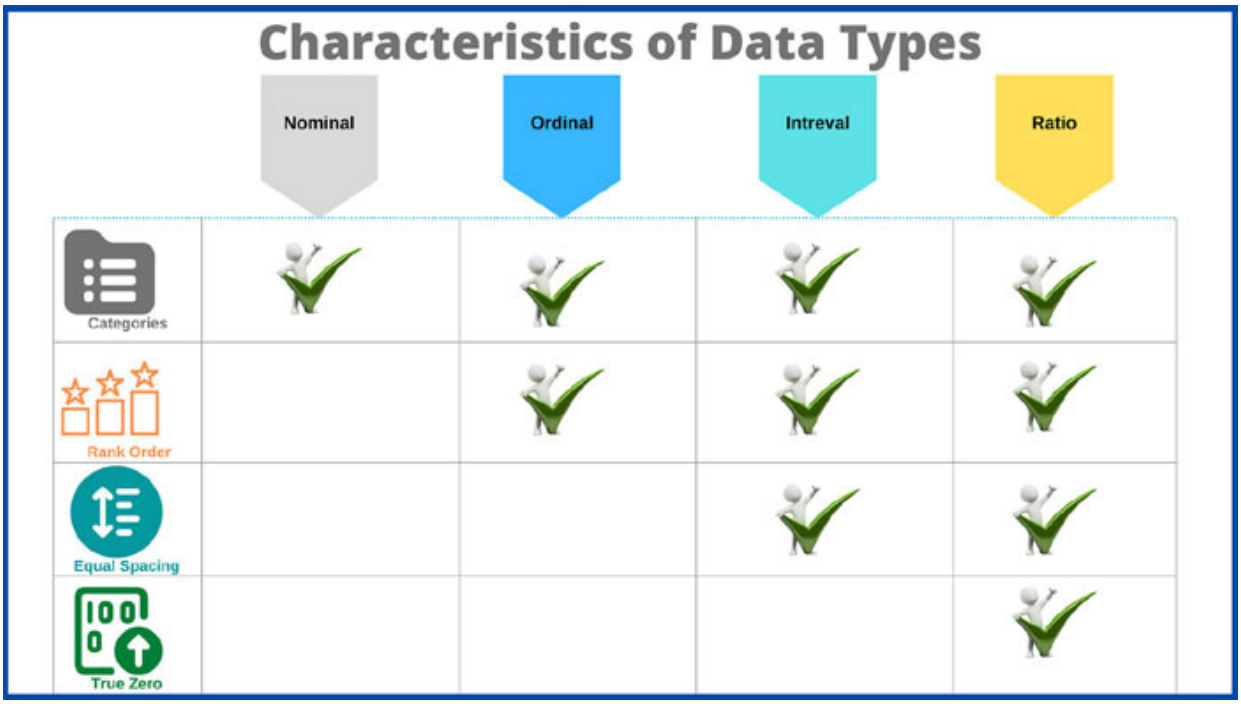


Figure 2.1: *Characteristics of Data Types*

Data preparation and understanding the criticality of the process

With the growing digitalization of business processes, organizations must empower as many individuals as possible to get meaningful insights from high-quality data. Data preparation is the only way to retrieve high-quality data. Effective data preparation enables more efficient analysis, minimizes data errors and inaccuracies while processing, and increases the accessibility of all processed data to users.

Data preparation is the process of cleaning and modifying raw data before processing and analysis. It is a critical stage prior to processing that frequently includes data reformatting, data modifications, and data integration to enhance data. For data experts or business users, data preparation might be time-consuming, but it is necessary to put data in context to turn it into insights and reduce bias caused by poor data quality.

Standardizing data formats, enhancing source data, and/or reducing outliers are all common steps in the data preparation process.

Data preparation strategy: Data preparation is not a stand-alone process; it needs a *data preparation strategy*, without which the whole objective of retrieving quality data would be meaningless. The data preparation strategy starts with formulating a workflow process which will cover all the steps which needs of your project and gives you clarity on what kind of data is available what is not available. Let us think of it as cooking a new dish from scratch using a recipe; the recipe will have all the ingredients and their quantities listed so that you know what would be going inside the pot and what exactly you need to do. Let us look at the strategy for data preparation:

- **Data cleansing**: After you have defined the data preparation strategy, the next step is to move into data cleansing. This essentially means removing data that is inaccurate, damaged or corrupt, or erroneous in some or the other way, which makes it undesirable for consideration during analysis process. Some of the errors might be due to human error, some might be due to corrupt sensors, and some might get corrupted during the transfer to storage systems.
- **Metadata creation:** Metadata refers to creating tags that provide information about the data. For example, when a picture is taken on your smart phone, you also capture other information, like when the picture was taken, where it was take, its geographical location, and so on.
- **Data transformation:** This stage entails converting data into a format that your analytics tools can understand. This entails taking the data in whatever format it was ingested in – whether by scanners, sensors, cameras, or manual human data input – and converting it to a database format that your analytics engines can understand. At this point, data

can be compressed to conserve space and increase performance, and any parts that will not be read by your analytics systems can be removed.

- **Data standardisation:** Data standardisation is the process of making sure data is logical and consistent, which means each type of data has the identical content and structure. Data standardisation has become increasingly important as data sources have become increasingly diverse, independent of sector, industry, or business goal. And completing the data standardisation process on a large scale now means the difference between success and failure for a business. Prior to analysis, data standardisation aims to convert raw data into usable information. Variations in entries that are supposed to be the same in raw data can influence data analysis afterward. The data that needs to be standardised will be modified as part of data prep so that it is consistent across all entries. Variables will be consistent, and all values will be in the same format. Standardizing data can make it easier to spot regression, patterns, and outliers in a dataset. It will be much easier to examine and use the dataset once the information is consistent and standardised.
- **Data augmentation:** Data augmentation is the process of synthesising new data from existing data. This might be used with any type of data, including numbers and graphics. In most cases, the supplemented data is comparable to the existing data. Data augmentation is a method for augmenting the size of a training set artificially by altering the existing data. If you want to avoid overfitting, if your original dataset is too small to train on, or if you just want to get more performance out of your model, employing data augmentation is a suggested technique.

The following figure shows the overall steps involved in data preparation, like data cleansing, transformation, standardisation, augmentation, and so on.

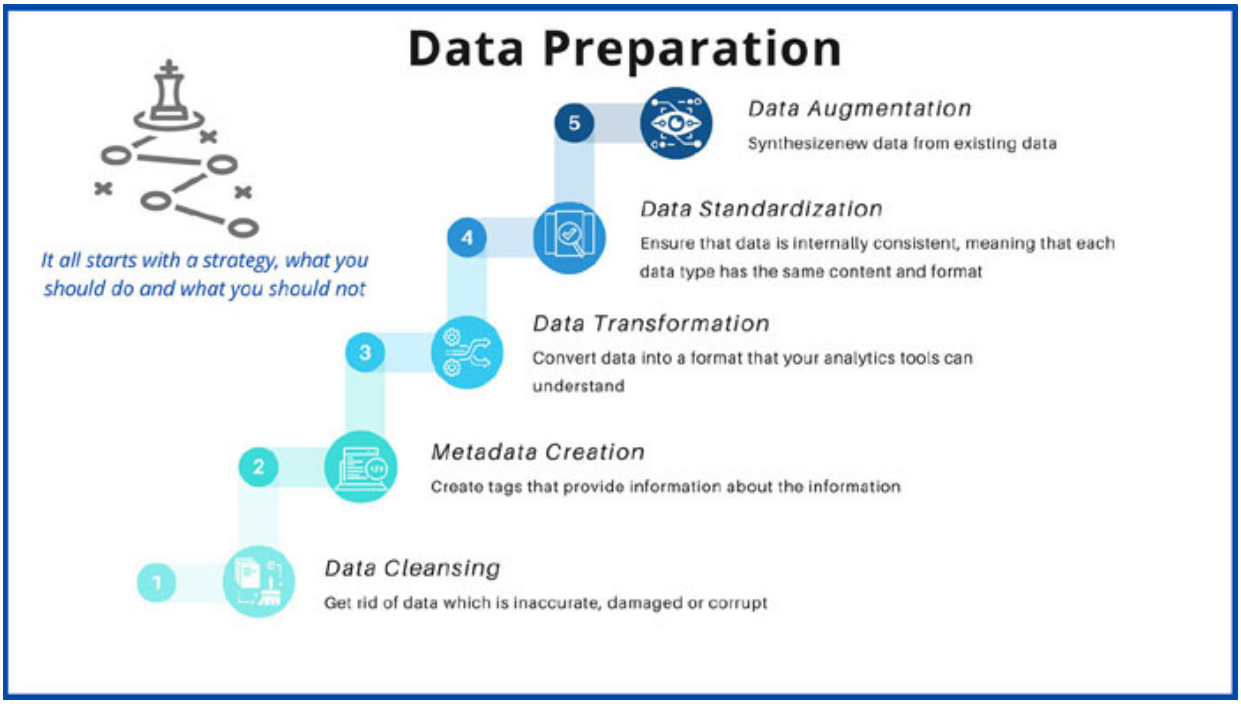


Figure 2.2: *Data Preparation*

Data preparation ensures data accuracy, resulting in accurate insights. It is likely that insights will be off owing to garbage data, an ignored calibration issue, or an easily corrected mismatch across datasets if data preparation is not done. Data preparation is a vital task that requires substantial time and resource investment.

According to data scientists and analysts, they spend around 80% of their time preparing data rather than analysing it. Data preparation helps eliminate errors in data before it is processed, resulting in higher quality data for analysis and other data management operations. It is vital, but it takes a long time and may necessitate specialised knowledge.

Business users can also establish trust in their data using data preparation technologies. This is accomplished by enhancing underwhelming datasets and merging data to address flaws. Users can use data preparation tools to check the accuracy of incoming data before devoting time and money to analysis. As a result, users can save a significant amount of time, allowing enterprises to accelerate their time-to-value.

The benefits of data preparation are numerous, and this step allows the business insights to be accurate. Some of the benefits are listed as follows:

- **Fixing problems quickly**: Data scientists can use data preparation to identify errors or fallacies before processing. When faults are discovered, they can be corrected right away.
- **Superior data generation:** Cleaning and modifying data streams before processing greatly enhances data quality. When data is of exceptional quality, it aids organisational success by allowing people to make decisions based on facts rather than habit, convenience, or human intuition.
- **Making better business judgments:** When data is correctly cleansed and then processed, it aids organisations in making high-quality business decisions. Companies can leave an unforgettable impact on their consumers and partners with this quick and effective decision-making capability.

The following figure will give an overview of the benefits of data preparation:

Figure 2.3: *Benefits of Data Preparation*

Some best practices that one needs to keep in mind while starting with data preparation are listed here; these can be considered as governing principles when one starts this process:

- Define business objectives and have a data governance policy in place. Data governance is an essential “wrapper” that defines the business goals, business lexicon, data quality, data lineage and auditing requirements that data preparation projects must adhere to. Finally, executive stakeholders must assume ownership of data governance activities, which requires them to perceive data as a strategic asset for their organisation. Some companies even have a Data Governance department alongside HR, Finance, Operations, and IT.
- Validate the reliability once there is clear understanding on the source of the data. Data sourcing is critical, and the format of the data source will have a great impact on the analytics.

  The data gathering process can divided into three simple steps:

  - Identifying the data required for a specific business task.
  - Identifying possible data sources and their business and IT owners(s)
  - Assuring that the data will be supplied at the needed frequency for the business task.

This stage normally involves some political bickering and negotiating, but it is vital to get a solid data source.

- Extract the data to a place where you can work on (a good work bench). This is a safe environment where one can analyse and manipulate the data. Text editors and spreadsheets can open smaller data files that retain a reasonable amount of their natural structure. For bigger and/or more intricate data sets, more advanced profiling approaches, such as those available in most **Extract Transform Load** (**ETL**) tools, advanced statistical software, or enterprise-class business intelligence services, will be necessary.
- Spend time understanding your data well so that you know how to use it. Along with a simple visual assessment, you must profile, visualise, detect outliers, and locate null values and other junk data in your data collection. The primary goal of this profile analysis is to determine whether the data source is valuable enough to include in your project.

- Always start small; take a random sample of the data for exploratory analysis and data preparation. By decreasing the latency associated with repetitive exploration of a massive data set, building data preparation rules on a representative sample of your data significantly accelerates your time to insight. This is a phase that combines science and art. To focus on the appropriate columns and rows to sample and ultimately prepare for further analysis, the data analyst should be thoroughly familiar with both the raw data and the business intelligence job at hand.
- Identify the data types in your data set and investigate the columns that contain them. Verify that the data types used in those columns correspond to the data that should be in them. Examine your data set's columns to confirm that the data type matches the data that must be in each column. For example, a field called "sales date" should have a value in the MM/DD/YY format. You should also be aware of the generic data type that each field represents. If it is a numeric field, is it ratio or ordinal?
- Graphing important fields may be an excellent method to familiarise yourself with your data. Histograms may be used to visualise the distributions of key variables, pie charts can be used to see values as a percentage of the total, and scatter plots can be used to do the all-important outlier identification. Graphing data has the extra benefit of facilitating and expediting the explanation of data profiling results to non-technical audiences.
- Verify that your data is reasonable. Understanding what particular columns mean, having reasonable amount of data that are appropriate for those columns, and applying common sense to the data set are all instances of sanity testing. Outliers can also be found with the help of automated techniques and graphing software. Outliers can significantly alter metrics that use the data's mean, which can lead to some difficult interactions with corporate stakeholders if they are not discovered and accounted for. Identify outliers, do studies with and without them, and provide the results to stakeholders as a starting point for a collaborative, constructive conversation about how to handle them.

- Understand that no data cleansing strategy is perfect, and it depends on the end business analytics goal. Experiment with various data cleansing procedures to obtain the relevant data in a format that can be used. Begin with a modest, statistically valid sample size and test different data preparation procedures iteratively, fine-tuning your record filters and consulting with business stakeholders.
- Take some time to rethink the subset of data you truly need to fulfil the business target once you have found what appears to be a suitable strategy. Running your data prep procedures on the complete data collection will take significantly longer, so talk to your business stakeholders about which columns you need and which ones you do not, and which records you can safely filter out.

The following figure summarizes the data preparation path from gathering, discovering, cleaning, transforming, enriching, and storing:

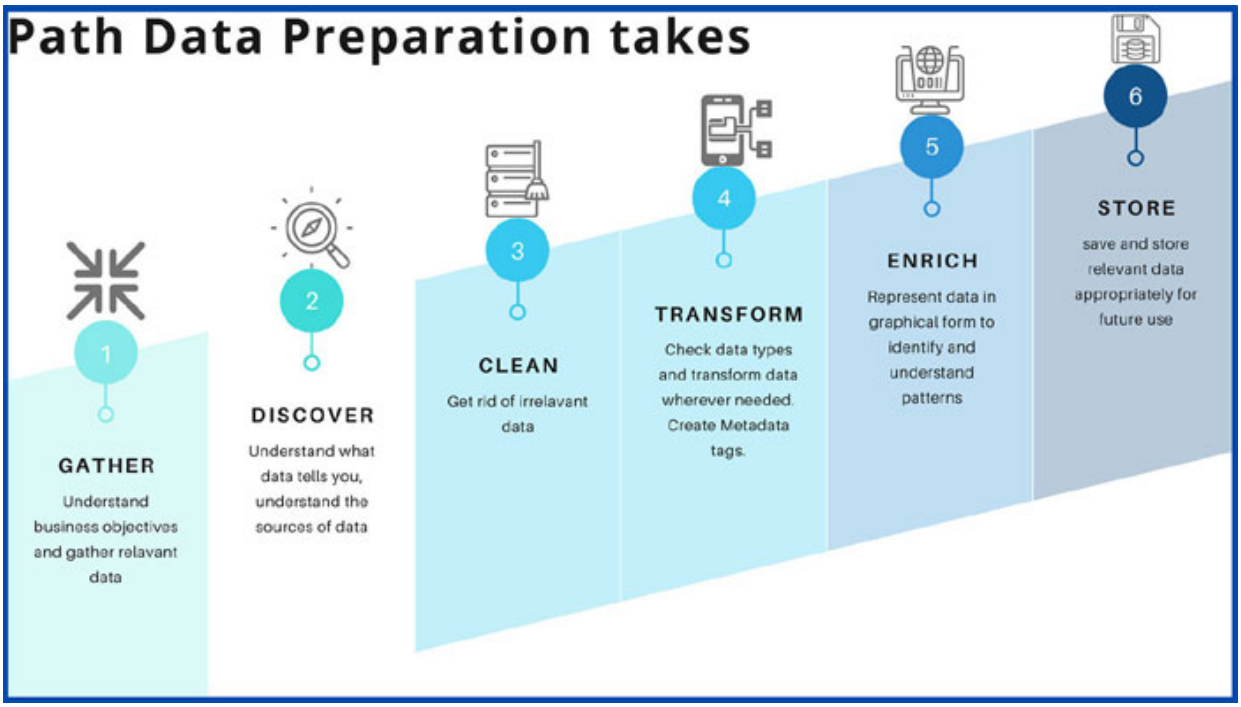


Figure 2.4: *Data Preparation Path*

Data preparation is a time-consuming, yet ultimately gratifying, activity. Making an initial investment in examining data sources and data types can save you a lot of time later in the analytics project. If you follow data governance principles and employ sample methods, profiling tools,

visualisations, and iterative stakeholder interaction, you can build an effective data preparation process that fosters trust in the data and earn the respect of business stakeholders.

Data science journey and impact of clean data

To get value from data, data science draws on various disciplines, including statistics, scientific methodology, **artificial intelligence** (**AI**), and data analytics. Data science refers to the process of cleansing, aggregating, and modifying data to undertake advanced data analysis. The results can then be reviewed by analytic applications and data scientists to uncover patterns and enable company leaders to make informed decisions.

Data science is one of today's most fascinating subjects. But why is it so crucial?

It's because corporations own a veritable goldmine of data. Data volumes have expanded as contemporary technology has facilitated the creation and storage of ever-increasing amounts of data. Every hour, for example, Facebook users post 10 million images. However, most of this data is languishing untouched in databases and data lakes.

The vast amounts of data collected and saved by these technologies has the potential to revolutionize businesses and communities all around the world, but only if we can understand it. This is where data science comes into play.

Data science identifies patterns and generates insights that businesses may utilize to make more informed decisions and develop more innovative products and services. Perhaps most crucially, it allows **machine learning** (**ML**) models to learn from the massive volumes of data that they are fed, rather than depending solely on business analysts to figure out what they can from the data.

While data is the bedrock of innovation, its worth is determined by the insights that data scientists can extract and act upon. The journey of data science is divided into five parts. Let's take a look at them:

- **Step 1**: Define your objective and frame your questions.

  What can be done to meet the various stakeholders' expectations? Is it necessary to include stakeholders in addition to the marketing

department? Who oversees the technical implementation?

Start jotting down questions, regardless of whether you believe the data needed to answer them exists. Create questions that will either qualify or eliminate prospective solutions for your problem or opportunity.

The first stage is to identify important stakeholders and outline the use case's objectives. The importance of this phase for overall success is sometimes overlooked. Data management can help with this phase by offering use case management principles and creating standards for documenting use cases. This implies another benefit: If all use cases are centrally documented, it will be easier to identify connections between use cases that are already in use or planned in different departments. This eliminates redundancy and enables the utilization of synergies. Furthermore, contact persons who are pertinent to the situation can be discovered more rapidly.

- **Step 2**: Explore the data

  What do the values 0 and 1 in the customer table's "active customer relationship" column mean? Why does the master data contain duplicate customers? Is the money spent at the online retailer a net or gross transaction? Is each sale displayed in a standardized or the local currency?

  It is important to assess current data and to weigh the relevance and quality of the data. It is also critical to understand whether the data available at hand will help you answer the questions you have framed in the first step.

- **Step 3**: Prepare the data

  Poor data quality or faulty data costs an average firm $13.5 million per year, according to a Gartner research report, which is an unacceptably high cost. Due to the fact that bad data or poor data quality can impact the accuracy of insights or even result in erroneous insights, data preparation or data cleaning is important, even if it is the most time-consuming and cumbersome task in the entire data science process.

  Data preparation needs to be codified as a best practice in the company. It will become a more efficient, consistent, and repeatable process,

thanks to shared metadata, persistent managed storage, and reusable transformation/cleansing algorithms. As a result, users will be able to easily find relevant data, and they will have the expertise they need to swiftly put that data to use. Business users can deal with data on their own, freeing up IT to focus on other duties, thanks to self-service data preparation tools. As a result, the entire organization gains productivity.

With bad data, even the best analytics are useless. It is critical to understand and work on all facets of the insights value chain.

- **Step 4**: Perform data modelling or value creation through the power of data science

  Once all data is accessible and understood, the primary work begins. The previous phase establishes a solid modeling foundation. Now is the time to select the appropriate model and calibrate it appropriately. The following are crucial success variables to consider:

  - The method is crucial. Finding the right method is critical, and the Data Scientist must present various options.
  - The first model will not resolve the issue. Applying a single strategy seldom resolves an issue immediately, and data-driven business models do not appear as instant outcomes, an agile approach, including feedback and iterative model development, is important.
  - Market conditions are always changing. The models should be able to learn and, in the best-case scenario, anticipate future events. As a result, a data model and the algorithms that go with it should be thought of as a product rather than a project.

- **Step 5**: Act on the derived insights for value generation

  How will stakeholders be able to see the results? How can the model be made to remain relevant in the future? How is the model capable of learning (automatically)? Which/how many of the organization's associated use cases can the model be used for?

  Many people conflate data with insights, despite the fact that the two terms have small but important distinctions. Data is a type of information that consists of numbers or words. Insights are the results

of examining data and drawing inferences from the data that can help your company. The input is data, and the result is insights.

Many distinct qualities can be found in actionable data insights. In general, insights are simple observations that may or may not lead to a decision or course of action. You will act because of the information you acquired if it is actionable. This serves as a model for increasing consumer value, while also increasing your company's efficiency.

The following image shows the mentioned phases of data science from objective till value generation:

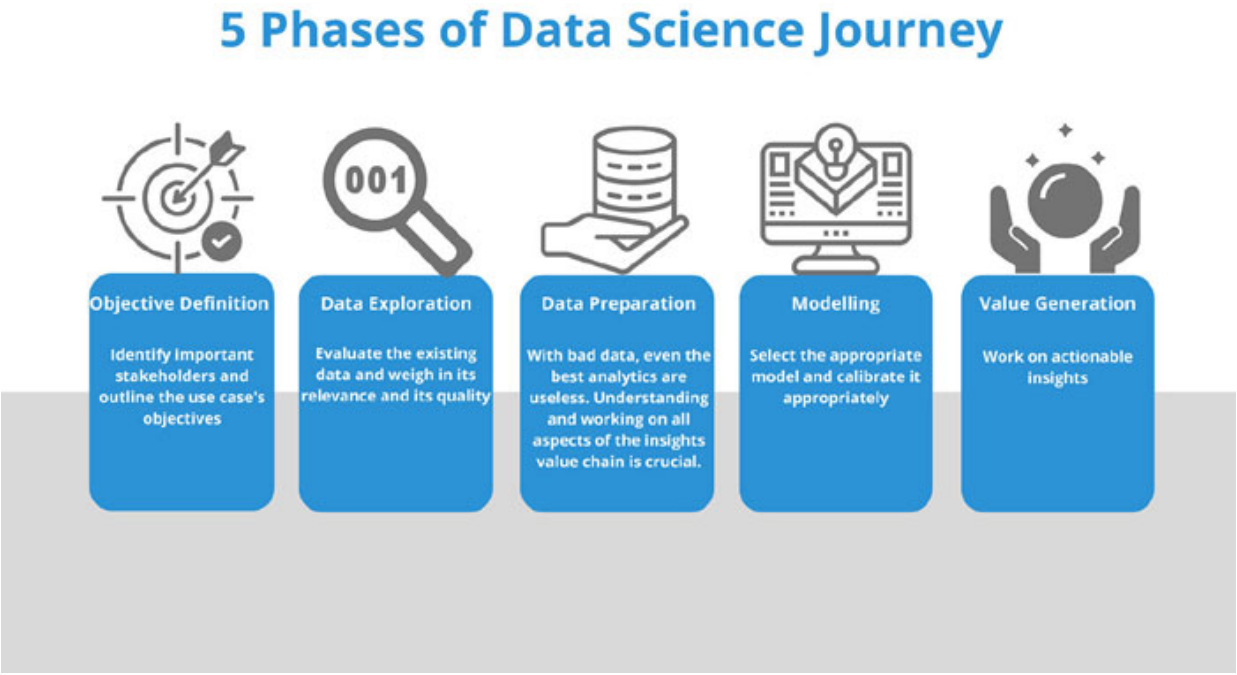


Figure 2.5: *Data Science Journey*

Numerous foundations throughout what we refer to as the insights value chain, which includes a diverse set of technical capabilities and strong business procedures, are necessary to maximize the value of data. In general, maximizing the value of a wealth of potential data begins with excellence in identifying, collecting, and storing that data; continues with technical capability for analyzing and visualizing that data; and concludes with an organization that can complement analytics with domain knowledge from a human and rely on a cross-functional, agile framework to implement pertinent analytics, as illustrated in the following image:

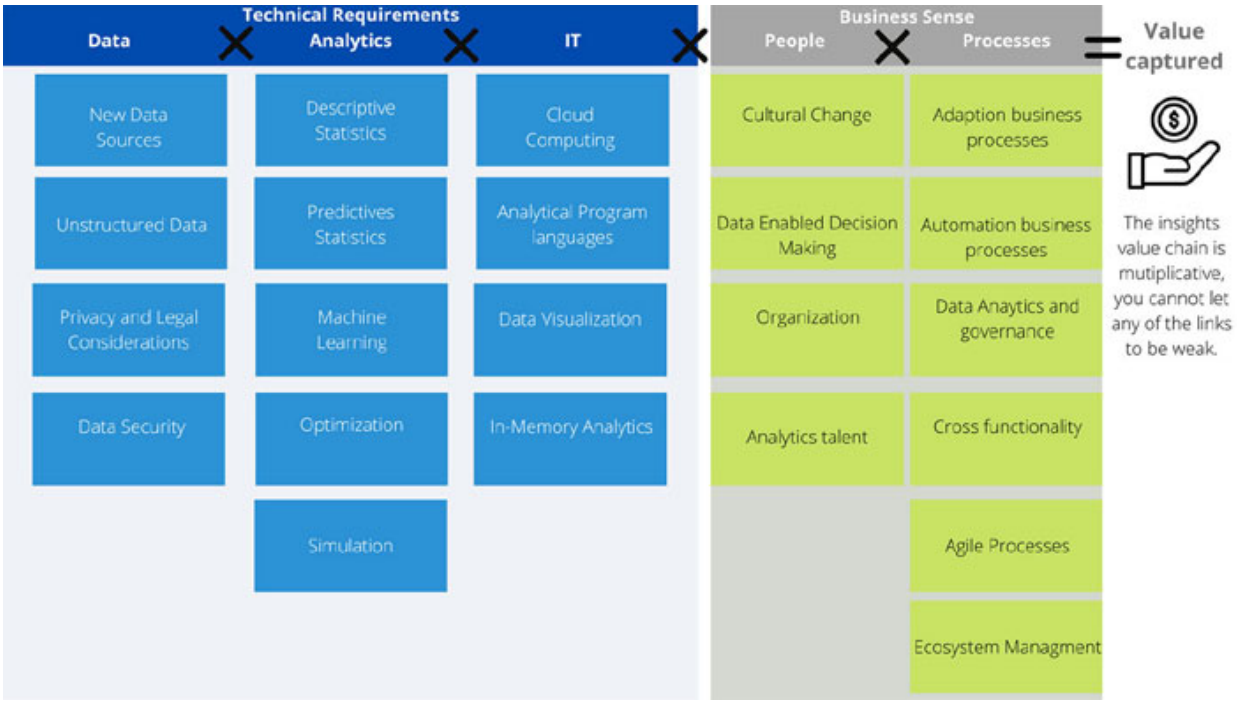


Figure 2.6: *Data Science Project Framework*

Mathematical concepts one must remember

Mathematics is at the heart of machine learning, which aids in the development of an algorithm that can learn from data and generate an accurate prediction. It might be as easy as identifying dogs or cats based on a set of images or recommending products to a consumer based on their previous purchases. As a result, a thorough understanding of the mathematical ideas underlying any central machine learning algorithm is critical. It assists you in selecting all the appropriate algorithms for your data science and machine learning initiatives. It is amazing to know that the shopping you do on Amazon passes through a lot of mathematical logics, isn't it?

Statistics, linear algebra, probability, and calculus are the four key principles that drive machine learning. While statistical ideas lie at the heart of all models, calculus aids in the learning and optimization of such models. When dealing with a large dataset, linear algebra comes in helpful, and probability aids in the prediction of future events.

- **Linear Algebra:** To design central machine learning algorithms, you must first learn how to form linear equations. These will be used to assess and monitor data collection efforts. Linear algebra is utilized in machine learning methods like loss functions, regularization, covariance matrices, **Singular Value Decomposition** (**SVD**), matrix operations, and support vector machine classification. It is also used in linear regression and other machine learning algorithms. These are the principles that are required to comprehend machine learning optimization strategies. We employ linear algebra to do Principal Component Analysis, which is a technique for reducing the dimension of data. In neural networks, linear algebra is employed extensively for network processing and representation.
- **Statistics:** When working with classifications like logistic regression, decision trees, hypothesis testing, and discrimination analysis, descriptive statistics is a critical skill for any aspiring data scientist to acquire. Combinatorics, Axioms, Bayes' Theorem, Variance and Expectation, Random Variables, Conditional, and Joint Distributions are only a few of the essential statistics for machine learning.
- **Probability:** It is plausible to conclude that probability is required to effectively complete a machine learning predictive modelling project. Machine learning relies on uncertainty, yet it is one of the aspects that newbies, especially those with programming experience, struggle with the most. There are three basic sources of uncertainty in machine learning: noisy data, inadequate coverage of the problem domain, and, of course, flawed models. We can, however, estimate the answer to the problem using the appropriate probability techniques. Hypothesis testing and distributions like the Gaussian distribution and the probability density function require probability.
- **Calculus:** Many students who disliked calculus in school will be surprised to hear that it is an essential component of machine learning. Fortunately, you may not need to master calculus; all that is required is for you to learn and comprehend its basics. During model construction, you will also need to comprehend the practical applications of machine learning using calculus. Therefore, if you understand how the derivative of a function gives the rate of change in calculus, you will

comprehend the idea of gradient descent. In gradient descent, we need to find the local minima of a function. Unless you start from numerous locations, a gradient descent may locate a local minimum rather than a global minimum if you have saddle points or several minima. Differential and Integral Calculus, Partial Derivatives, Vector-Valued Functions, and Directional Gradients are a few of the subjects that one must master in order to ace the data science calculus part.

The following image gives a picture of the importance of mathematics in data science. Statistical concepts like descriptive and inferential statistics, linear algebra and calculus will enable you to understand and build better models:

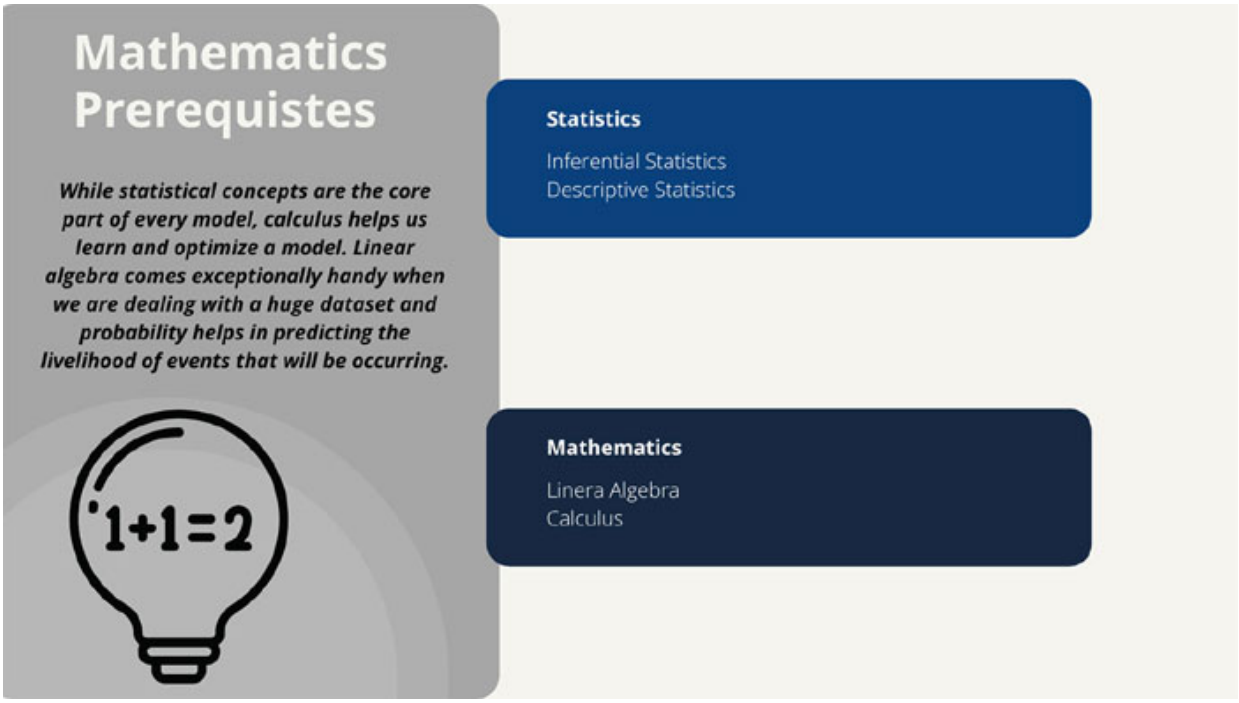


Figure 2.7: *Data Science Project Framework*

Conclusion

Let us recall what we covered so far in this chapter. The various types of data and the corresponding characteristics, the importance of data preparation and the strategy that can used for data preparation were discussed. Finally, we looked at the five phases of data science journey and the importance of mathematics.

In the next chapter, we will look at the various tools available in the market for data analytics.

Points to remember

- The most common data forms or types are as follows:
  - **Ordinal**: For example, Likert scale
  - **Nominal**: For example, gender, region
  - **Ratio**: For example, BMI
  - **Interval**: For example, temperature in Celsius
- The data science journey has the following five steps:
  - Define your objective and frame your problem statement
  - Prepare the data through data cleansing, metadata creation, data transformation, data standardization, and data augmentation
  - Explore the data (exploratory data analysis - EDA)
  - Data modeling, the core part where you use various machine learning algorithms
  - Value generation, the last and most important part
- Having a clear understanding of mathematical topics like linear algebra, statistics, probability, and calculus is essential

Multiple choice questions

1. Pick all that are ratio variables:
   a. 41-50 kilograms
   b. Male
   c. 5-6 feet
   d. Grade scale
2. What is data augmentation?
   a. Data augmentation is the process of synthesizing new data from existing data

b. It is the process of removing data that is inaccurate or erroneous in some or the other way

c. It is the process of making the data uniform and consistent

d. None

3. What is creating tags that provide information about the data known as?

a. Nominal data

b. Meta data

c. Ordinal data

d. Raw data

Answers

Question Number	Answer
1	a
2	a
3	b

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 3

Up and Running With Analytical Tools

Data analysis is the process of transforming raw data into meaningful statistics, insights, and explanations to make data-driven business choices. Data analysis has evolved into a critical component of current company operations. Choosing the finest data analytics tool is a difficult undertaking because no product satisfies every need. Every tool that you consider needs to suit one key criteria: “Will it fix my problem”. Hence, the one-size-fits-all approach will not be useful when it comes to data analytics. Before analysing the various tools, there are a few factors to consider. You should understand the sort of data your company wishes to evaluate and, as a result, your data integration needs.

Structure

In this chapter, we will cover the following topics:

- The different types of easy-to-learn and simple-to-implement analytical tools and hardware requirements
- Python workbook and **Auto Machine Learning** (**Auto ML**) libraries, and using their key features to solve business problems
- Steps to use analytical tools

Objectives

After going through this chapter, you’ll be able to employ appropriate data analytics approaches for addressing your most pressing business issue more quickly. You will be able to use various data analytics tools and auto ML frameworks to solve crucial business challenges. Finally, you will be able to perform predictive modelling without any prior coding expertise. Predictive modelling activities like data cleansing and visualisation may be automated using several open-source and commercial solutions. You will also understand how you can use Excel, Tableau, KNIME, Weka, Rapid Miner, Orange, and other tools to your advantage.

Analytical tools that matter and their hardware requirements

An enormous quantity of data sits beneath the inner workings of any firm. Data analytics tools are used to derive value from data in the form of insights that influence the customer experience, financials, and company decisions. Analysis tools combat information overload by combing through all your data, metrics and dimensions, and alerting you to any unexpected changes. Consequently, they aid you in posing questions you were unaware of, ensuring that you have no blind spots. A simple and easy-to-understand data to insight flow can be understood from the following image. It starts with data collection, followed by storage, visualization and finally, analysis/reporting.

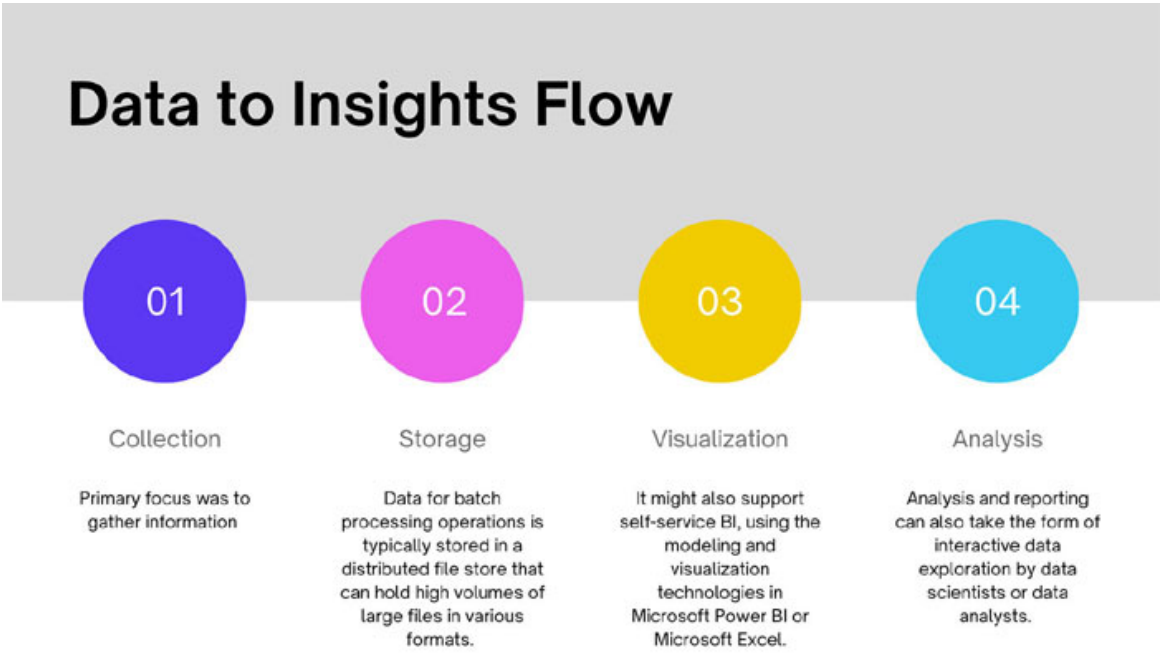


Figure 3.1: *Data to Insights Flow*

Auto analysis tools must have the following features:

- **High on fidelity**: Fidelity is a term used to describe the quality. These technologies are useless if they provide you with hundreds of insights every day because it would take all day to go through them. Seek out a product that summarises the most critical data.
- **Low on intensity**: The more personalization and customization these technologies entail, the less probable it is that they will expose facts you did not previously know.

When selecting a tool to employ, keep the following factors in mind:

- **Repurposed tools**: Numerous technologies originally developed to monitor manufacturing processes or devices have been "*repurposed*" for business analysis automation in recent years. They are referred to be "*Anomaly Detection*" tools or called by similar terms, but they were not designed for business data and hence, deliver a poor user experience.
- **False automation**: Numerous data visualization tools have sought to increase their perceived value (and thus, their pricing) by claiming the ability to analyse your data automatically. In actuality, they still necessitate a person's exploration and discovery of insights; it does not happen automatically. This may be appropriate for you, but it is not the same as automated analysis.

Let us start with the most basic non-programming tools:

- **Excel or Spreadsheet**: If you're new to data science or have been doing it for a while, you'll know that even after all these years, Excel is still an essential part of the analytics business. Even now, this software is used to tackle most difficulties encountered in analytics projects. With more community support, tutorials, and free resources than ever before, mastering this technology has become very easy. It supports all major functions, such as summarizing and visualizing data, data wrangling, and so on, and it is powerful enough to investigate data from all perspectives.
- **KNIME**: KNIME provides an open-source analytics platform for data analysis, which can later be implemented and scaled with the help of other KNIME solutions. This program provides a plethora of data mixing and visualization tools, and it also offers powerful machine learning techniques. Yes, you can create models with this program.

  KNIME Analytics Platform is the most powerful and free platform for drag-and-drop analytics, machine learning, statistics, and ETL. All major providers include connectors to data sources (both on-premise and in the cloud), making it simple to move data between settings. SQL Server to Azure, Google Sheets to Amazon Redshift, and so on. It is also possible to use a machine learning algorithm and filter/transform the output.
- **WEKA**: Weka is a library of machine learning algorithms for use in data mining jobs. The algorithms can be applied on a dataset directly. Weka includes data pre-processing, classification, regression, clustering, association rules, and visualization tools. As it is a machine learning tool, its interface is simple enough for you to perform machine learning task quickly.

It offers data pre-processing, classification, regression, clustering, association rules, and visualization choices. Weka can handle the majority of the stages you might consider when creating a model. With its graphical user interface, it provides the simplest entry point into the realm of Data Science for someone who hasn't coded in a while. And the library is written in Java, so those with Java experience can include it in their own code.

- **Rapid Miner**: In the 2016 Gartner Magic Quadrant for Advanced Analytics, this tool emerged as a leader. For the fourth time in a row, RapidMiner was named a Gartner Peer Insights Customers' Choice for Data Science and Machine Learning Platforms in August 2021. It broadens its experience in developing machine learning models.

  It is more than just a graphical user interface; it provides support for model builders that use Python and R. Rapid Miner has made significant progress in the AI world as it is commonly used by non-programmers and researchers.

  Rapid Miner has its own dataset collection, but it also offers the opportunity to build up a cloud database for storing massive amounts of data. Data can be stored and loaded from Hadoop, the cloud, RDBMS, NoSQL, and other sources. Apart from that, you can quickly load your CSV data and begin utilizing it. Drag-and-drop choices allow for the conventional implementation of procedures like data cleansing, visualization, and pre-processing without the need to write a single line of code.

- **Orange**: This application is intended to provide interactive data visualizations and data mining jobs. There are numerous YouTube tutorials available to help you understand how to use this tool. It offers a large data mining task library that includes all classification, regression, and clustering methods. Orange enables you to:

  - Display a data table and choose features,
  - Contrast learning algorithms and predictions, and
  - Consider visualizing data items.

  Developers frequently claim that Orange is more engaging than other tools and has a playful mood that makes normally dull and tedious analytics fascinating. It immediately formats the data into a widget-movable pattern.

  Orange widgets have been created by a graphical programmer to support Orange's data processing and machine learning approaches. They include for data entry and pre-processing, classification, regression and association

rules, and a set of widgets for model assessment and the mental image of assessment findings. The user will interact with visualizations or feed the selected collection into other widgets.

- **Data**: These are the widgets for information input and data filtering.
- **Visualize**: These are the widgets for common and variable mental images.
- **Classify**: This is a set of algorithms for classification.
- **Regression**: This might be a collection of supervised machine learning regression techniques.
- **Examine**: These are the cross-validation and sampling-based processes.
- **Unsupervised**: These unsupervised algorithms are used for clump and information projection techniques.

In a production scenario, the best analytics performance may necessitate more resources than the minimum specification. The following image will give you an idea about the easy-to-use analytics tools (no or low code platforms) available in the market:

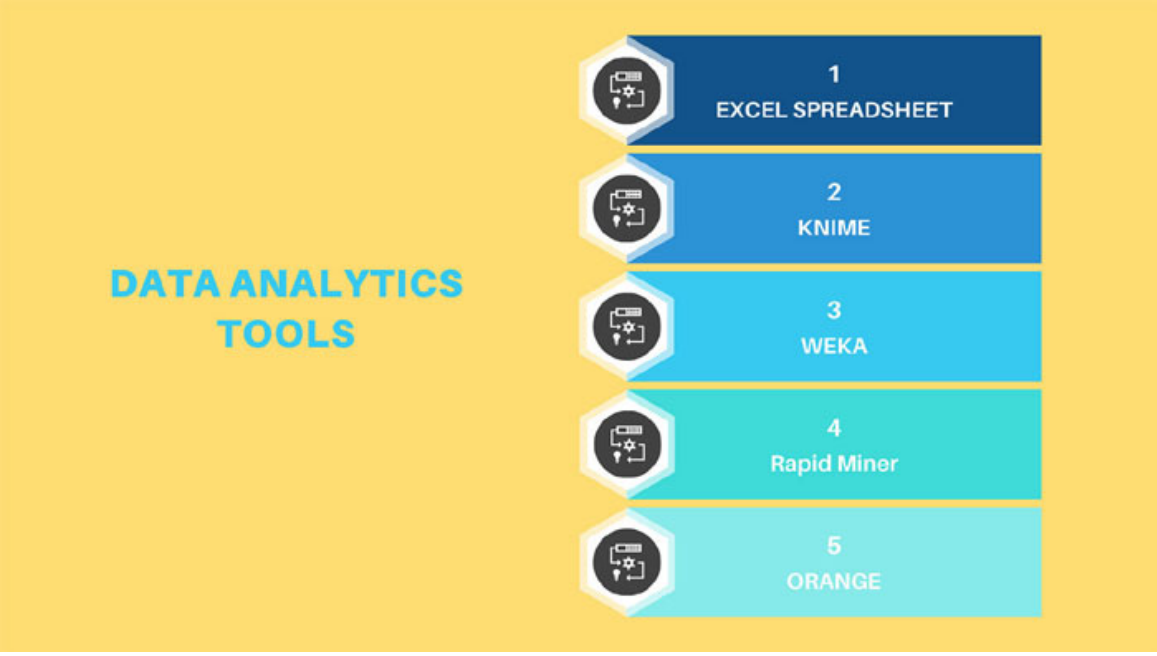


Figure 3.2: *Data Analytics Tools*

Though these analytics tools are easy to use, you might have to ensure that you meet certain basic system requirements. The following table will guide you with

the minimum recommendation in terms of computational power:

Component	Minimum	Recommendation
Processor	1.8 GHz	
Memory (RAM)	2 GB	• 64-bit operating systems: 8 GB or more, especially if sorting large files • 32-bit operating systems: 4 GB, especially if sorting large files
Hard disk space (Analytics application files)	1.1 GB	
Hard disk space (software prerequisites)	8 GB	
Hard disk space (data storage)	100 GB or more	In addition to the hard disk space required to install Analytics application files and prerequisites, significant additional space is required if a computer will be used to store data extracts, flat files, and results.

Table 3.1: *Prerequisites for using Analytics Tools*

Python workbook and auto ML libraries

Jupyter Notebook is an interactive computing platform that allows users to create notebook documents, such as real-time programming, interactive widgets, plots, equations, captions, text, images, and videos. These documents give a comprehensive and self-contained record of a computation that may be translated to different forms and shared via email, dropbox, version control systems (such as git/GitHub), or nbviewer.jupyter.org.

The Jupyter Notebook is made up of three parts:

- **The notebook web application**: It is an interactive web application that allows you to write and run code interactively while also producing notebook documents.
- **Kernels**: These are independent processes launched by the notebook web application that run users' code in a specified language and provide results to the notebook web application. The kernel also supports interactive widget computations, tab completion, and introspection.
- **Notebook documents**: These are self-contained documents that provide a representation of all information accessible in the notebook web application. Every notebook entry has its own kernel.

Jupyter Notebooks are a fork of the IPython project, which had its own IPython Notebook project. Jupyter is named for the three primary programming languages that it appears to support: Julia, Python, and R. Jupyter comes with the IPython kernel, which allows you to develop Python programs, although there are presently over 100 alternative kernels available. Auto ML libraries allow non-experts to use ML models and techniques without getting into the depth of machine learning algorithms.

Auto ML will take care of data pre-processing, manual model selection, hyperparameter tuning, and model evaluation chores, and eventually, models will be optimized and ready for prediction without prompting the user. It is all about making Machine Learning jobs easier to utilize with less code. To create strong models, machine learning model building necessitates domain knowledge, sophisticated programming ability, mathematics, statistics, and resources. With Auto ML, you can save time and increase productivity, which can be clearly seen from the following image that shows the differences in ML process flow between Auto ML and traditional ML:

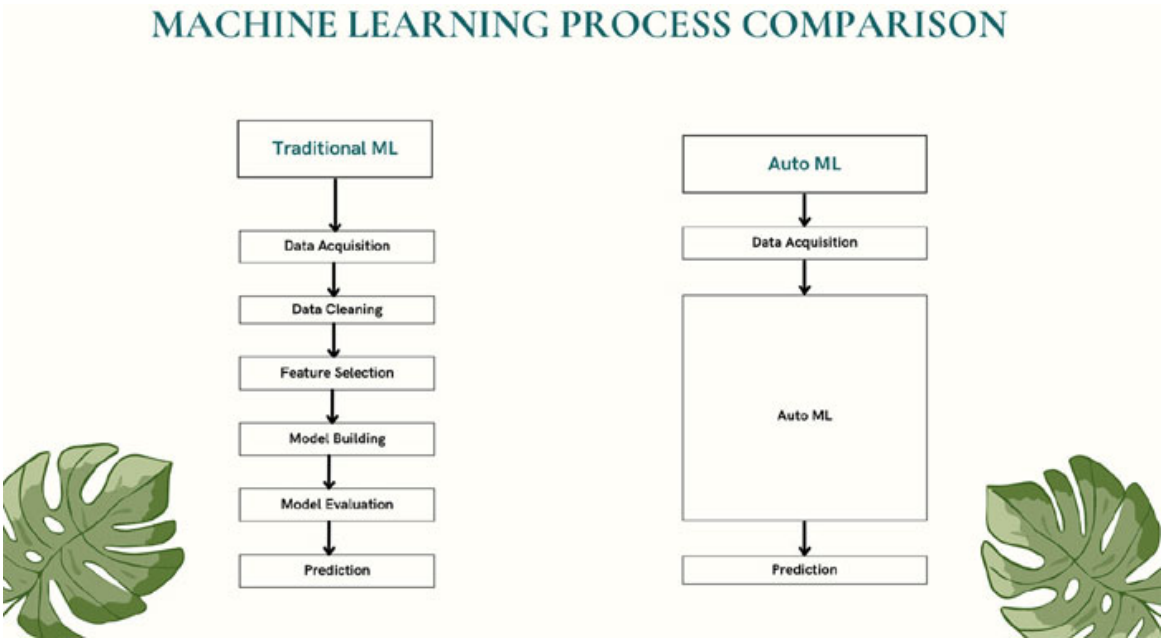


Figure 3.3: *Auto ML vs Traditional ML Process Flow*

The decision to bring Auto ML into a business area is based on the technical team and business stakeholders because there are both important benefits and inevitable drawbacks. Furthermore, there is little doubt that Auto ML enables everyone to use ML features far more quickly and efficiently than actual programming. "*Time-saving, organized execution, and increased development activity*" are crucial factors for the necessity of Auto ML, and we can truly say YES to Auto ML. Keep the following Auto ML considerations in mind:

- **Supports Data Scientists**: Normally, data scientists must be involved in an end-to-end life cycle. However, the Auto ML platform can manage the required machine learning life cycle phases, which simplifies integration and productivity.
- **Routine job/task in the ML model life cycle**: By automating repetitive activities in the machine learning life cycle, instead of concentrating on the models, data scientists may concentrate on data storytelling, tracking success metrics, continuous deployment and integration of models, providing best practices, and directly helping customers save money.

Some of the most common Auto ML Libraries used in various industries are as follows:

- **H2O.ai**: H2O was formed in 2012 and offers an open-source package; in 2017, it introduced a commercial Auto ML service dubbed Driverless AI. It

has been widely implemented in various areas, including finance and retail. To enable less experienced data scientists to train and deploy the most accurate models possible, driverless AI (formerly known as H2O.ai) was designed to be a helpful digital assistant for data scientists of all skill levels. It also needs to be able to utilise best-of-breed open-source machine learning platforms on corporate hardware, both on premise and in the cloud, and deploy in minutes.

- **TPOT**: TPOT, or Tree-based Pipeline Optimization Tool, is a Python package for automated machine learning. TPOT represents a model pipeline for a predictive modelling problem using a tree-based structure, which includes data preparation and modelling techniques as well as model hyperparameters. TPOT's purpose is to automate the creation of ML pipelines by integrating a flexible expression tree representation of pipelines with stochastic search strategies like genetic programming, which can be understood from the following image as well:

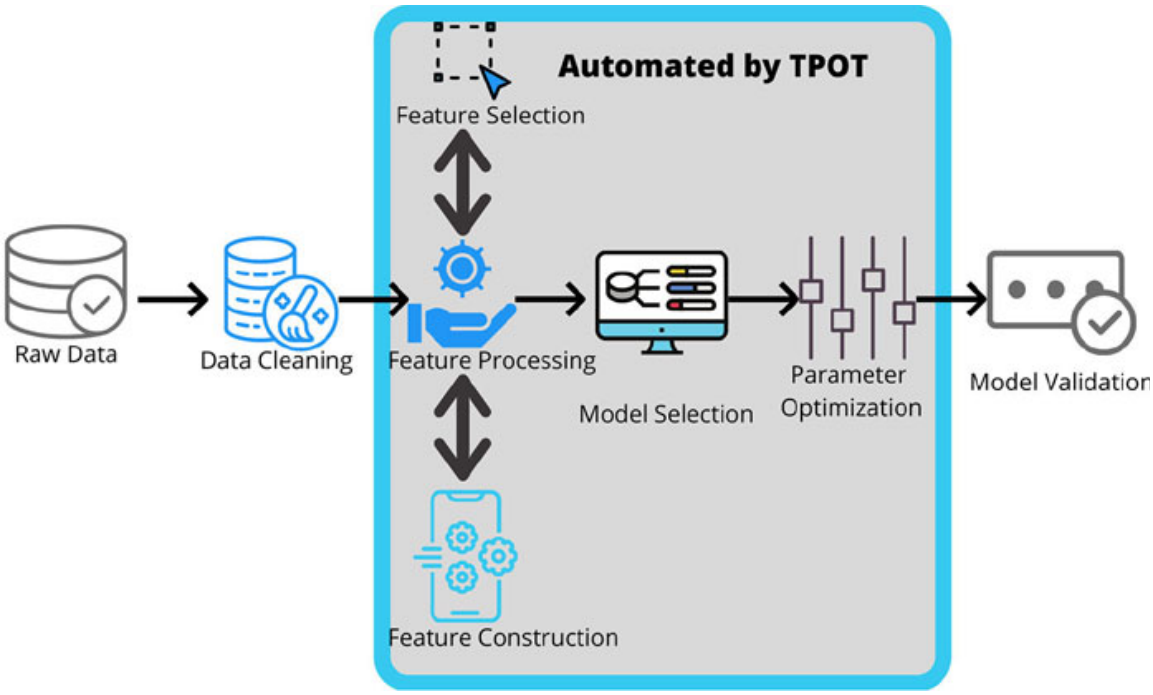


Figure 3.4: *Advantage of TPOT*

- **Google Cloud Auto ML**: Google Cloud Auto ML, a machine learning software suite, was unveiled in January 2018. As of now, it only has one publicly available product, i.e., Auto ML Vision, which is an API that identifies and classifies objects in images. Cloud Auto ML Vision uses transfer learning and neural architecture search. Transfer learning is a useful

approach that allows people with smaller datasets or less computational capability to get cutting-edge findings by leveraging pre-trained models built on similar, larger data sets. A model taught via transfer learning does not have to start from scratch, so it typically achieves better accuracy with less data and computation time. Many pictures contain underlying characteristics (such as corners, circles, dog faces, or wheels) that occur in many distinct kinds of images.

- **Microsoft Azure Auto ML**: Azure Automated ML enables enterprises to deploy ML models with built-in data science knowledge. A non-technical person with little knowledge of data science can also implement models using Automated ML. This method of deploying models reduces work, risk, and the time needed. Azure Automated ML enables businesses in many areas, including healthcare, finance, and banking, to harness ML and AI technology. During the training process, Azure Machine Learning generates a number of pipelines at the same time to forecast which ML algorithm is most suited to the underlying data. It also handles feature selection and other necessary pre-processing. The following figure shows how Auto ML automates multiple tasks to produce a model:

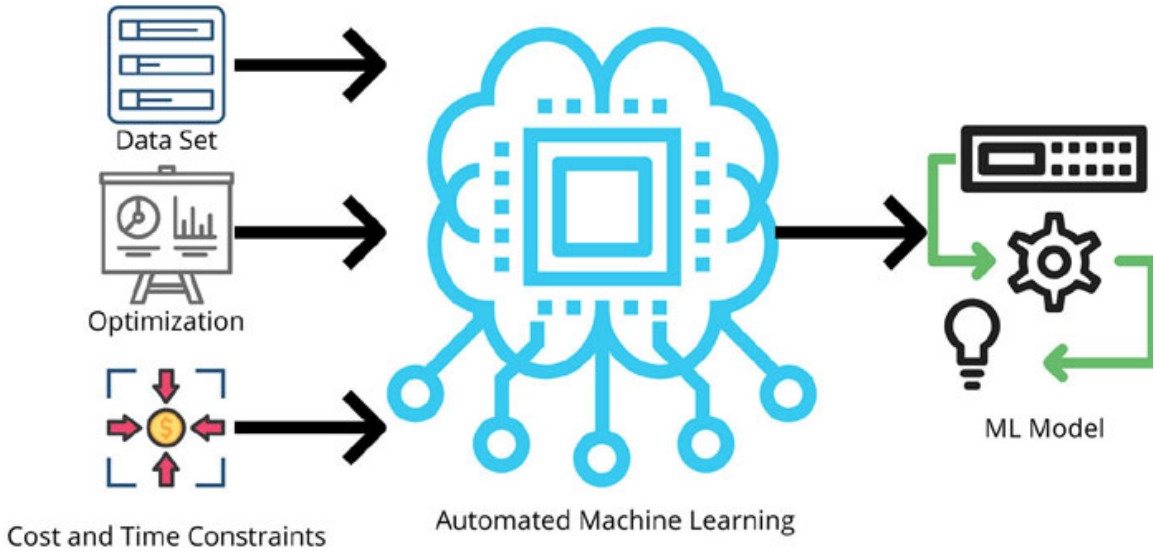


Figure 3.5: *Advantage of Auto ML*

There are certain challenges with Auto ML libraries, like data wrangling, feature engineering, model selection, etc., which are shown in the following figure:

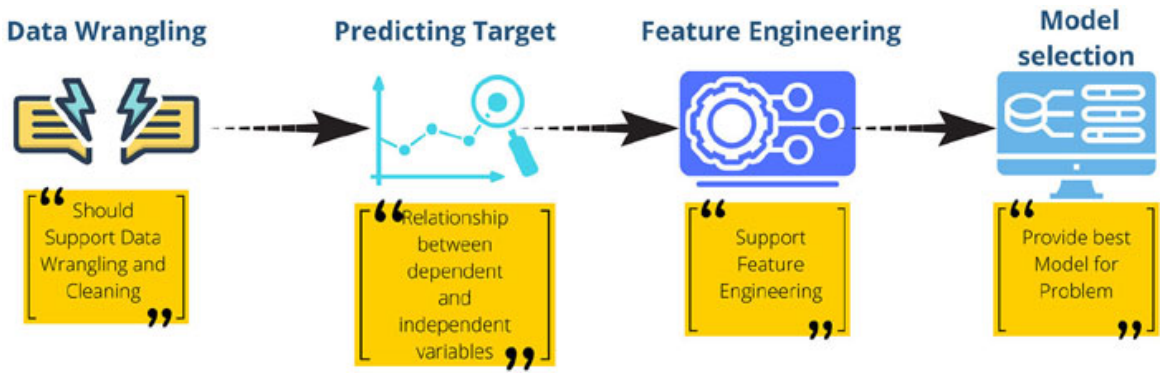


Figure 3.6: *Challenge Area with Auto ML*

Steps to use analytical tools

In this section, we will learn how to install, use and interpret the results of some of the most useful auto analytics tools.

Weka: Waikato University in New Zealand is developing Weka, a data mining/machine learning program. The user interface of Weka is simple and has the following options:

1. A basic **command line interface** (**CLI**) is given to run Weka functions directly.
2. Explorer is a data discovery environment.
3. Experimenter is a platform for conducting experiments and statistical tests on different learning methods.
4. Knowledge flow is a Java-Beans-based interface for machine learning and tuning experiments.

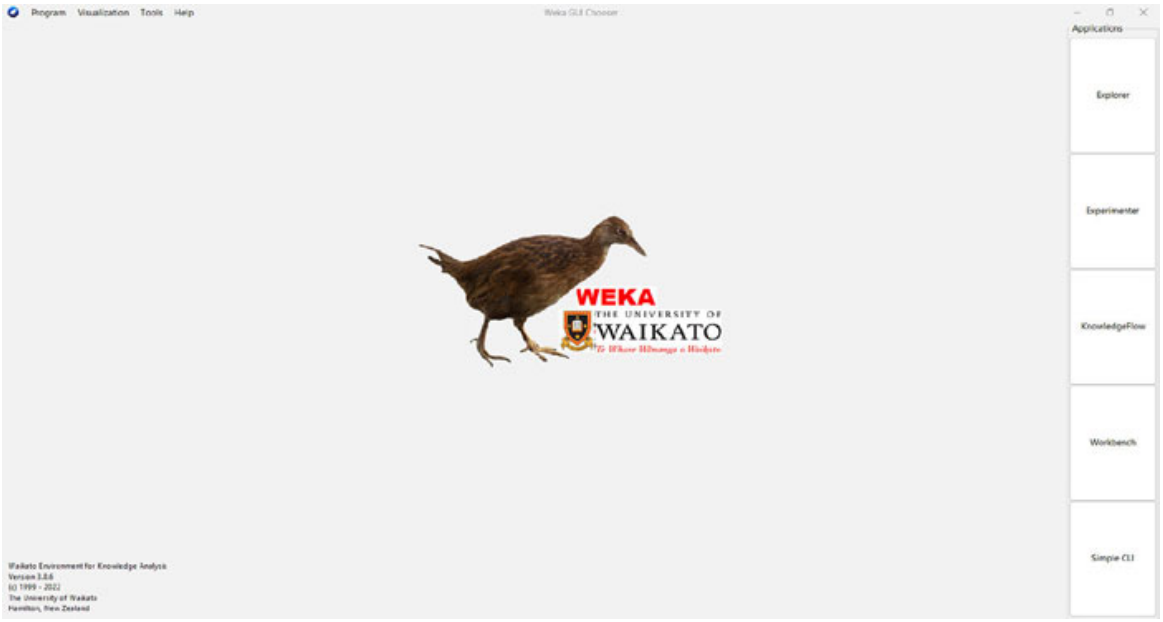


Figure 3.7: *Weka User Interface*

In this section, we will use the “Explorer” for the experiments. We will start with the pre-processing step because as learnt in the previous chapters, data will not be in ideal state for us to perform the experiments that we want to.

Weka makes pre-processing simple. You can simply click on the “Open file” button to load your file as one of the following file types: Arff, CSV, C4.5, binary, LIBSVM, XRFF; you can even load SQL database files through URL and then apply filters to them, as shown in the following image:

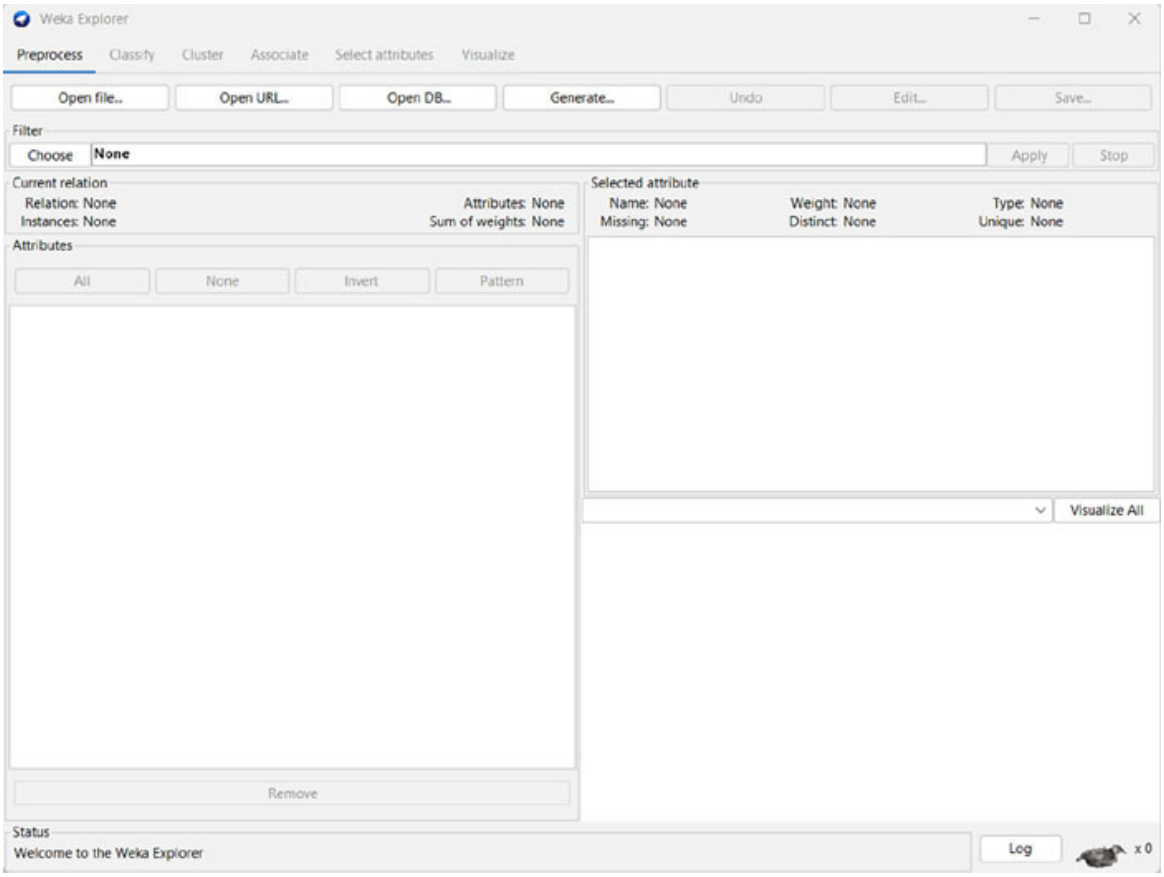


Figure 3.8: *Weka Data Pre-Processing*

Remember, if you have an **.xls** file, you will have to convert it to a .csv file before loading it. Once you have converted the file into .csv form, take the following steps before loading the file. A sample raw data file can be similar to the following screenshot:

	A	B	C	D	E
1	sepal.length	sepal.width	petal.length	petal.width	variety
2	5.1	3.5	1.4	0.2	Setosa
3	4.9	3	1.4	0.2	Setosa
4	4.7	3.2	1.3	0.2	Setosa
5	4.6	3.1	1.5	0.2	Setosa
6	5	3.6	1.4	0.2	Setosa
7	5.4	3.9	1.7	0.4	Setosa
8	4.6	3.4	1.4	0.3	Setosa
9	5	3.4	1.5	0.2	Setosa
10	4.4	2.9	1.4	0.2	Setosa
11	4.9	3.1	1.5	0.1	Setosa
12	5.4	3.7	1.5	0.2	Setosa
13	4.8	3.4	1.6	0.2	Setosa
14	4.8	3	1.4	0.1	Setosa
15	4.3	3	1.1	0.1	Setosa
16	5.8	4	1.2	0.2	Setosa
17	5.7	4.4	1.5	0.4	Setosa
18	5.4	3.9	1.3	0.4	Setosa
19	5.1	3.5	1.4	0.3	Setosa
20	5.7	3.8	1.7	0.3	Setosa
21	5.1	3.8	1.5	0.3	Setosa
22	5.4	3.4	1.7	0.2	Setosa
23	5.1	3.7	1.5	0.4	Setosa
24	4.6	3.6	1	0.2	Setosa
25	5.1	3.3	1.7	0.5	Setosa
26	4.8	3.4	1.9	0.2	Setosa
27	5	3	1.6	0.2	Setosa
28	5	3.4	1.6	0.4	Setosa

Figure 3.9: *Raw Data File*

- Open your CSV file in any text editor and add **@RELATION** database name to the first row.
- Use the following definition to add attributes: **@ATTRIBUTE** attr name.
- If attr type is a number, it should be defined as REAL; otherwise, values must be added between curly parentheses.
- Finally, add a **@DATA** tag right above your data rows.

- Then, save your file with the **extension .arff**. The following figure illustrates this:

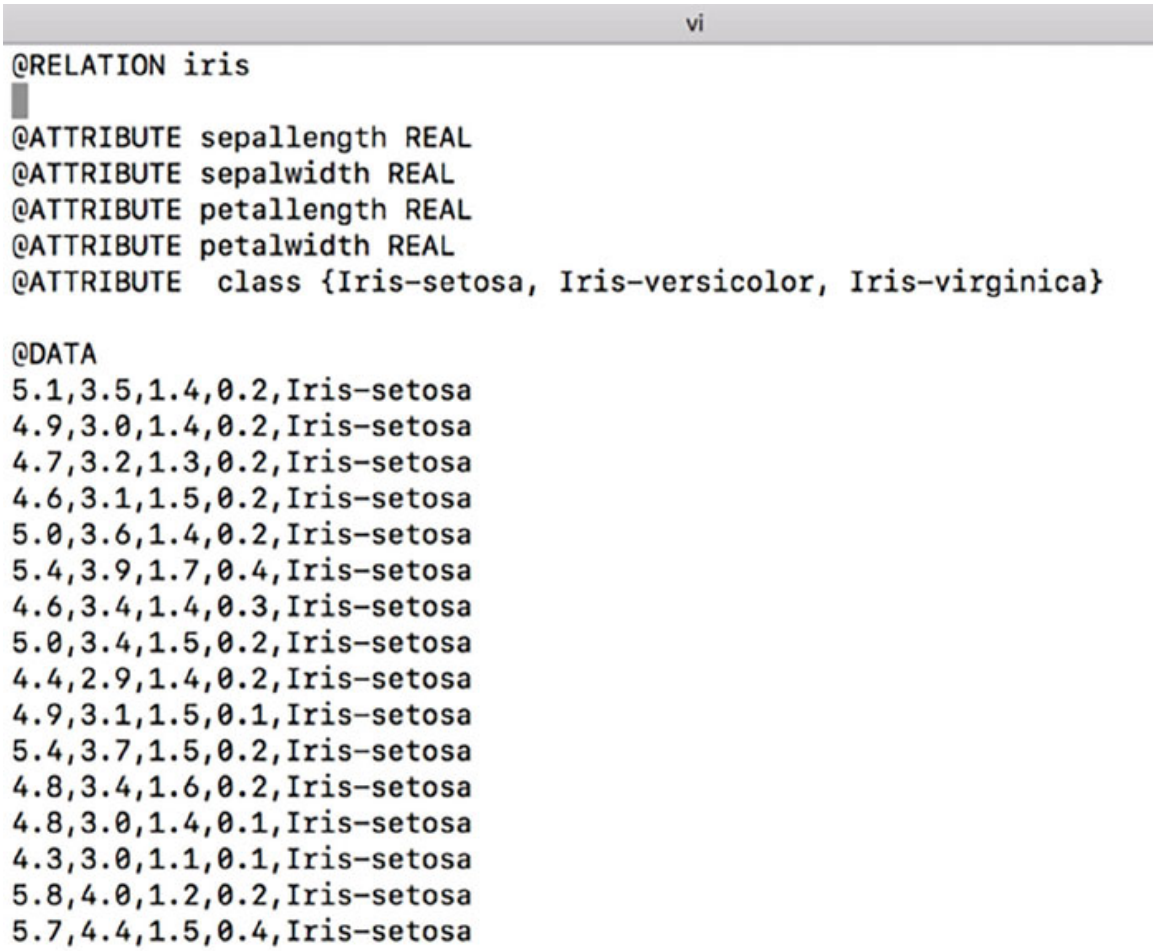


Figure 3.10: *Weka Data Load*

After pre-processing comes loading:

In the Pre-process section, click on the "**Open file**" button to load your **.arff** file from your local file system. Don't worry if you couldn't convert your .csv file to **.arff** since Weka will do it for you. Voila! Now you can see a summary of the raw data as a visual, as shown in the following figure:

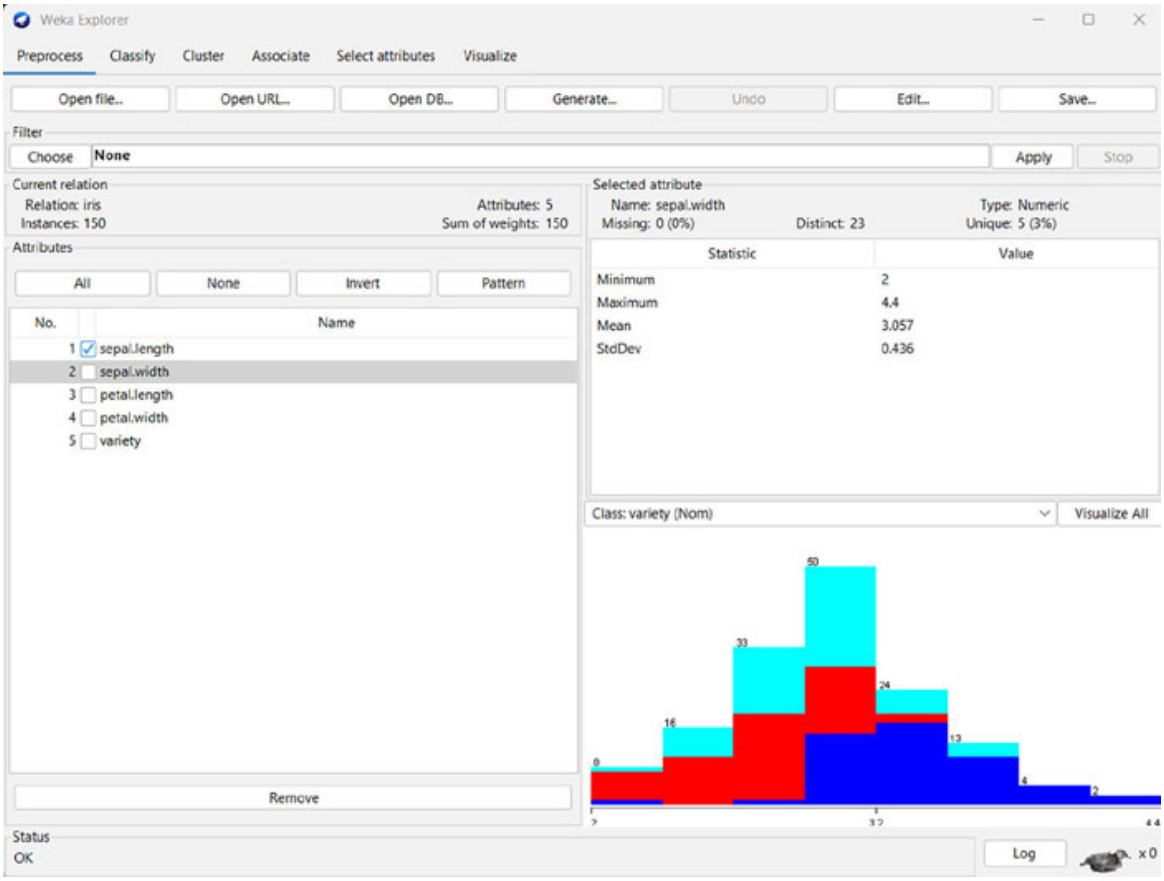


Figure 3.11: *Weka Data Summary Window*

In Weka, the pre-process stage is called Filter, and you can apply any filter you want by clicking on the **`Choose`** option from Filter. If you want to utilize Association Rule Mining as a training model, for example, you must separate the numeric and continuous attributes. You can do so by following the steps here: Select | **`Filter`** | **`Supervised`** | **`Attribute`** | **`Critique`**.

After the pre-processing and loading step, let us now move on to Classification step:

- Classification refers to distributing data among the various classes defines on a data set.
- Classification algorithms learn this form of distribution from a given set of training data and then attempt to accurately categorize it when presented with test data for which the class is not provided. The values on the dataset that identify these classes are labelled and utilized to determine the class of data to be presented during the prediction.

- For the purpose of understanding classification, we will use the iris dataset, and this data set can be easily downloaded using the link here: **Iris**
- The Iris dataset does not need any pre-processing; we can directly move to the classification step.
- After loading the dataset, click on the **`Classify`** section, where you can switch to another window; this can be seen in the following figure:

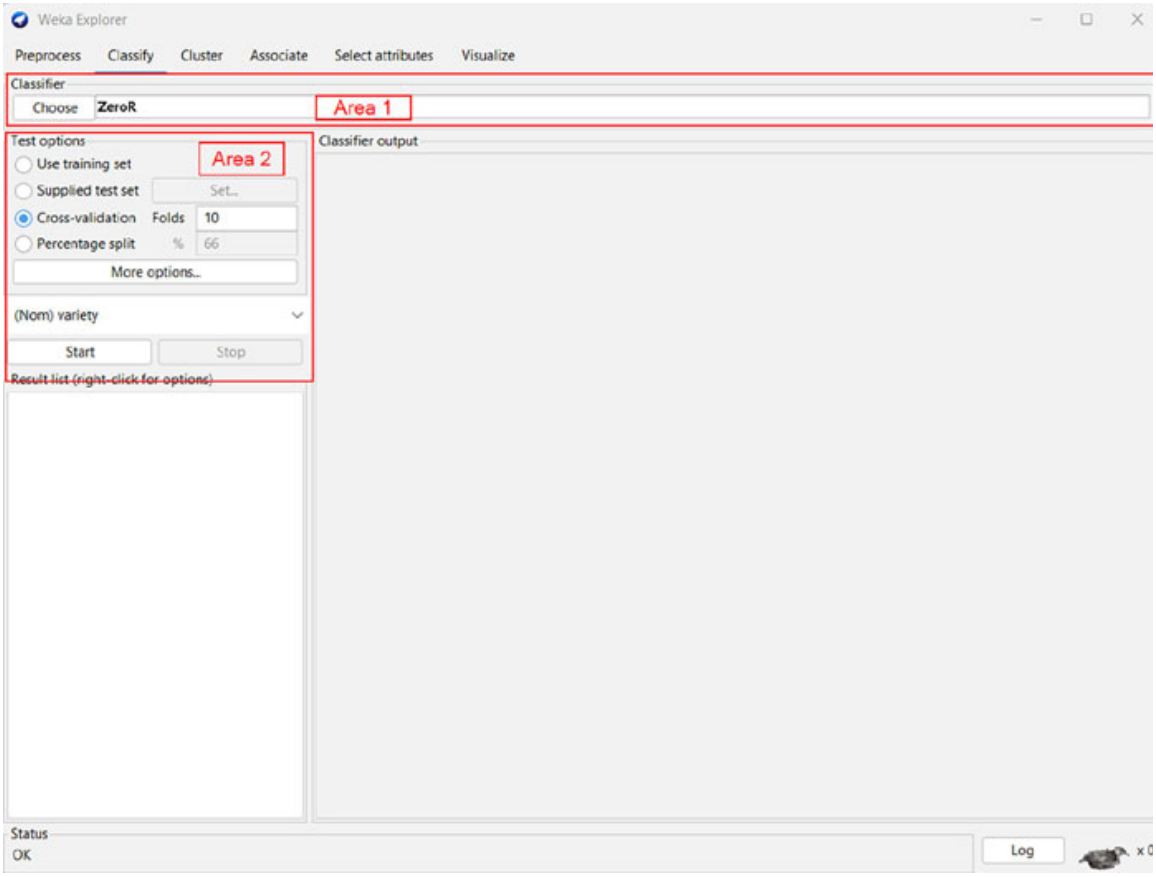


Figure 3.12: *Weka Split Data for Training & Testing Model*

- In the Classify section, as shown in Area 1 of the preceding image, ZeroR is the default classifier for Weka. However, because the ZeroR method performs poorly on the Iris dataset, we will replace it with the J48 algorithm, which has a high success rate on our dataset. A new algorithm can be picked from the list by clicking on the Choose button in Area 1 of the preceding figure. In the Classifier list, the J48 algorithm is located within the trees directory. We must first select the test alternatives from Area 2 before executing the algorithm.

  There are four test options:

- **Use training set**: Classifies your model depending on the dataset with which it was originally developed
- **Supplied test set**: Controls how your model is categorised based on the external dataset you supply; by clicking on the Set button, you may select a dataset file
- **Cross-validation**: A popular choice, especially if you have a small number of datasets; the number you specify in the Fold area is used to divide your dataset into Fold numbers (for example, 10). The original dataset is divided into 10 parts at random. Then, for the first training, Weka utilizes set 1 for testing and 9 sets for training, then set 2 for testing and the remaining sets for training, and so on for a total of 10 times, incrementing the testing set number each time. Finally, the user is informed of the average success rate.
- **Percentage split**: Using the number you enter, divide your dataset into train and test. By default, the percentage value is 66 percent, which implies that 66 percent of your dataset will be utilized as a training set, while the remaining 33 percent will be used as a test set.

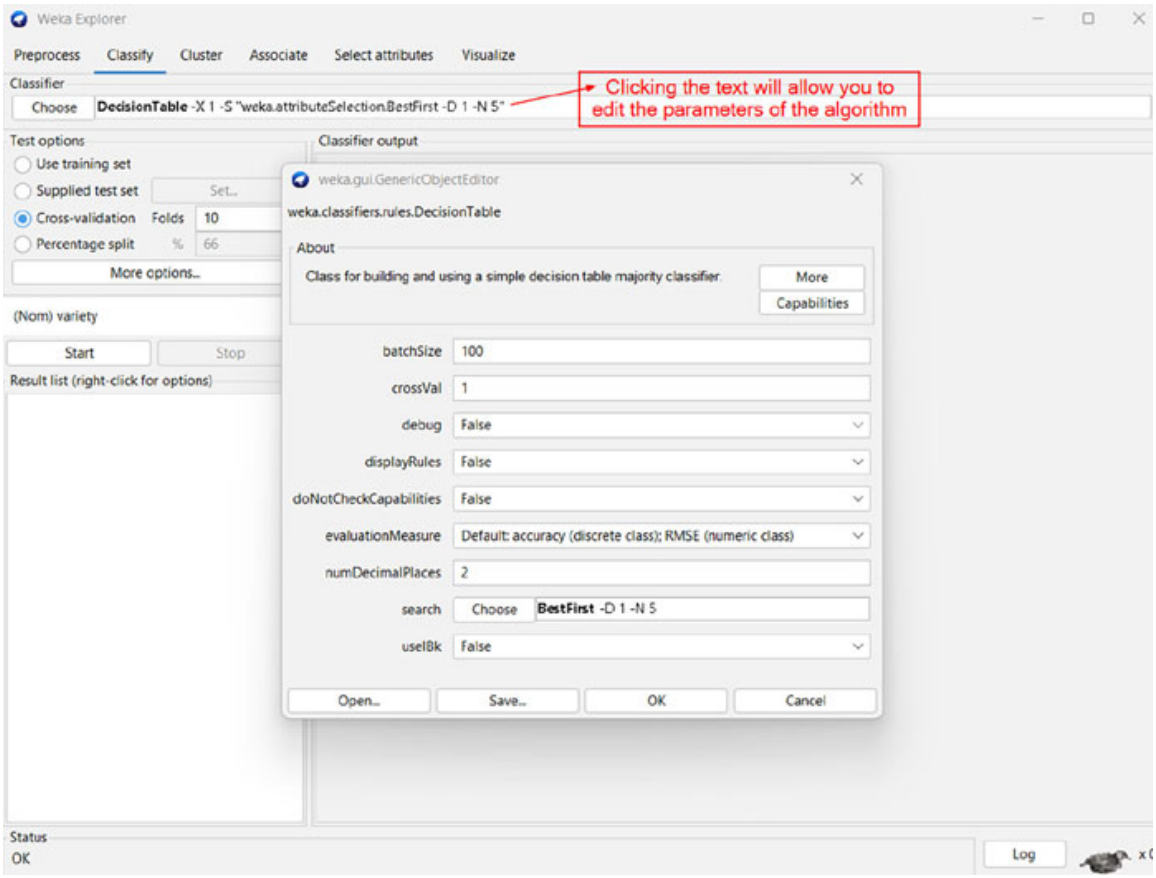


Figure 3.13: *Weka Model Training*

- The J48 algorithm was used to do 10-fold cross validation, which was chosen from the Test Options shown in the preceding image. The class feature was selected from the dropdown box, and then the “Start” button just above Area 2 in the following figure was clicked on. According to the results, the success rate is 96 percent, as shown by the Classifier Output in Area 1 in the following screenshot:

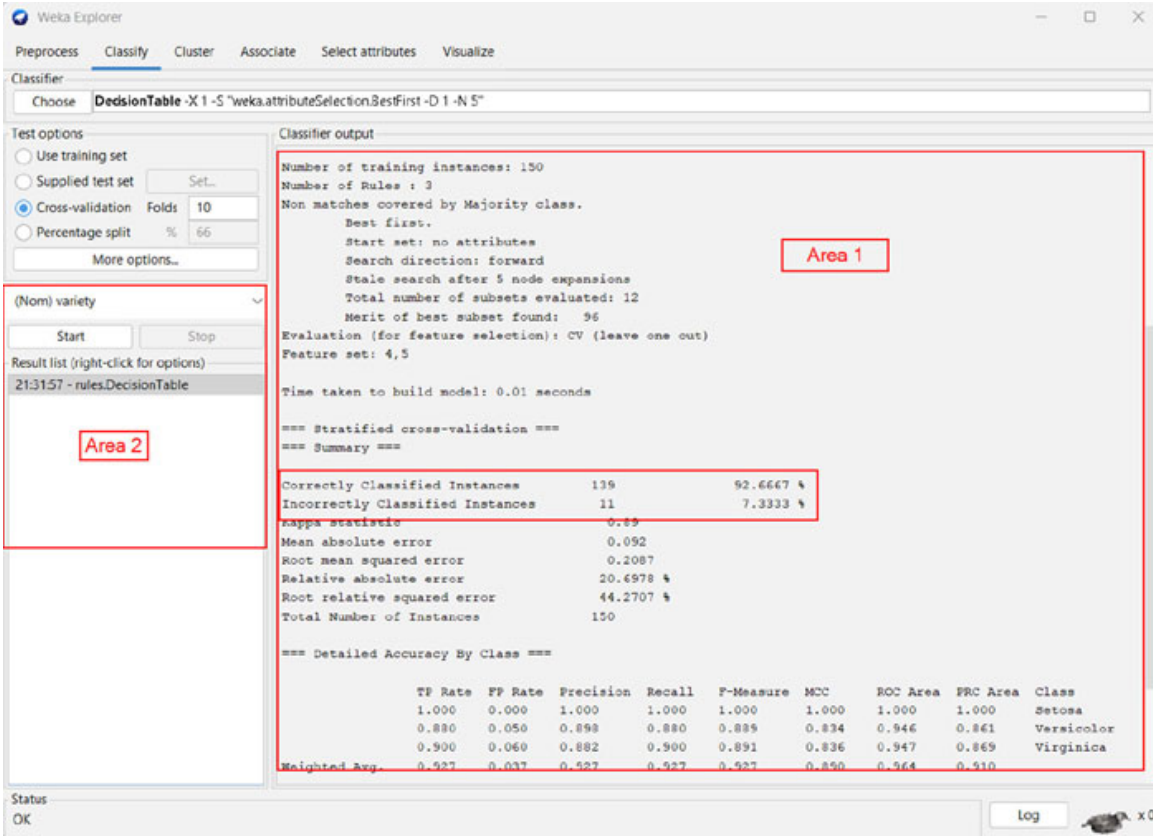


Figure 3.14: *Weka Model Output Summary*

As seen in the preceding figure, running information in Area 1 will yield detailed results. It is divided into five sections, the first of which is Run Information, which contains specific information about the dataset and the model you employed.

- As shown in the preceding figure, we utilized J48 as a classification model, our dataset was Iris, and its features were sepal length, sepal width, petal length, petal width, and class.
- We use 10-fold cross-validation as our testing method. Our model generated a pruned tree because J48 is a decision tree. It’s a tree like model that uses

the most important characteristics and features, the data is divided recursively into subsets until each subset mostly contains one type. The goal of the technique is to generate the shortest possible classification or prediction tree given a set of input.

- As seen on the tree, the first branching occurred on petal length, which displays the petal length of the flowers; if the value is less than or equal to 0.6, the species is Iris-setosa; otherwise, another branch checks another specification to determine the species.
- The class label is represented by the symbol ‘:’ in a tree structure.
- The classifier model section depicts the model as a tree and provides information about the tree, such as the number of leaves and the size of the tree.
- The stratified cross-validation section follows, and it displays the error rates. This section tells you how successful your model is. For example, our model properly identified 96% of the training data, and our mean absolute error rate is 0.035, which is acceptable based on the Iris dataset and our model.

At the bottom of the report, you’ll find a confusion matrix and a thorough accuracy table. The details of these terminologies will be discussed in the upcoming *chapters 8*.

After the classification step, let us move to visualizing the result section. For the graphical representation of the results, use the option shown in the following figure:

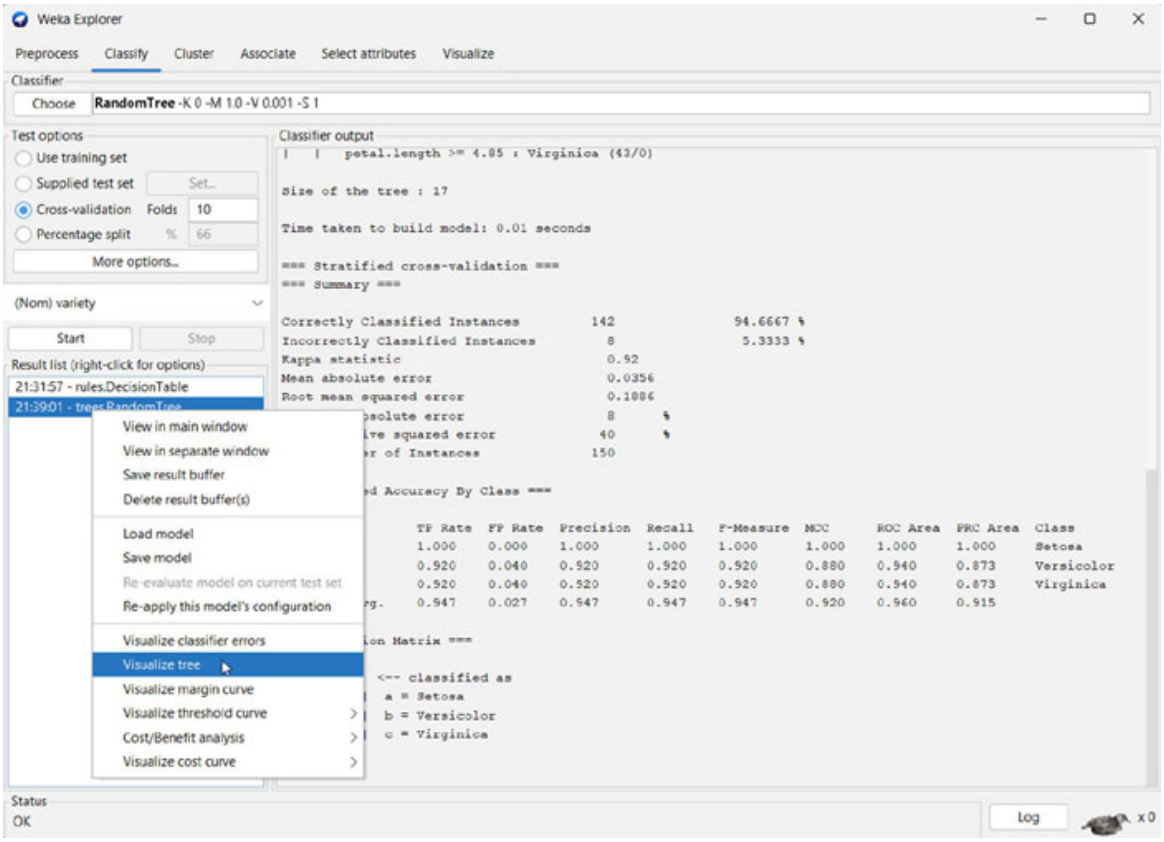


Figure 3.15: *Weka Visualize Model Option*

By right-clicking on visualize tree, you may see your model's illustration, as shown in the following figure:

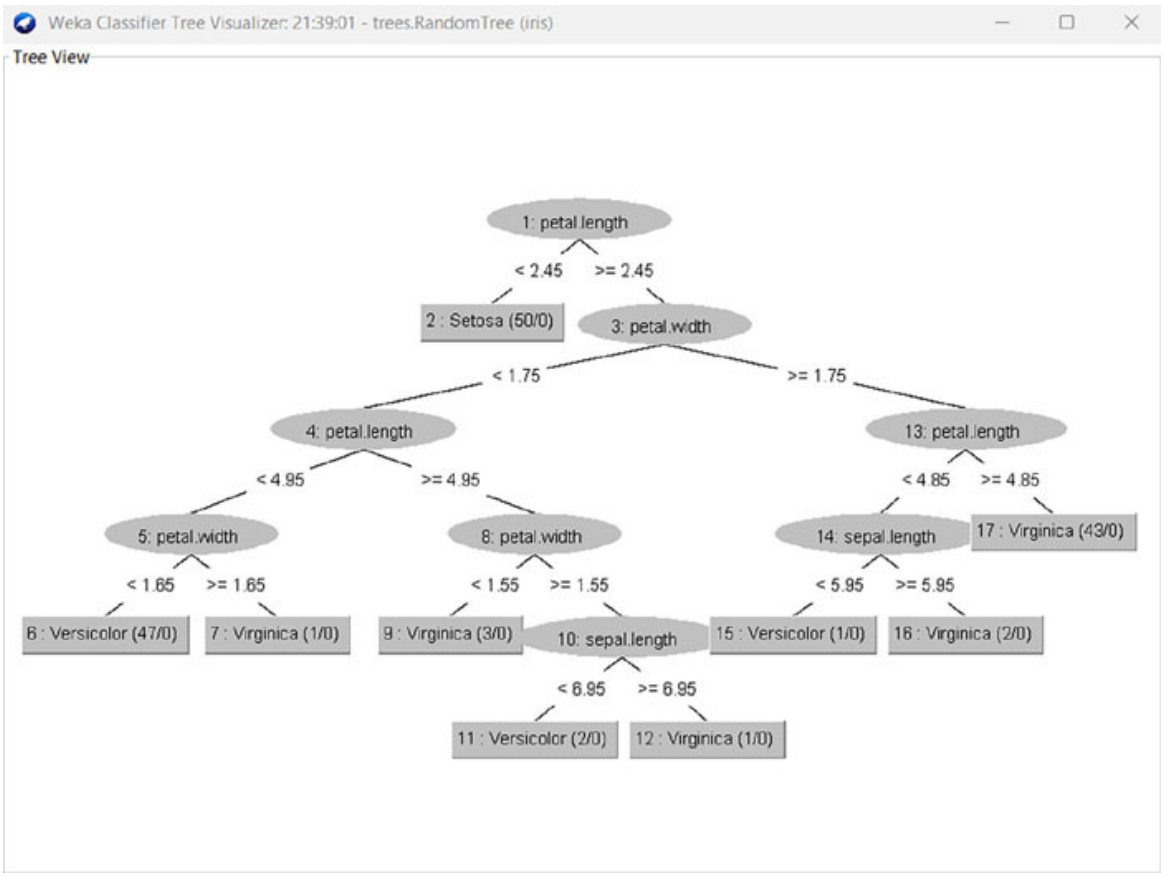


Figure 3.16: *Weka Model Illustration*

Select visualize classifier errors from the same option if you want to see classification errors depicted. All samples on the coordinate plane can be seen by sliding jitter (as shown in Area 1 of the following figure). The x-axis shows the anticipated classifier results, whereas the y-axis shows the actual classifier results. The squares reflect samples that were incorrectly categorised, and the stars represent the true categorized samples.

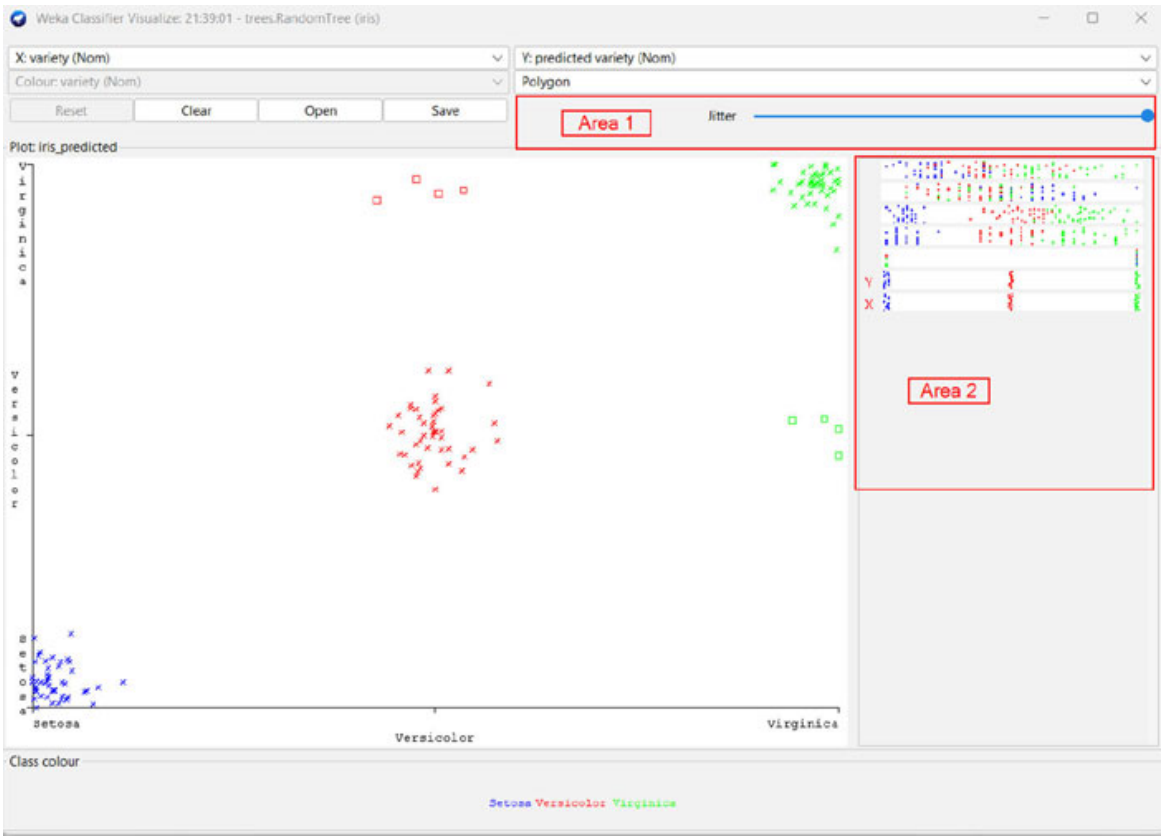


Figure 3.17: *Weka Visualize Model Error*

Choosing the right classifier depends on the user's knowledge of data, so understanding the data is critical. Even though it is referred to as machine learning, a human is required in most cases to oversee the data in datasets.

Orange: Orange is a platform designed for mining and analysis using a **graphical user interface** (**GUI**). This means that you don't need to know how to code to mine data, crunch numbers, and generate insights using Orange. You can do everything from basic graphics to data manipulation, transformations, and data mining. It combines all the process's functions into a single workflow. The following steps will give you a taste of Orange and show the end to end ML modelling process.

Step 1: Set up Orange, and it comes as a built-in application on Anaconda tool. Use the following link to download and install the application.

Orange (https://orangedatamining.com/download/#windows)

Step 2: Set up the working directory for Orange to store its files. This is how the Orange start-up page appears. You can start by creating new projects, opening existing ones, or looking at examples. A widget is the starting point for any data

modification. It can perform various tasks depending on what you select in the widget selector on the left side of the screen. A workflow is the series of processes or actions you take in your platform to complete a specific activity. Once you've established your first workflow, you can go to "Example Workflows" on your start-up screen to look at more. For the time being, click on "**`New`**" and begin creating your first process.

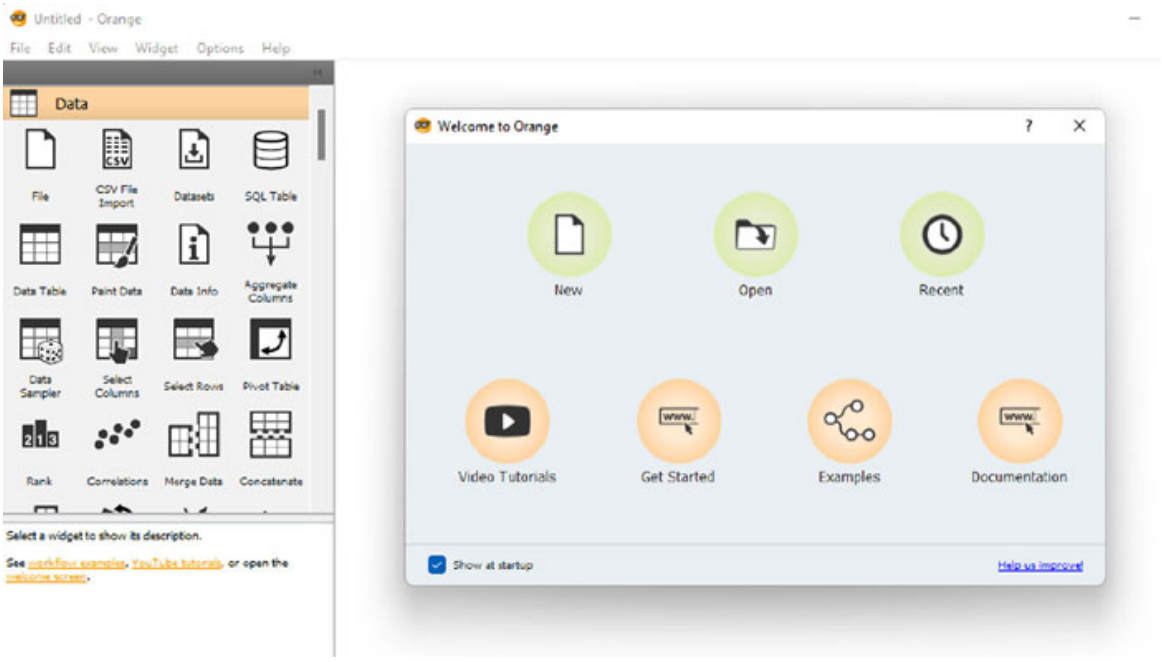


Figure 3.18: *Orange User Interface*

Step 3: Import data files

1. On the widget selector menu, click on the "**`Data`**" tab and drag the "**`File`**" widget to our blank workflow.
2. Double-click on the "**`File`**" widget and choose the file to load into the workflow.
3. Once you've seen the structure of your dataset using the widget, return to the previous screen by closing this menu.
4. Now that we have the **`raw.csv`** details, we need to transform them to a format that can be used in mining. Click on the dotted line enclosing the "**`File`**" widget and drag; then, click anywhere in the empty space.
5. We require a data table to better show our findings, so we select the "**`Data Table`**" widget.
6. Double-click on the widget to see your table in action.

Step 4: Visualize and understand the data

1. Drag the semicircle in front of the “File” widget to an empty spot in the workflow, and then choose the “**`Scatter Plot`**” widget, as shown in the following figure:

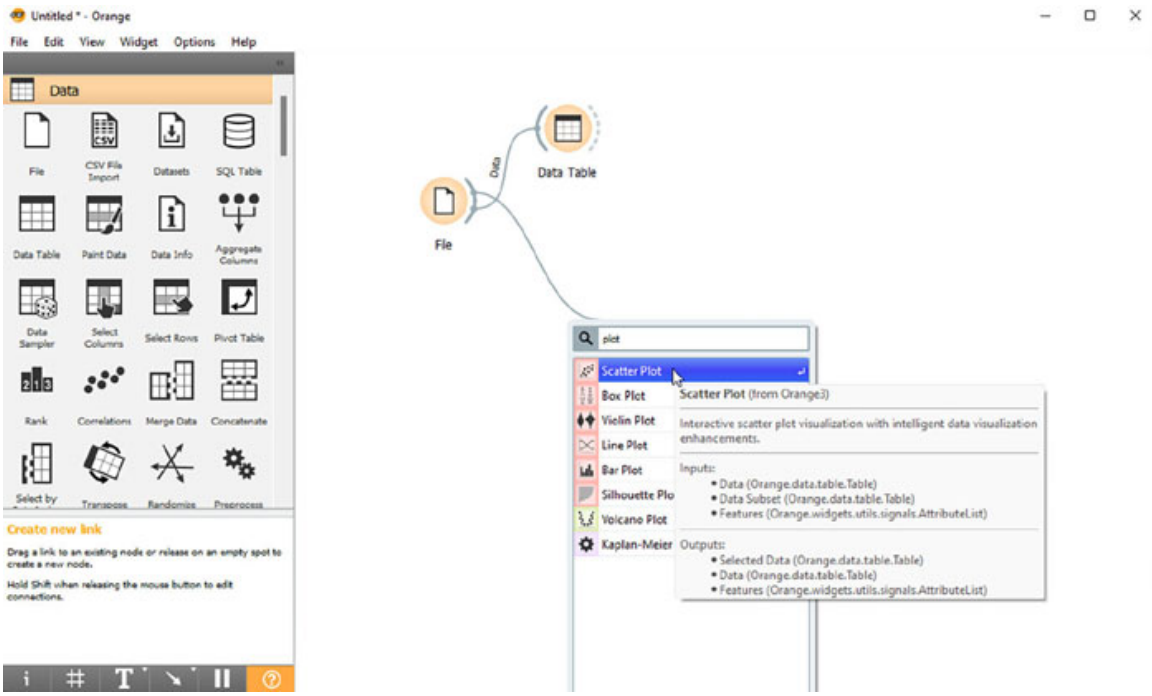


Figure 3.19: *Orange Layout*

2. After you’ve created a scatter plot widget, double-click on it to study your data using the visual. You may change the X and Y axes, colours, forms, and sizes, among other things.

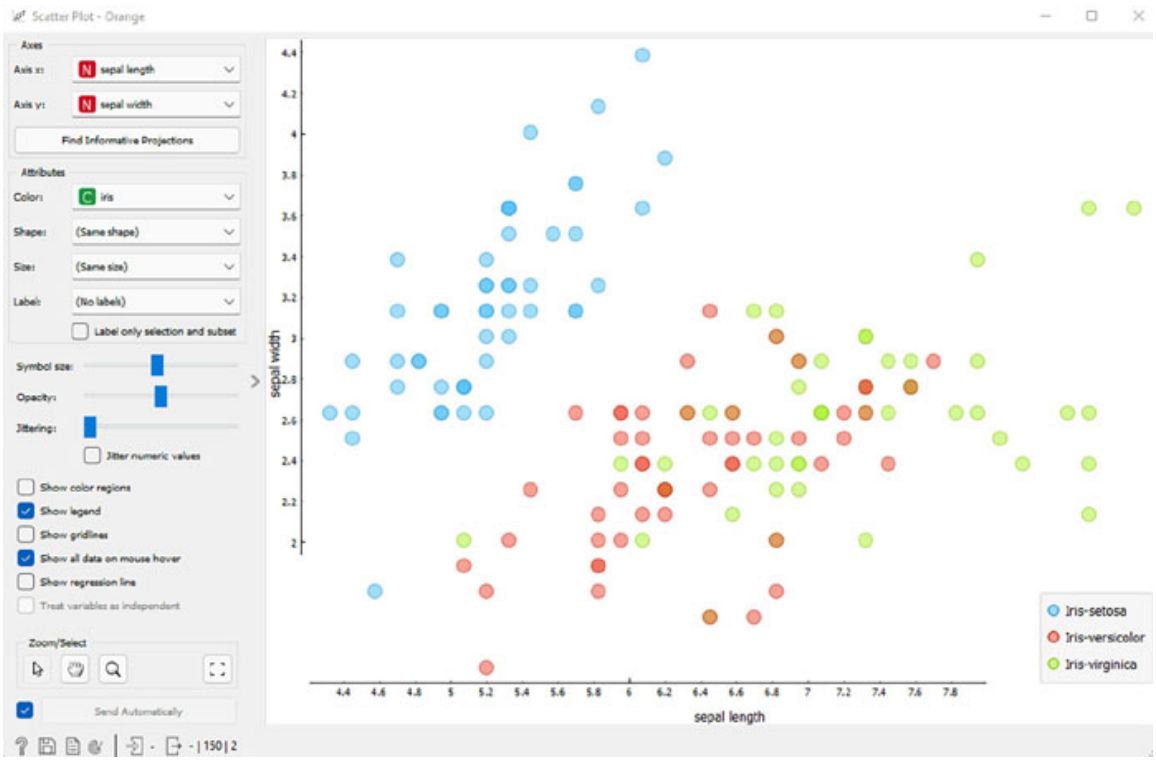


Figure 3.20: *Orange Data Visualization-1*

3. The “Distributions” widget is another option for seeing our distributions. Click on the semi-circle again and drag it to find the “Distributions” widget. Then, double-click on it to visualize it; this can be seen in the following screenshot:

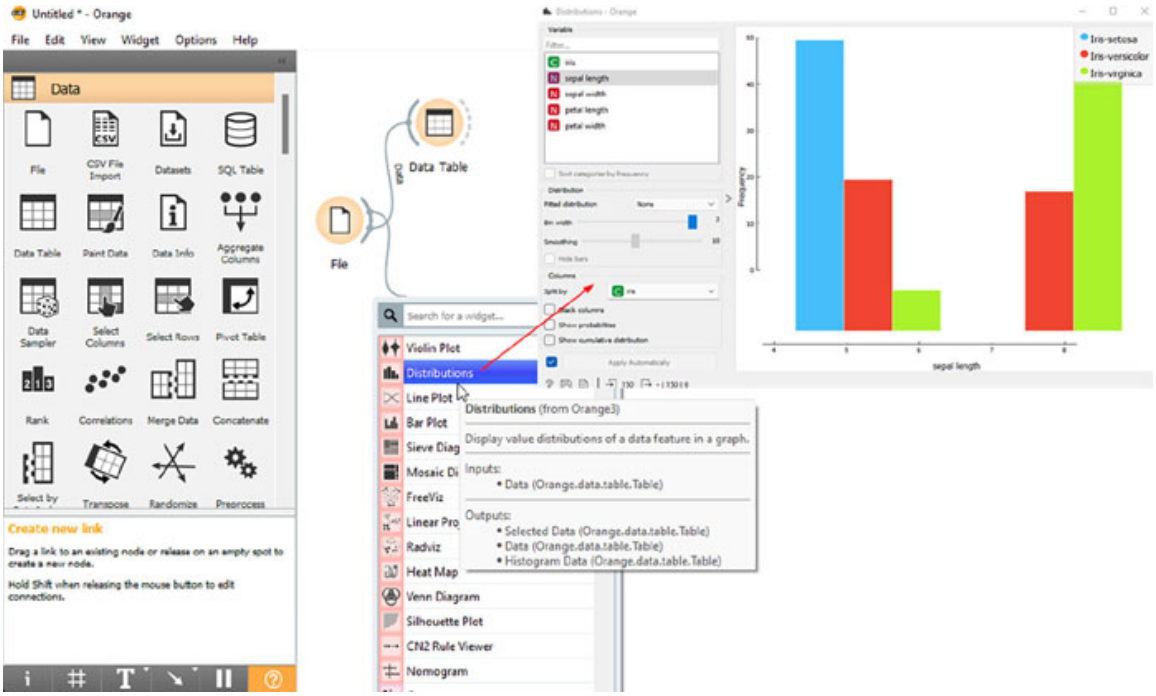


Figure 3.21: *Orange Data Visualization-2*

Rapid Miner: In order to use Rapid Miner, you need to first download the tool in your local system. The link to download the tool is **https://rapidminer.com**.

Ensure that you choose the 'Rapid Miner Studio' option. Select the type of operating system and then proceed to account setup once the downloading process is complete.

Create an account and choose a template to use. Use the green button to load data from the samples folder, which has a comprehensive list of datasets. You can also utilize the import option to load a personal dataset into the system for RapidMiner data mining.

To visualize the data, use the dataset's result button.This data visualization shows how the data points are related to one another. The widgets and home screen will look as shown in the following screenshot:

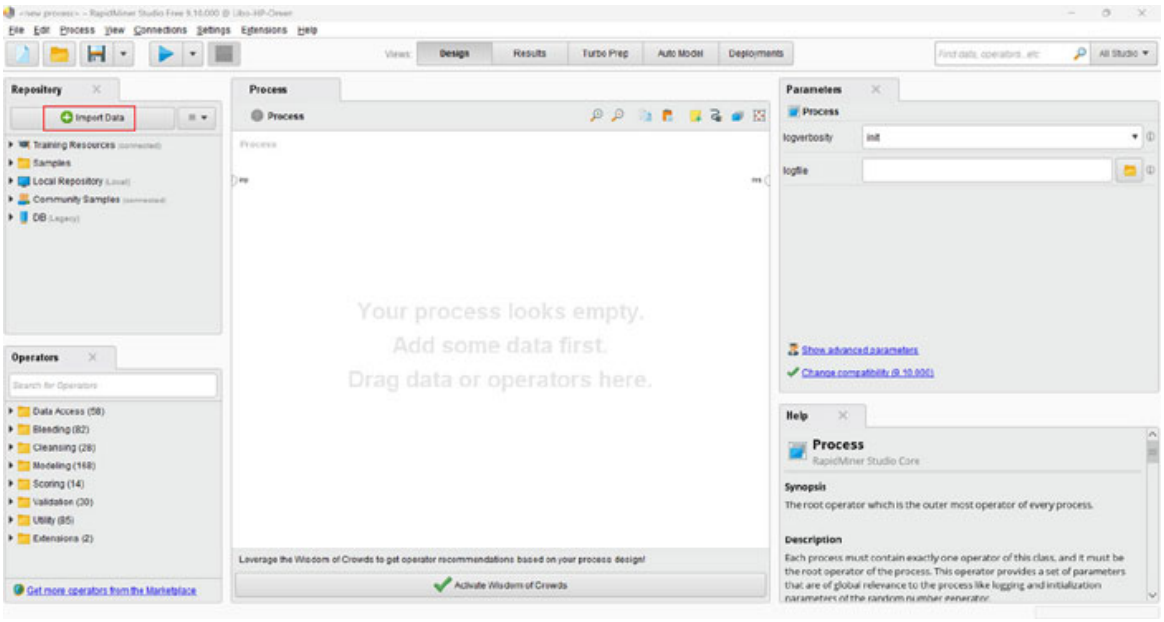


Figure 3.22: *RapidMiner Home Screen View*

Next, look at the data processing RapidMiner examples and options for cleaning, transforming, generating new datasets, merging columns, and statistically analysing data. The drag-and-drop choices can be used to move columns to be examined in order to target them together and find numerous analysis alternatives, such as the desired outcome's average, median and aggregate values. The clean option can be used to automatically clean the dataset structure and format.

After cleaning, select a label as the target column, and Rapid Miner will analyse the data using the correlation option across sets to deliver the least important column as the highlighted column. If the ID column is highlighted, this column is automatically discarded when the dataset is examined.. The next step is to do dataset standardization and PCA. This stage generates clean data, which may now be used for data modelling.

The auto-model option must be used with the processed dataset during the RapidMiner data modelling process. This provides options from which to choose, find outliers, predict, cluster, and so on. Use a prediction set, such as the popular Iris dataset, and then select the target column and click on next to see the target distribution. When the target analysis is finished, the user is presented with column alternatives. For optimum predictive efficiency, the key columns should be chosen based on one's needs. Select the model kinds as well to compare performance and choose the best model. One can choose which model to execute and whether the execution should take place on a local machine or in the cloud. When the best logistic regression in the RapidMiner model is executed, one

receives comparative findings and can choose to inspect the model's mistakes, confusion matrix, accuracies, and so on. All these steps can be seen as a process flow chart in the following screenshot:

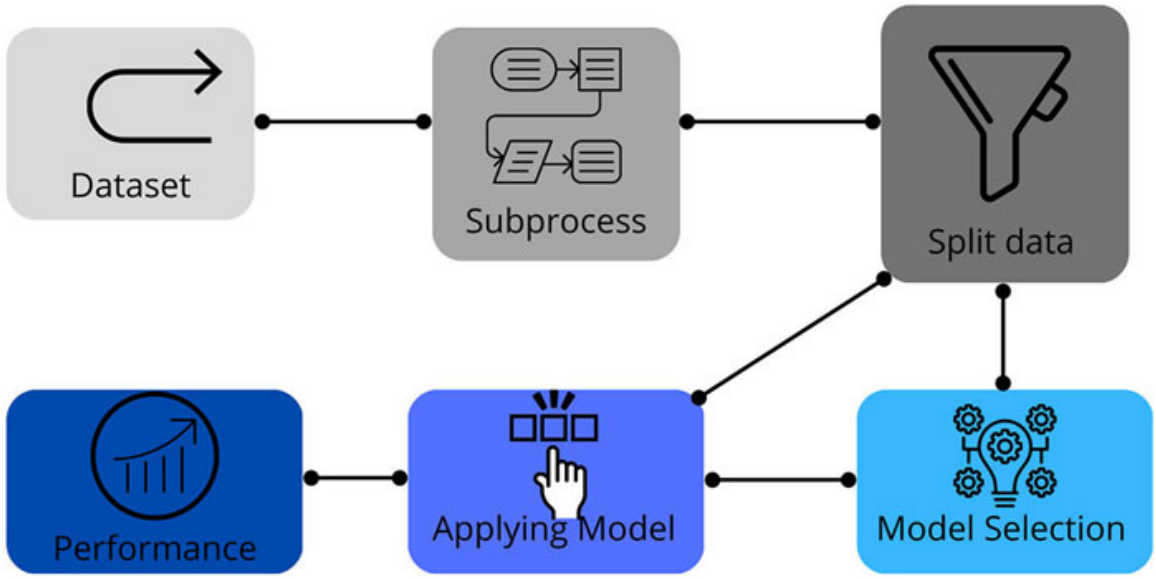


Figure 3.23: *Rapid Miner Process Flow*

Conclusion

Analytics tools are essential for every data professional, but they may be hard to decipher and execute for those who are not skilled. In this chapter, we saw some of the analytical tools available for handling large amounts of data. These analytical tools help solve many business problems, but they are often used to provide answers to queries such as the following:

- What is the hidden pattern in the data?
- How can you use it to enhance your business process?
- What is the most effective approach to displaying the information?

In the next chapters, you will look at the influence of machine learning on business results, recognizing business needs, understanding the math, and selecting the right algorithm or ML method.

Points to remember

- Analytical tools are important for everyone working with data, but they can also be challenging to understand and implement.

- Keep in mind the purpose and hardware requirements while choosing the right analytical tools.
- Excel spreadsheet has lots of untapped predictive ML capabilities, which can come in handy for students who are working on their research thesis and domain experts or analysts who work on micro-level data day in and day out.
- Apart from this, Knime, Weka, RapidMiner, Orange and low code Python libraries are available for use, and depending on the purpose and data availability, these tools can be used for rapid prototyping.

Multiple choice questions

1. What key feature(s) does auto analysis tools possess?
    a. Cloud-based
    b. High fidelity
    c. High intensity
    d. High speed
2. Which of the following is an advantage of using auto ML tools?
    a. Can speed up data analysis (i.e., after the data is transcribed)
    b. Many operations that are seldom done manually owing to time restrictions may be automated
    c. Non techies and domain experts can quickly generate insights
    d. All of the above
3. Orange can be used for image analysis.
    a. True
    b. False

Answers

Question Number	Answer
1	b
2	d
3	a

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 4

Machine Learning in a Nutshell

Machine learning is a part of AI that aims to offer accurate predictions based on data without being explicitly programmed to do so. The job is to make the decision-making process easier, faster and accurate.

The core idea behind machine learning is to create algorithms that can take in input data and apply statistical analysis to predict a future event, while also updating its output predictions when new data is added. The goal is to enhance prediction accuracy as the volume of input data rises. Machine Learning is commonly confused with Artificial Intelligence, but they are not the same. The goal of artificial intelligence is to create a machine that can mimic the behaviour and capabilities of a human mind, which includes things like reading words and languages, the ability to analyse photographs, videos, and motion pictures, and whatever other jobs people can do. Machine learning is a subset of artificial intelligence.

Structure

In this chapter, we will cover the following topics:

- Machine learning life cycle and its impact on business outcomes
- Understanding business need
- Couple business need with data
- Understand and finalize the math
- Choose the right algorithm: decision tree to select the algorithm.
- Break the myth: only machine learning can empower you to take data-driven decisions

Objectives

After studying this chapter, you will understand how machine learning can help you solve your business problem and how it helps you anticipate the future problem. When you anticipate the problem before it happens, you can take

corrective steps to either eliminate it or reduce its impact. It is very important to understand business problem well enough to choose the right data that would help you incorporate the right machine learning algorithm in order to arrive at the right decision-making step. In this chapter, you will learn how everything from the business problem to the solution implementation process is highly interrelated.

Machine learning life cycle and its impact on the business outcomes

Data scientists and data engineers follow the three phases of the Machine Learning Life Cycle to develop, train, and test models using the massive amounts of data involved in various applications so that an organisation can derive practical business value from artificial intelligence and machine learning algorithms. The following figure will give you a generic overview of the various phases of machine learning.

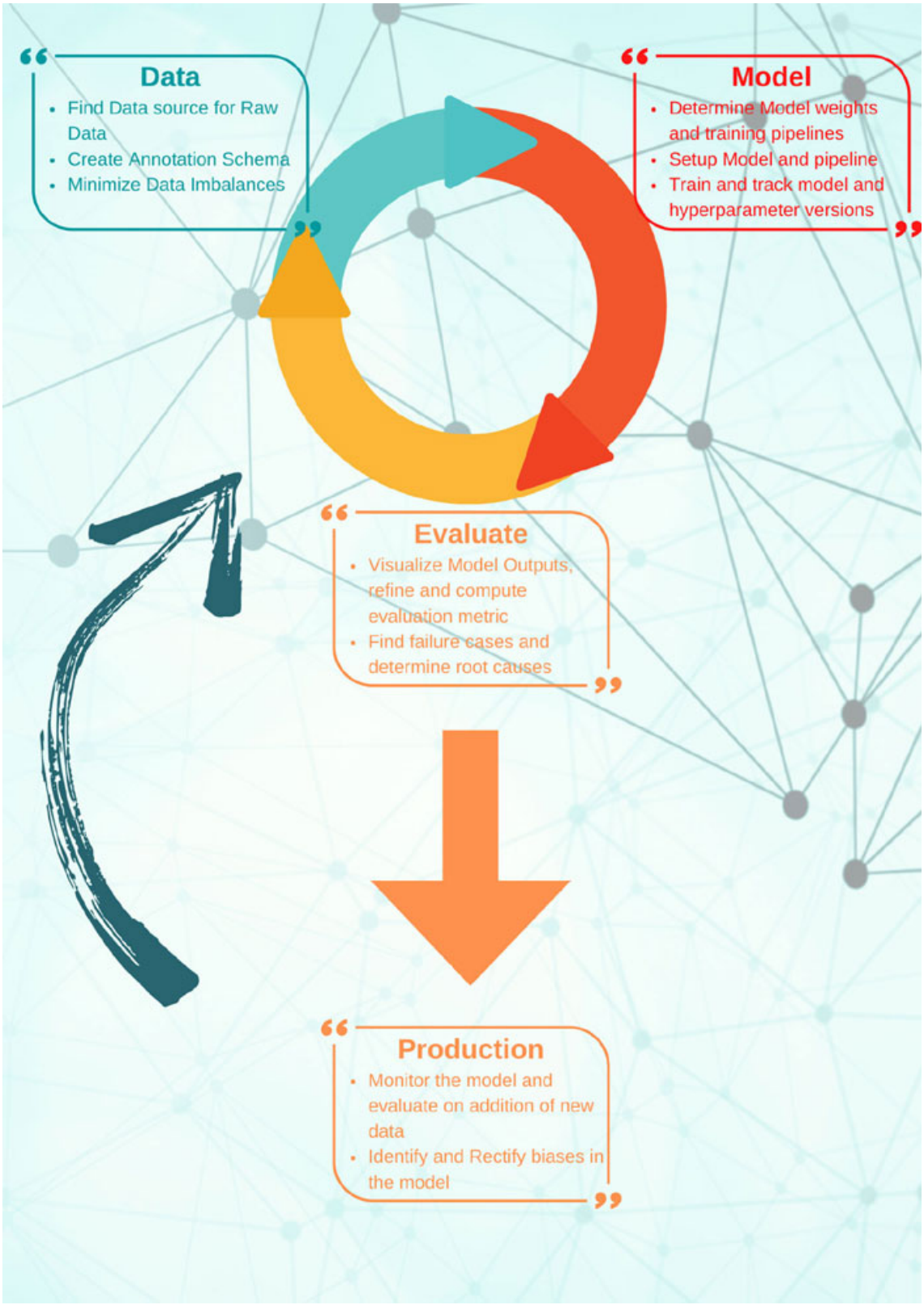


Figure 4.1: *Phases of Machine Learning*

The Machine Learning life cycle begins with a business problem and then moves on to identifying a solution and implementing the model. An ML model's life cycle can be broken down as follows:

Step 1: Establish a business context and outline a problem: Understanding the business problem and the criticality of solving it forms the base of defining the problem statement. Understanding the problem and defining the problem statement are two different things. Most of the times, the problem statement is not laid down clearly, which makes the subsequent steps difficult to accomplish. The step starts with asking the relevant and right questions to the business stakeholders who understand the business issue. This sets the base of defining the problem statement and brings in clarity about the problem at hand.

It is very important to understand the business context and identify the challenges that the business is facing. For example, if there is a Churn Analysis model being considered for a banking industry, it is very important to understand the banking industry. Here, the main problem to be addressed would be to understand what causes customers to churn and then anticipate which customers would churn soon. This can be used to plan the retention strategy.

Step 2: Translating the business problem into a Machine learning problem statement: A business problem needs to be translated into a proper machine learning problem. It is critical so that the model moves from conceptualization to development and then to the deployment stage successfully. Although this seems simple, it is not easy. Many models that are built aren't production ready, mainly because there is a mismatch of expectation from the business perspective. This primarily happens when the business problem is not translated into a machine learning problem.

For example, in Churn analysis, the machine learning model that needs to be considered is a classification problem, which means users with the same behavioural patterns should be grouped together. This forms the basis of the retention policy to be followed.

Step 3: Planning and Milestones: Keeping track of milestones and creating a timeline aid in understanding project progress, resource planning, and deliverables. This is critical to keep the whole process within budget. It is very important to have a clear understanding of the computational power and human resource need for a particular machine learning project. If the data that needs to be ingested is too huge, then the computational requirement of the local machines will be high. In such cases the, the deadlines and budget need to be planned accurately to avoid any hassles in the future.

Step 4: Data Collection and Analysis: Data is not always readily available in appropriate formats or with the necessary features to develop a model. The necessary data must be acquired from businesspersons. Data comprehension entails having a thorough understanding of the features contained in the data, what they precisely represent, and how they are obtained.

Typically, data scientists rely on their organization's databases to gather information for their commercial projects. However, there may also be cases where valuable data cannot be categorized or stored in a relational database, such as raw log files, images, and video recordings. . **Extract, Transform, Load** (**ETL**) pipelines are used by data engineers and scientists to do extensive processing on these datasets. Either a data lake or a database is used to hold these types of data (either relational or not). To get the data they require, data scientists can scrape websites, buy data from data providers, or collect data from surveys, clickstreams and sensors and cameras, among other methods.

Step 5: Data Preparation: Data scientists must prepare the raw data, undertake data exploration, visualize data, transform data, and repeat the procedures until the data is ready for modelling. Processing of raw data for analysis is known as data preparation. Before developing any machine learning model, data scientists must first comprehend the available data. Raw data can be disorganized, duplicated, or erroneous. Data scientists investigate the data at their disposal, and then cleanse it by detecting faulty, erroneous, and incomplete data and replacing or deleting it.

Step 6: Clean Up the Data: Furthermore, data scientists must identify whether the data has labels. To build a detection model to determine whether there is a dog in an image, you'll need to have a set of photos tagged with whether there is a dog in them as well as bounding boxes around the dog(s) in the image. If the photographs are unlabelled, human annotators will need to label them. There are open-source technologies and commercial suppliers that offer data labelling platforms as well as human annotators for hire.

Once the data has been cleansed, data scientists study the features (or variables) in their dataset to identify any correlations between feature. Tools for exploratory data analysis are available in open-source libraries, and analytics and data science platforms. This stage requires the use of a software that does statistical analysis on a dataset and creates graphs of the data and its characteristics. Also, you must determine what type of data or features are included in the dataset before you begin analyzing the data. For example, find the type of data you have; floating

point and integer features are examples of numerical features. The goal here is to provide a simple and concise summary of the main features of a dataset.

- Categorical features have a limited number of possible values and are often used to categorize data. Gender is an example of categorical data.
- Ordinal features are categorical features that have a predefined order or scale. Customer satisfaction responses, for example, have a predetermined order: extremely satisfied, satisfied, indifferent, unhappy, and very dissatisfied. This ordering can be expressed in numerical format as well.

After determining the type of features, establishing a distribution of values for each feature and producing summary statistics for each feature would be the next step. This would aid in answering the following dataset-related questions:

- Is the dataset biased toward a specific range of values or a subset of categories?
- Do you know the feature’s average, lowest and highest values?
- Are there any missing or invalid values, such as null? If so, how many?
- Is the dataset riddled with outliers?

Plotting features against each other during data exploration can help you discover patterns in your dataset. This aids in determining whether data transformation is required.

Some of the questions you must answer are as follows:

- How do you deal with missing values?
- Do you wish to fill in the data? If yes, how will you fill in the missing values? The mean, median, mode, the value of a neighbouring entry, and the average of the values of surrounding entries are some of the methods to impute missing values.
- How will you deal with outliers?
- Are some of the characteristics related?
- Is it necessary to normalize the dataset (for example, log transformation)?
- How do you deal with categorical variables?
- Do you utilize the data exactly as they are, organize them in some meaningful way, or disregard a portion entirely?

Step 7: Analyse exploratory data: This stage provides more information about the data and how it relates to the target variable. This step mostly entails the following:

1. **Analysis of univariate data**: Frequency distribution tables, bar charts, histograms, and box plots can be used to identify individual feature data trends. When analysing specific features such as monthly charge distribution or number of churners and non-churners, frequency distribution tables, histograms, and box plots can be used to identify trends and patterns in the data, such as whether the data is normally distributed or skewed, and whether there are any outliers or extreme values. Bar charts can be used to compare the number of churners and non-churners, as well as to identify any differences between different months or time periods. Overall, univariate data analysis provides a useful way to summarize and understand the characteristics of a single variable.
2. **Bivariate and multivariate analysis**: Scatter plots, Correlation Coefficients, and Regression analysis (these terms will be explained in upcoming chapters) can be used to determine the behaviour of the target variable depending on independent factors. Identifying churn based on user tenure and total charges is an example.
3. **Pivots**: It helps us to quickly extract insights from data.
4. **Data visualization and insights**: Based on the data visualization in steps 2 and 3, you can determine hidden patterns and insights. These data insights are the key takeaways from this step. Identifying the gender and age group of churners is an example.

Step 8: Feature design and selection: Identifying the correct drivers/features that affect the goal variable and deriving new features based on current features is what feature engineering entails. Some features can be deleted from the data based on their importance, which helps reduce data size. For example, feature importance can be determined using correlation, **information gain** (**IG**) and the random forest model. It is necessary to find the appropriate drivers/features that influence the target variable. This necessitates a thorough understanding of the business context and use case.

Step 9: Model assumptions and verification: Some ML models require assumptions to be checked before proceeding with model construction. In most cases, model assumptions do not hold true when applied to real-world data. Based on the following assumptions, a linear regression model is constructed:

1. Data is distributed normally.
2. There is a linear relationship between the dependent and independent variables.
3. The variance of the residuals is considered to be constant. (Homoscedasticity)
4. There is no multicollinearity in the data. When two or more independent variables (also known as predictors) are highly correlated in a regression model, the term “multicollinearity” is used.

Step 10: Preparing Data for Modelling: Data preparation is the next step in the process. It is necessary when the data available for modelling is not adequate or there is lack of variety in the available data. Some of the techniques that are usually used are creating synthetic data by creating dummy variables, creating data sets to avoid over sampling or under sampling when the data is imbalanced, and ensuring the right split between the training and testing data sets.

For example, in the churn analysis, a customer’s gender is 0 if she is female and 1 if he is male. This categorical data must be addressed in regression by introducing dummy variables (Customer 0, Customer 1). This is an example of creating dummy variables. Here’s another example of over sampling or under sampling: if the churn to non-churn ratio is 95:5, then the model will learn highly churn behaviour, and therefore, the non-churn behaviour would be predicted incorrectly. So, this model will not be feasible in real-life cases.

Step 11: Create a Model: The next step in the process is developing the probabilistic model that would describe the relationship between the independent and dependent variables. The key is to base the model on the problem statement defined in the first step.

Step 12: Validation and Evaluation of the Model: The process doesn’t stop at the model building process; the actual authenticity and the power of the model start with the model validation and evaluation. The model validation and evaluation consist of distinct steps:

- Model testing
- Tuning the model
- Cross-validation
- Model evaluation metrics trade-off
- Model overfitting and underfitting

Step 13: Predictions and model implementation you might consider when creating a model:

Prediction and reviewing the outputs after fine-tuning the model is the final step. This step also covers deployment and identification of real-time production issues.

Some of the critical steps in the process are as follows:

- Model scaling
- Model deployment
- Adoption by business and real-time consumption of the outcomes
- A/B testing
- Mapping business KPI
- Measuring the performance and monitoring the accuracy
- Getting feedback from the real-time predictions and closing the loop

The following figure will give you a basic understanding of workflow related to machine learning algorithms. The process does not end with analysing the data and predicting the outcomes; there are several intermediate steps that will help get a high-quality final outcome.

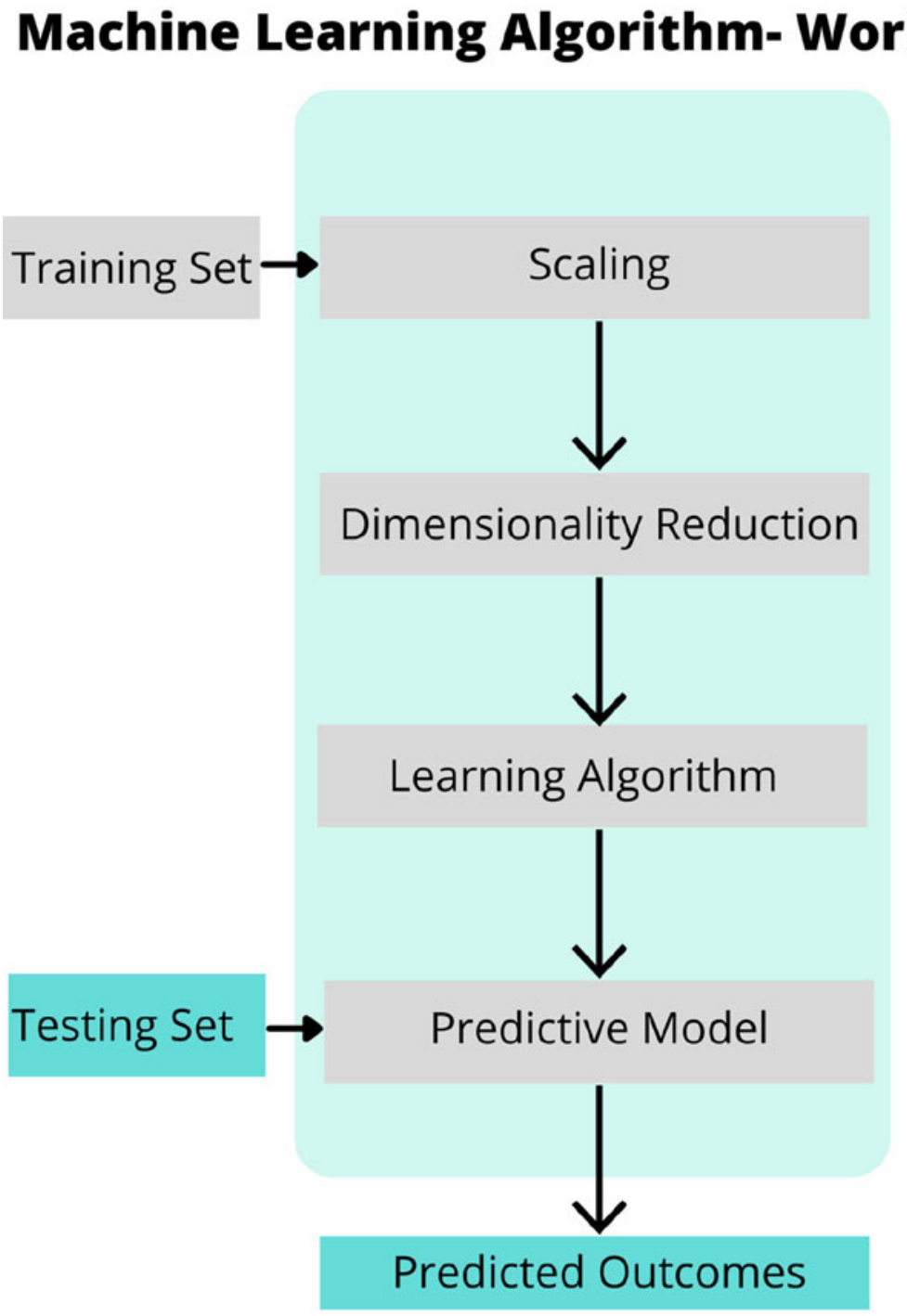


Figure 4.2: *Machine Learning Work Flow*

Understanding business need

A data scientist listens for crucial words and phrases when interviewing a line-of-business expert about a business challenge. Understanding the problem and the data, and getting an awareness of the many AI and data science approaches that

can be used to solve the problem are all part of the data scientist's routine. Putting it all together yields a collection of iterative trials, modelling tools, and assessment of models against business objectives.

The emphasis must remain on the business. Specifying all business inquiries as precisely as possible is critical. Any additional business requirements have to be determined. One such example can be avoiding client loss while enhancing cross-sell potential. Predicted business benefits must be specified clearly, such as a 10% reduction in attrition among high-value clients.

Understanding the business needs also means determining which task or aspect to focus on and which one to neglect because of the value that the aspect would bring to the business.

The outputs of the AI and data science models are frequently needed to be post-processed in order to solve the business challenge. For example, in order to determine the appropriate next step, the list of customers who may churn must be arranged. A customer with a low value and a high churn rate may not be worth the effort and money it takes to keep. Customers with a high lifetime value may be prioritized above customers with a lower lifetime value but a higher churn score.

Let's take an example here, say of a bank, where the focus is on optimizing the loan underwriting process. The current process includes application of filters on the loan applications automatically to reject riskier ones. However, the bank data shows that the bank has been approving too many applications that have had repayment issues in this case.

Looking at the bank data, the following observations are made:

There are 150 fields, which are split into distinct groups like loan demographics, loan amount, term of loan, interest rate and the reason for taking the loan. There are groups for applicant demographics as well, like age, salary, employment duration, and home ownership. Also, risk factor information is available, like public records, credit score, history of bankruptcy, etc.

Based on the available data, the goal is to create a model that can predict whether a particular loan is a bad loan. Note that in the available data, there is no field that tells whether a loan is good or bad.

Now we need to understand the business problem, so the first step is to articulate the business problem, asking questions like "*Can you detect any attributes about a person or the type of loan that can be used as a flag to tag a particular application as risky?*" But before we decide what is risky, we need to understand what a risky loan is or identify which loans are problematic. The next step would

be to segment loans into different categories using information like the purpose of the loan, the loan amount and the term of loan. Also, customers need to be sorted into various demographics, and patterns that cross over the mentioned parameters need to be identified.

Developing effective AI and data science solutions requires a thorough understanding of the business problem, including establishing clear measurements for baselining and validation, as well as identifying relevant patterns to address the underlying issues. By breaking down complex business needs into manageable and repeatable techniques, solutions can be developed that target the root cause of the problem and drive tangible results.

Couple business need with data

The journey of insights to informed business decisions starts with matching the right, relevant data with the business need. In the previous chapters, we looked at the different types of data and how to narrow down on the relevant data from the billions of data points available; now, it's time to match the data with the business problem identified.

Let's take an example of Amazon, which is the largest e-commerce platform. When a regular user logs into their account, the platform suggests a list of products they would be interested in buying; this is a simple example of a Recommender System. Now, let's understand what problem the recommender system is solving.

- First of all, it is providing an improved user experience to the customer, who feels valued when Amazon suggests the right products worth buying. This means Amazon cares for the user preferences.
- Secondly, it helps in higher sales conversion and the user, who would have thought of buying a single product, would end up buying more sometimes because the suggestions invoke their interest.
- Lastly, the whole experience not only improves business growth but also ensures that the user returns to the platform again for their next buy.

It doesn't matter if the customer buys the things, adds them to their cart, or simply looks at them, Amazon utilizes the activity information to understand what each customer wants and enjoys. It offers the same or comparable products to them when they return to the e-store. This is how the corporation obtains 35% of its annual sales. This is how important data is when it comes to solving the business

problem, which, in this case, is to improve sales; it works wonders for a company like Amazon in terms of revenue.

Data analytics and machine learning can help overcome higher user churn or problems with customer retention as well. When the customer behaviour is analysed and trends are interpreted, it helps determine when a customer would churn, and informed decisions can be taken faster to reduce probability of churn.

While planning to reposition a product or launch a new product, one of the most important things to focus on would be to take the product to the right target audience. You can determine what people think of your brand and products by reading reviews and comments left on your company's social media channels. These are data points that will help you shape the right ad campaigns, improvise your marketing techniques, and take them to the target audience for a better conversion rate.

Businesses may now traverse massive datasets from internal and external sources like social media, sales, customer experience, and the environment. They cah entirely restructure their operations. They can create customer profiles based on millions of individualized datasets with supplementary semantic data, allowing them to understand why a consumer picks their product over a competitor's and vice versa. The following image depicts how the decision-making process works based on the insights or results communicated from the machine learning models:

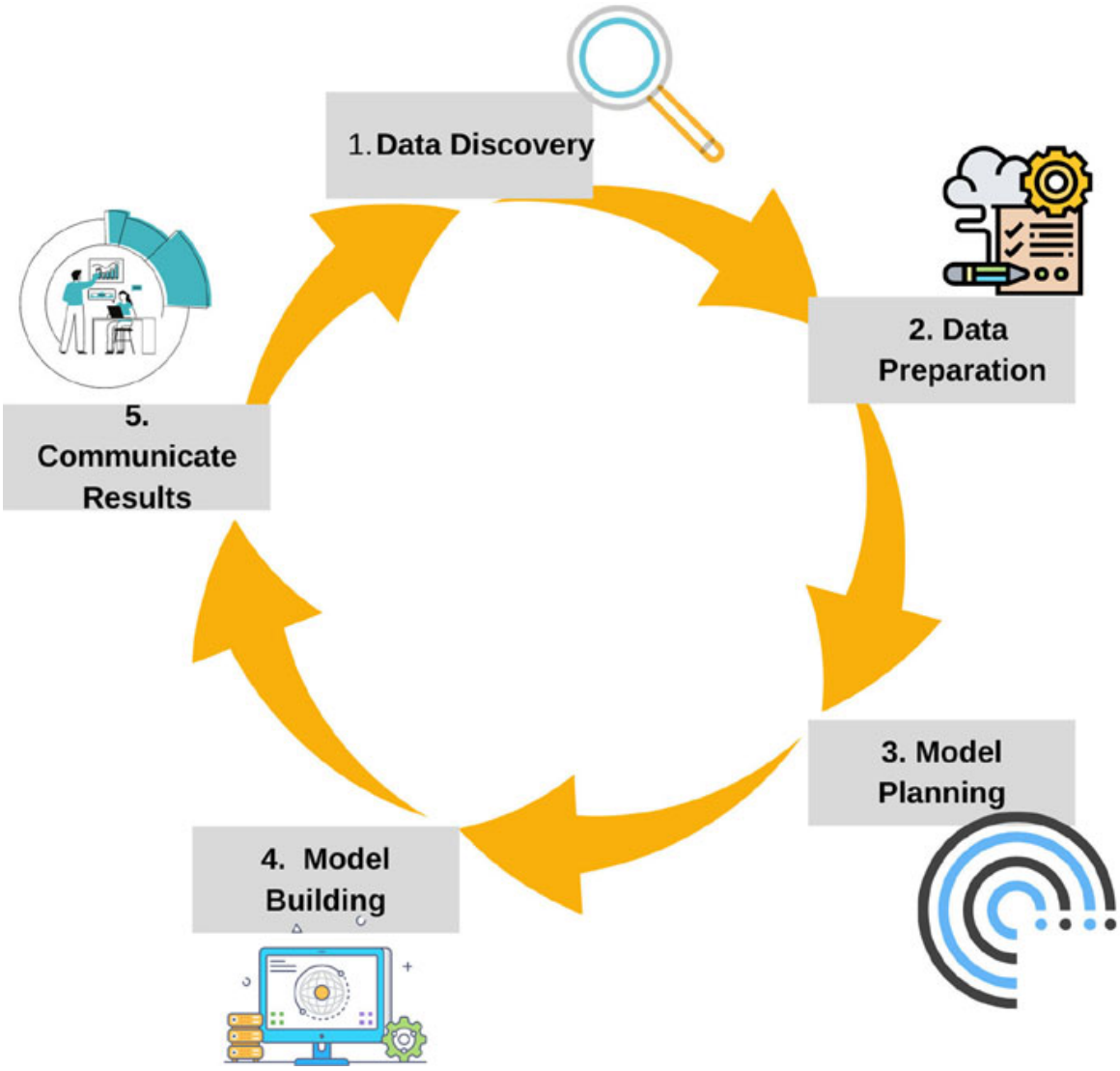


Figure 4.3: *Journey of Informed Business Decisions*

Understand and finalize the Mathematics

As young kids in school or college, most of us have had dreadful fear of one subject: Mathematics. When we talk about Machine Learning, we need to understand the importance of Mathematics. But the question is, “Do we need every aspect of Mathematics if we want to venture into Machine Learning?” The answer is no, but it is important to know what matters.

Calculus, Matrix Decompositions, and Statistics are all essential, as is linear algebra. The clarity of these concepts helps in creating intuitive machine learning models and improves the applicability of the deployed models.

- Machine learning heavily relies on Linear Algebra as it serves as the main component for almost all complex processes considered while model building. Computation is the primary focus of Linear Algebra, which deals with vectors and matrix algebra. Deep learning and machine learning both rely on it heavily. It is not uncommon for **machine learning** (**ML**) experts and researchers to employ Linear Algebra in the development of algorithms like linear regression, decision trees, and support vector machines.
- In machine learning, probability forecasts the collection of outcomes, whereas statistics drive the favourable outcome to completion. The event could be as basic as a coin toss. Probability is divided into two types: conditional probability and joint probability. Conditional probability arises when one event takes precedence over the other, whereas joint probability emerges when two occurrences are not connected in any way. Given the likelihood of a previous occurrence, probability concepts can help calculate the chance of a future event. This is called Bayes' Theorem; it can be seen in the following figure:

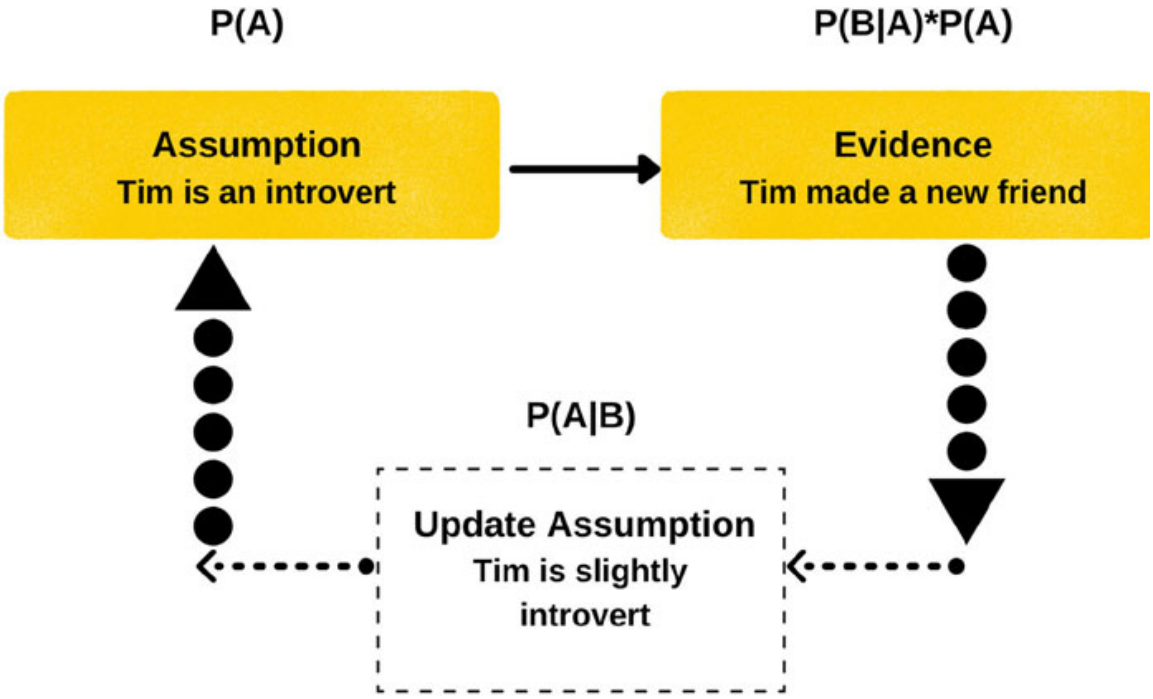


Figure 4.4: *Probability of an Event Based on Evidence*

- Calculus is at the heart of machine learning algorithms. It would be impossible to anticipate outcomes with a given data set if one cannot understand the concept behind it. It is possible to better comprehend the pace at which values change by using calculus, which is concerned with the

optimal performance of machine learning algorithms. Calculus principles like integrations, differentials, limits, and derivatives aid in the training of deep neural networks. A computer can understand linear algebra because it is a systematic representation of knowledge, and all linear algebra operations follow the same set of rules. Mathematically, a function may be optimised using multivariate calculus, or partial differentiation, to be precise.

- An algorithm's quantitative and qualitative characteristics are examined in statistics. It aids with the identification of goals and the transformation of the obtained data into precise observations by presenting it in a concise form. Descriptive and inferential statistics are the two types of statistics used in machine learning:
- Descriptive statistics is focused on describing and summarizing a model's limited dataset. Methods used in this study include the mean, median, mode, variance, and standard deviation. The results are provided in the form of graphical representations.
- Inferential statistics is concerned with gaining insights from a particular sample while dealing with a big dataset. It enables machines to analyse data that extends beyond the scope of the information presented. Some features of inferential statistics are hypothesis tests, sample distributions, and analysis of variance.

It is not enough to just know how to look at the numbers; comprehending what is happening, why it is happening, and how we can try other variables to get the desired results is equally important.

Choose the right algorithm

Machine Learning isn't easy, so choosing the right algorithm that can potentially solve a business problem is a tough task. Determining the right algorithm depends on multiple factors from understanding the problem statement and the output we are expecting, to whether it is a Yes or No problem or a range that the model needs to predict and how big the available data set is, everything plays a role in determining the right model.

Some of the most important factors that should be taken into account while choosing the algorithm are listed below:

1. **Size and volume of the training data**: Everything starts with the data and the size of the data. A good amount of usable data guarantees reliable and

accurate business predictions. However, as seen previously, the most difficult problem to fix is the availability of data. If the dataset is limited, then a model based on high bias and low variance, like Linear Regression, Naïve Bayes or Linear Support Vector Machines, would be useful. If the training data is ample and multiple observation are available when compared to the number of features, a model based on low bias/high variance, such as KNN or decision trees, can be used.

2. **The output's accuracy and/or interpretability**: A model's accuracy means it predicts a response value for a given observation that is close to the genuine response value for that observation. A highly interpretable method (restrictive models like linear regression) means that each individual predictor can be clearly understood, whereas flexible models provide more accuracy at the expense of low interpretability, as it can be seen in the following figure that gives you a simple matrix of various models evaluated for accuracy and interpretability.

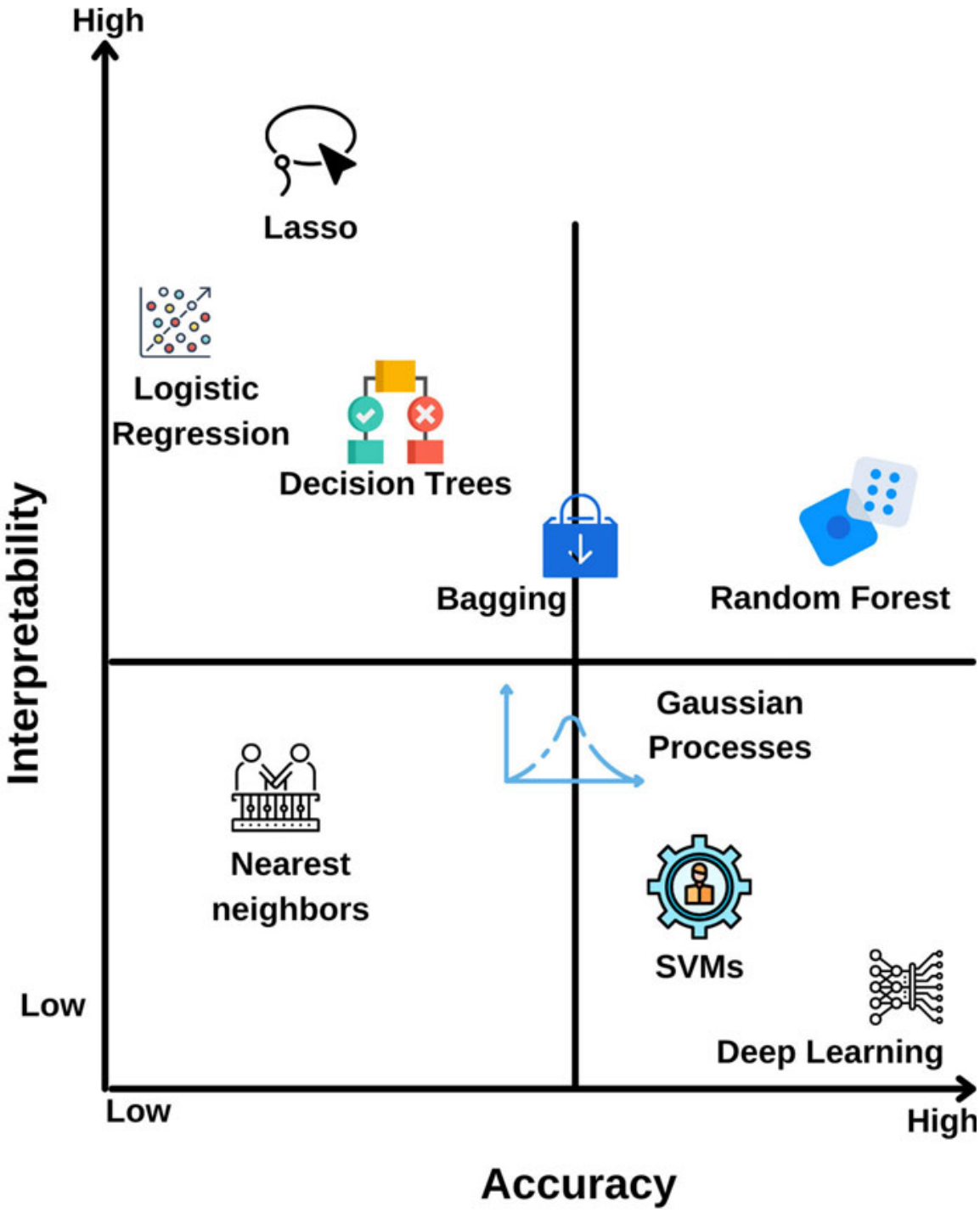


Figure 4.5: *Accuracy vs Interpretability*

Some algorithms are referred to as restrictive because they yield a limited number of mapping function shapes. Linear regression is a restricted technique as it generates linear functions like lines. Flexibility refers to the ability to construct a wide variety of mapping function shapes with the same technique. Some algorithms are flexible, like KNN. Every input data point

is considered while creating the mapping function for KNN with k=1. The following figure gives you a matrix of various models comparing the flexibility or predictivity and interpretability.

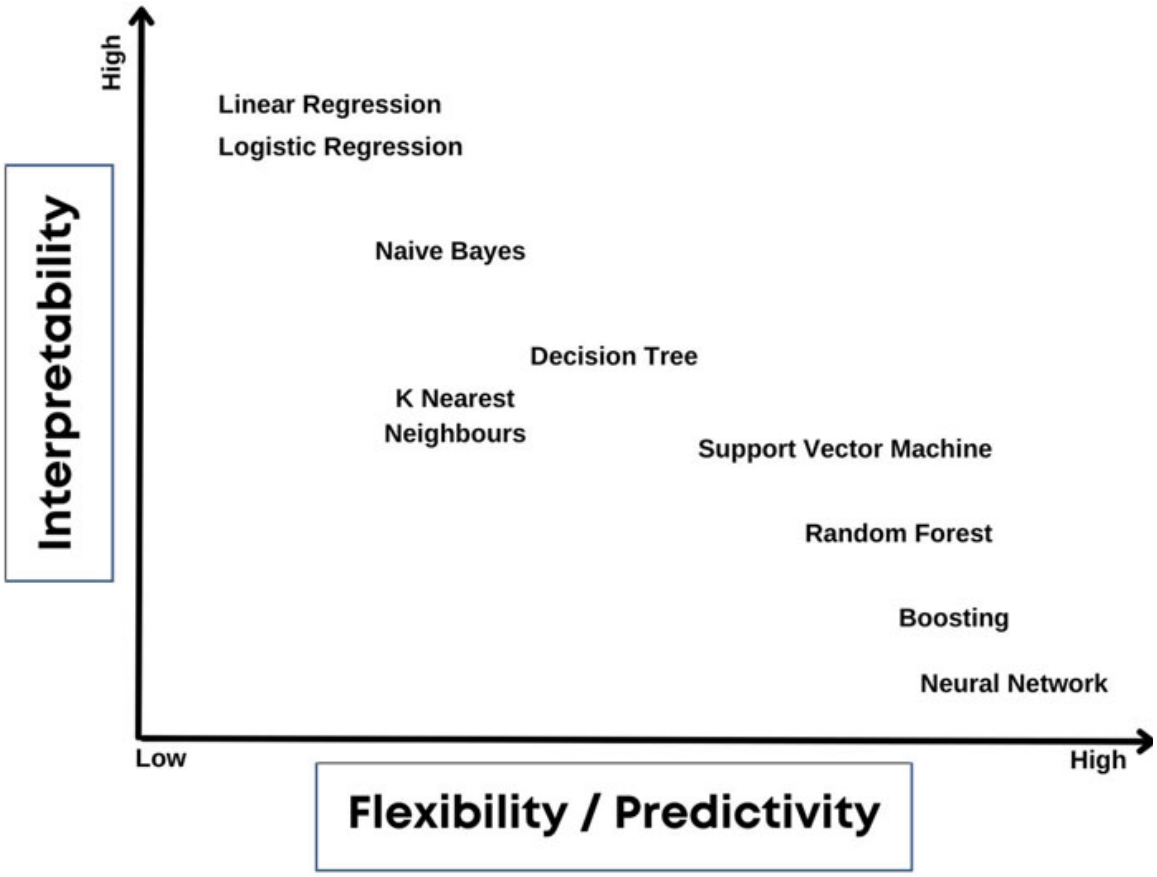


Figure 4.6: *Flexibility vs Interpretability*

Now, the business challenge determines which algorithm to use. Restrictive models are preferred if inference is the objective, as they're more interpretable. If greater accuracy is desired, flexible models are preferable. In general, flexibility and interpretability are inversely proportional.

3. **Trade-off between training time and speed**: Higher accuracy usually necessitates more training time, and algorithms take longer to train on big amounts of training data. These two criteria are the most important determining factors in algorithm selection in real-world applications. It's easy to implement and execute algorithms like Naive Bayes, Linear and Logistic regression. Algorithms, such as SVM, which requires parameter tuning; neural networks with a high convergence time; and random forests require a significant amount of time to train the model.

4. **Linearity**: Classes can be split by a straight line in the construction of several algorithms, for example, logistic regression and **support vector machine** (**SVM**). There's an implicit expectation that data trends would follow a straight line when using a linear regression technique. When the data is linear, these techniques perform admirably. However, not all data is linear, necessitating the use of alternative algorithms capable of dealing with high-dimensional and complicated data structures. Kernel SVM, random forest, and neural nets are some examples.
5. **Number of features**: The dataset might have a lot of features that might or might not be relevant to the business problem. A huge number of characteristics can prevent the progress of some learning algorithms, increasing the training time significantly. SVM is better suited for data with a large feature space but few observations. PCA and feature selection algorithms should be used to reduce dimensionality and identify key features.

The following image will give you an idea of various machine learning methods, their purpose and their real-world applications. Classification can further be classified into binary and multiclass based on the number of categories to be predicted.

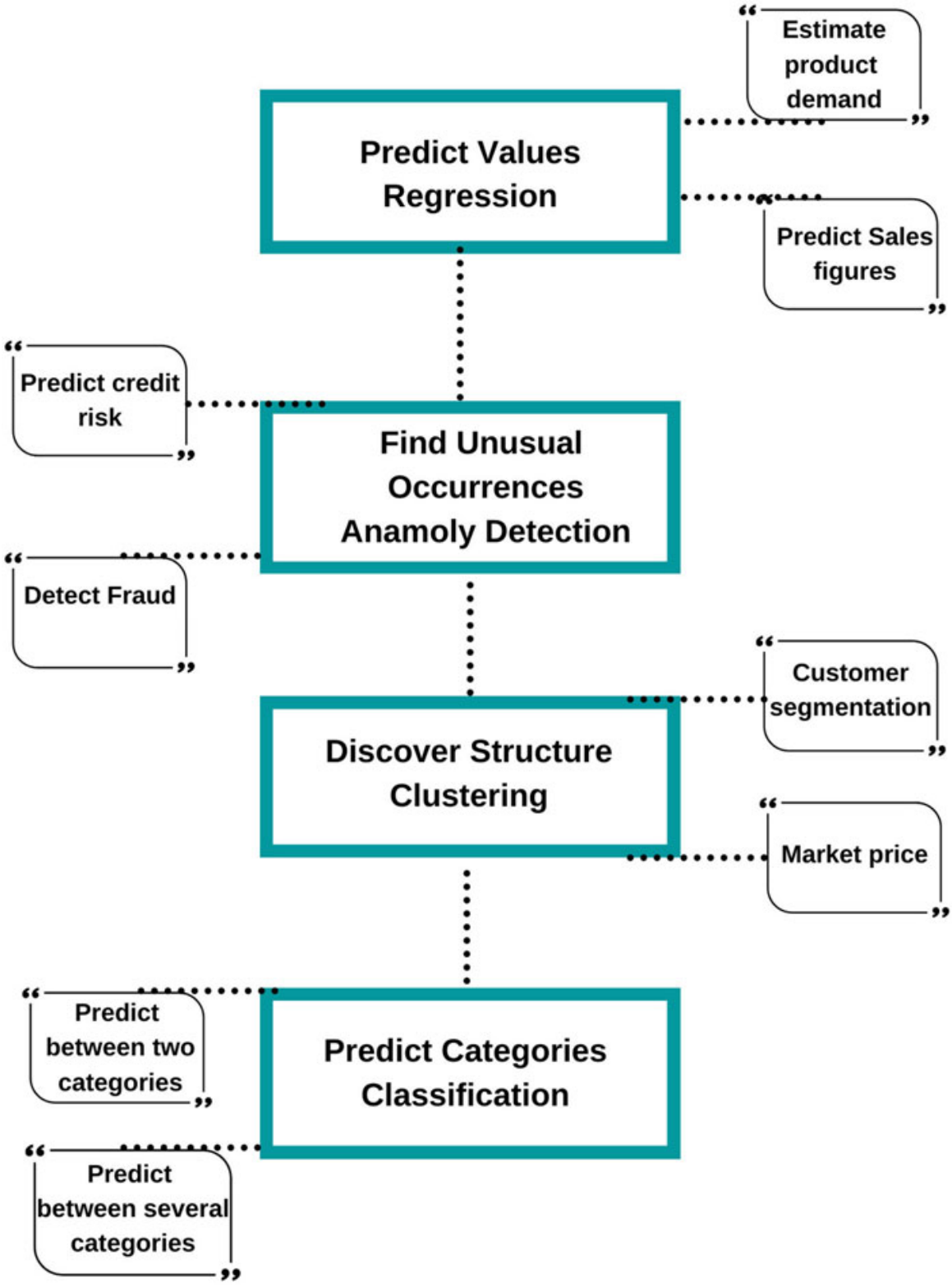


Figure 4.7: *Classic Machine Learning Methods and its Applications*

Break the myth; gone are the days of intuition-based decision-making processes

A company's bottom line may benefit significantly from the commercial outcomes produced by machine learning practical applications. New techniques in the field are continuously growing, opening up practically unlimited possibilities for the use of machine learning. Machine learning has been embraced as the greatest approach to construct models, strategize and plan by industries that rely on massive amounts of data and require a system to evaluate it fast and effectively.

Online retailers and social networking sites employ machine learning to analyse your prior purchases and search history to propose products based on your preferences. There is a strong belief among retail professionals that AI and machine learning will play a vital role in the future of retail as deep learning apps become more adept at collecting data for personalised shopping experiences and generating bespoke targeted marketing campaigns for retailers.

Machine learning and artificial intelligence are already being used to discover new energy sources and assess mineral resources in the subsurface, detect refinery sensor failure, and streamline oil transport to boost efficiency and save costs.

Machine learning, with its case-based reasoning, reservoir modelling, and drill floor automation, is also reshaping the business. In this market, the insights supplied by machine learning enable investors to spot fresh opportunities or know when to trade. Data mining identifies high-risk clients and assists cyber surveillance in detecting and mitigating symptoms of fraud.

Using machine learning, financial portfolios can be better calibrated, and risk assessments for loans and insurance applications can be performed more accurately. Machine learning is also not new to the large manufacturing business.

Machine learning applications in manufacturing aim to improve operations from conception to completion by drastically lowering error rates, enhancing predictive maintenance, and boosting inventory turn.

Using machine learning, like in the transportation sector, firms have been able to improve their supply chain and inventory management. By evaluating the uptime, performance, and reliability of assembly equipment, machine learning may also enhance **overall equipment effectiveness** (**OEE**). Experts believe that it's critical to grasp the value of adopting machine learning into your organization.

Conclusion

Machine learning is a process of deriving a predictive model from a set of data. It is usually applied to large datasets and is used for a wide range of business problems. Data scientists and data engineers follow the three phases of the Machine Learning Life Cycle to develop, train, and test models using the massive amounts of data involved in various applications so that an organisation can derive practical business value from artificial intelligence and machine learning algorithms.

Points to remember

- There are several steps that must be taken in order to maximise the influence of machine learning on the bottom line. These include:
    - Clear business understanding
    - The math behind machine learning, which is the core and fundamental part, and needs to be understood
    - Choosing the right algorithm based on the business problem and data at hand
- In a nutshell, the machine learning lifecycle consists of three phases: modelling, model validation, and model deployment.

Multiple choice questions

1. What does Bayes' theorem state?
    a. The probability of the impossible event is zero.
    b. Probability of the simultaneous occurrence of two events A and B is equal to the product of the probability of one of these events and the conditional probability of the other, given that the first one has occurred.
    c. Conditional probability of an event A, given the occurrence of another event B, is equal to the product of the likelihood of B, given A and the probability of A.
    d. None of the above
2. Which algorithm(s) works well for non-linear data?
    a. Logistic regression
    b. Support Vector Machine

c. Random Forest

d. Neural Network

3. Predicting product demand is an anomaly detection ML method.

a. True

b. False

4. The Machine Learning life cycle begins with data gathering.

a. True

b. False

Answers

Question Number	Answer
1	C
2	B, C, D
3	B
4	B

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 5

Regression Analysis

An easy way to establish a relationship between variables is through regression analysis.

When we want to establish a relationship between data and variables, the easiest and most effective way to start is by performing a regression analysis. It is a statistical technique to determine the relationship between a single dependent variable and several independent variables.

Structure

In this chapter, we will cover the following topics:

- Types of Machine Learning
- Basics of regression analysis
- Regression process flow
- EDA and Statistics for Regression
- Summary with tables, cheat sheets and handouts for easy reference

Objectives

After studying this chapter, you will understand the different types of machine learning algorithms and specifically, grasp the nuances of Regression Analysis. This chapter will help you understand when to use regression and what kind of business problems can be solved using it.

Types of Machine Learning

Arthur Samuel created the phrase “machine learning” to describe the ability of computers to learn without being explicitly directed (IBM Journal of Research and Development, Volume: 3, Issue: 3, July 1959). In order to forecast output values within a predefined range, programmable algorithms are used to accept and analyse input data. As more and more data is fed into these algorithms over time, they become smarter and better at what they do.

There are four types of machine learning algorithms:

- Supervised
- Semi-supervised
- Unsupervised
- Reinforcement

Supervised learning

Imagine a situation when as a kid, we were given an example of a role model or a studious kid by our parents to encourage us to study hard. Similarly, with supervised learning, the machine is trained using an example, which comes in the form of data. User inputs and outputs are given to the machine learning algorithm, which then must figure out how to get to the desired endpoints, which is the task of the algorithm. While the user has a fair understanding of the data, an algorithm searches for patterns or relationships, gathers knowledge from past experiences, and uses that knowledge to make predictions. As the algorithm learns and improves, the procedure of feeding new inputs followed by learning patterns is repeated until the algorithm's accuracy and performance are good enough to be useful.

Supervised learning branches out to three different aspects:

- **Classification**: To properly execute classification jobs, the machine learning system must make a judgement from observations and decide which category fresh observations fall into. Before deciding whether an email is "spam" or "not spam," the algorithm, for example, must take into account previously collected observational data.
- **Regression**: A machine learning algorithm must be able to predict and understand the relationships between variables to complete regression tasks. When used for forecasting and prediction, regression analysis is ideal since it can focus on one dependant variable and a sequence of changing variables.
- **Time Series Forecasting**: It is the practice of predicting the future based on historical information that has a time stamp attached to it. Time series forecasting, or the extrapolation of time series data, is necessary for many prediction issues due to the inherent temporal nature of the data.

Semi-Supervised Learning

When a person learns anything, they use both labelled and unlabelled material to accomplish the goal. Labelled data is information that has relevant tags added to it so that the algorithm can comprehend it, while unlabelled data does not have such tags attached. When working with a large amount of data and finding it challenging to extract valuable features, semi-supervised learning can be a lifesaver.

Unsupervised Learning

In this case, a machine learning technique is utilized to detect patterns in the data, without the use of labeled examples or the guidance of a human expert. Instead, an analysis of available data is used to uncover connections and relationships, allowing the algorithm to identify patterns autonomously, without external assistance. Unsupervised learning is a method in which the machine learning algorithm is left to understand enormous data sets so that it can uncover details that were previously unknown. The algorithm attempts to organize the data in order to describe its structure. This means finding similarities and clustering it in a way that appears more organized.

New data is analysed each time in the online learning mode, leading the decision-making abilities to continuously improve and become more polished. On the other hand, offline learning is more of a one pass where all data is used to learn similarities.

The following fall under "*unsupervised learning*":

- Clustering is the way of bringing together sets of data that are similar. It is useful for segmenting data into various groups, and statistical analysis can be performed later, if needed, to discover deeper meaning and patterns.
- To visualize the data patterns, you must reduce the number of variables you analyse. This is the essence of dimensionality reduction.

Reinforcement Learning

As a kid, my teacher used to reward me when I completed a given assignment and punished me on days I didn't complete my assignments. Similar to this method of learning and teaching, machine learning can be used to develop organised learning processes if it is given several objectives, parameters, and final values. Following the guidelines, the machine learning model tries to explore many alternatives and possibilities, constantly monitoring and evaluating each outcome to determine the optimum choice. It's called "*reinforcement learning*" when a

computer learns by making mistakes and trying again. It incorporates lessons learned from previous experiences and begins to tailor its approach in response to the current situation to get the best potential outcome. The following image will help you remember the mentioned points:

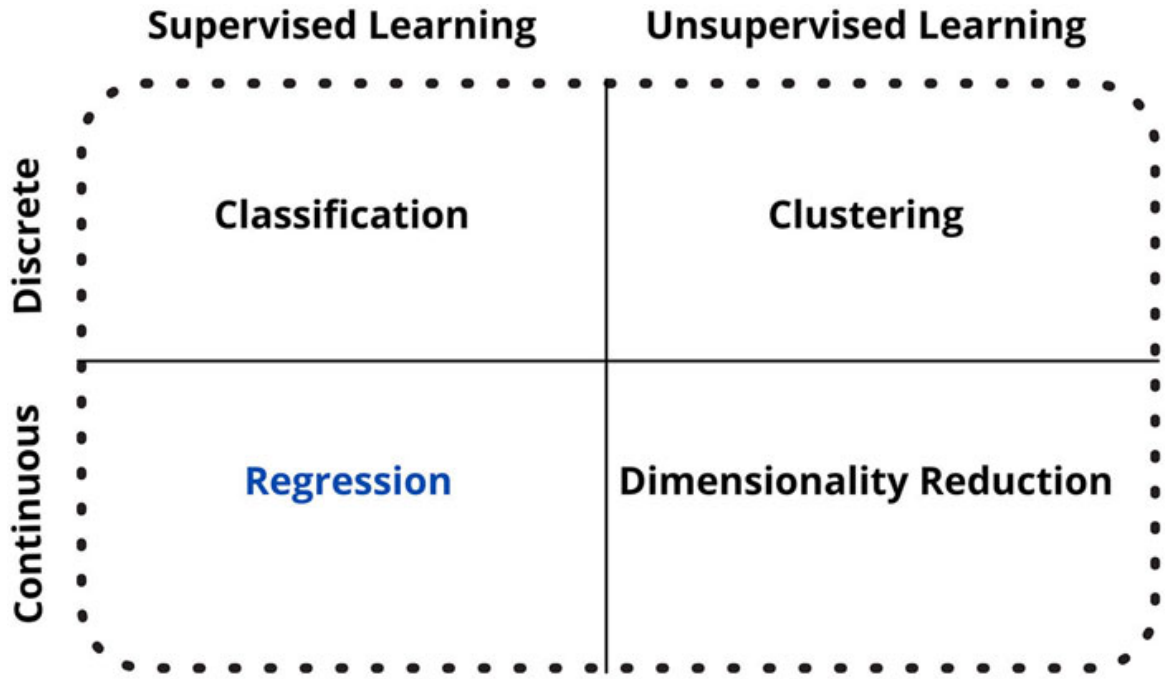


Figure 5.1: *Data Type vs Machine Learning Methodology*

Basics of Regression Analysis

Regression analysis is a technique used to find a correlation between data points, for example, between the number of calories consumed and the weight of a person. For the sake of simplicity, let's say that regression analysis is the process of finding the relationship between two variables. To put it another way, it is the process of finding the function that fits a subset of the available data. Regression analysis is one of the fundamental techniques in machine learning and is used to predict future events. For forecasting the future, you fit a function to the existing data and fit it to the data available using regression.

- Regression can be applied in a broad range of practical scenarios. This machine learning technique is suitable for tasks that involve continuous or numerical values, encompassing a wide range of applications, such as those involving large or small data sets.Time series forecasting
- Predicting the financial health of an organisation (like house price projections or stock prices)

- Predicting sales and promotions

The following concepts must be understood to fully understand regression analysis:

- **Dependent Variable**: As the name suggests, this is what you are trying to understand or predict.
- **Independent Variables**: The factors you hypothesize will have a direct impact on the dependent variable and are known as independent variables or predictors.

Five types of regression may be outlined as follows:

Linear Regression: One of the most fundamental forms of regression in machine learning, it includes the use of a predictor and a dependent variable that are linearly related (straight line relationship when graphed). As mentioned previously, linear regression requires the use of a best fit line to determine the best fit.

When your independent variables are linearly related to one another, you should utilize linear regression. When predicting the effect of increasing advertising expenditure on sales, for example, you should consider that the presence of outliers in large data sets can impact the regression analysis. Here are a few important points to remember regarding linear regression:

- It is quick and simple to model, and it is especially beneficial when the data to be modelled is not excessively complex.
- Understanding and interpreting it is intuitive.
- Outliers have a significant impact on the results of linear regression.

Ridge Regression: Ridge regression, on the other hand, is more appropriate if there is a strong degree of correlation between the variables being studied. It is referred to as a regularization approach, and it is employed to reduce the complexity of the model under consideration. It introduces only a little amount of bias (known as the ‘ridge regression penalty’), so the model is less prone to overfitting.

In cases where there is a high degree of collinearity among the feature variables, standard linear regression may fail to generate accurate predictions, leading to substantial prediction errors. . Collinearity is defined as the presence of near-linear relationship between the independent variables in each situation. It is

possible to observe or guess the existence of high collinearity using the following tips:

- Although a regression coefficient is theoretically expected to be significantly associated with the dependent variable, it may not exhibit statistical significance in practice.
- Modifying the set of independent feature variables in regression can lead to a significant change in the regression coefficients.
- The independent feature variables in regression may exhibit strong pairwise correlations, as evident from the correlation matrix.

Lasso Regression: LASSO stands for Least Absolute Shrinkage and Selection Operation. As the name implies, LASSO uses the "shrinkage" technique, in which coefficients are found and then shrunk toward the center point, which serves as the mean of the distribution. LASSO regression, which is used in regularization, is based on basic models with a small number of parameters. Because of the shrinkage process, we can acquire more accurate predictions. In addition, the shrinkage process allows for the identification of factors that are highly associated with variables that correspond to the target variables. This strategy is commonly used in machine learning to pick a subset of variables to be utilized in the analysis. When compared to other regression models, it predicts the future with better accuracy. The use of lasso regularization improves the interpretation of models.

Less significant features of a dataset are penalised by the lasso regression method of classification. The coefficients of this dataset have been set to zero, resulting in their deletion from consideration. With its high dimensions and correlation, the dataset is well-suited for use with lattice regression.

Regression process flow

Linear regression analysis entails more than simply fitting a linear line across a set of data points. It also involves the analysis of many data points and consists of several steps not limited to exploring the data to identify the various correlation that exists, evaluate the performance of multiple models, analyse the validity and reliability of the model, and so on.

- Create a list of probable variables/features to consider; both independent (predictor) and dependent variables are considered (response).

- Obtain information about the variables. Examine the significance between each predictor variable and the response variable to determine its significance. This could be accomplished using scatter plots and correlations.
- Examine the relation that exists between the predictor variables. Again, scatter plots and correlations will be helpful. Independent variables shouldn't be correlated with one another. If independent variables included in a regression model are correlated, a phenomenon known as multicollinearity arises. The term "multicollinearity test" refers to the test of multiple correlations. To identify multicollinearity, the **Variance Inflation Factor** (**VIF**) is used. The extent to which an independent variable's variance is impacted by its correlation with the other independent variables is quantified by VIF.
- Simple linear regression is an ideal approach to analyse the relationship between two continuous variables. Here, use the non-redundant predictor variables. This is based on examining the degree of multicollinearity between the predictor variables in the model. If association exists, one of these variables should be changed.
- Analyse one or more models in accordance with some of the following criteria:
    - t-statistics values are computed to evaluate if a parameter's value is equal to zero, thereby testing the null hypothesis.To test the null hypothesis, the p-value is utilised to assess whether the dependent and independent variables are linked. The greater the statistical importance of a parameter, the lower the p-value must be set.
    - R^2 (R squared) or adjusted R^2 is used to determine how well the model fits the data: This procedure evaluates the suitability of the regression model.
    - A R^2 value of zero implies that the dependent variable cannot be predicted from the independent variable, while an R^2 value of one tells that the dependent variable can be predicted accurately using the independent variable.
    - With an R^2 of 0.8, 80% of the variation in "y" can be predicted using the variance in "x."
    - The adjusted R^2 informs you how much variance can be explained by only the independent factors that influence the dependent variable,

rather than all of them.

- Increasing the adjusted R^2 will penalise you for the addition of independent variables.
- The adjusted R^2 value of 0.99 indicates that the model is extremely well fitted and that the model makes absolutely accurate predictions. Similarly, if the "R^2" and "Adjusted R^2" values are close to one other, it indicates that selected features are relevant and performing admirably. If they are diametrically opposed to one another, it is a clear indication that the traits chosen are not relevant.

- Predictions based on the predictor should be made using the best-fitting model (independent variables). This is accomplished through statistical analysis of some of the previously mentioned statistics, such as p-value and R squared. Also, t-score and F-value can be used. In linear regression models, the F-value is used to determine whether a given regression coefficient is statistically significant. In analysis of variance, the F-value is a useful statistic. The t-test statistic is helpful in determining whether there is a connection between the independent and the dependent variables.

The following figure will give you a quick summary of the various steps involved in linear regression modelling in the form of a simple flow chart:

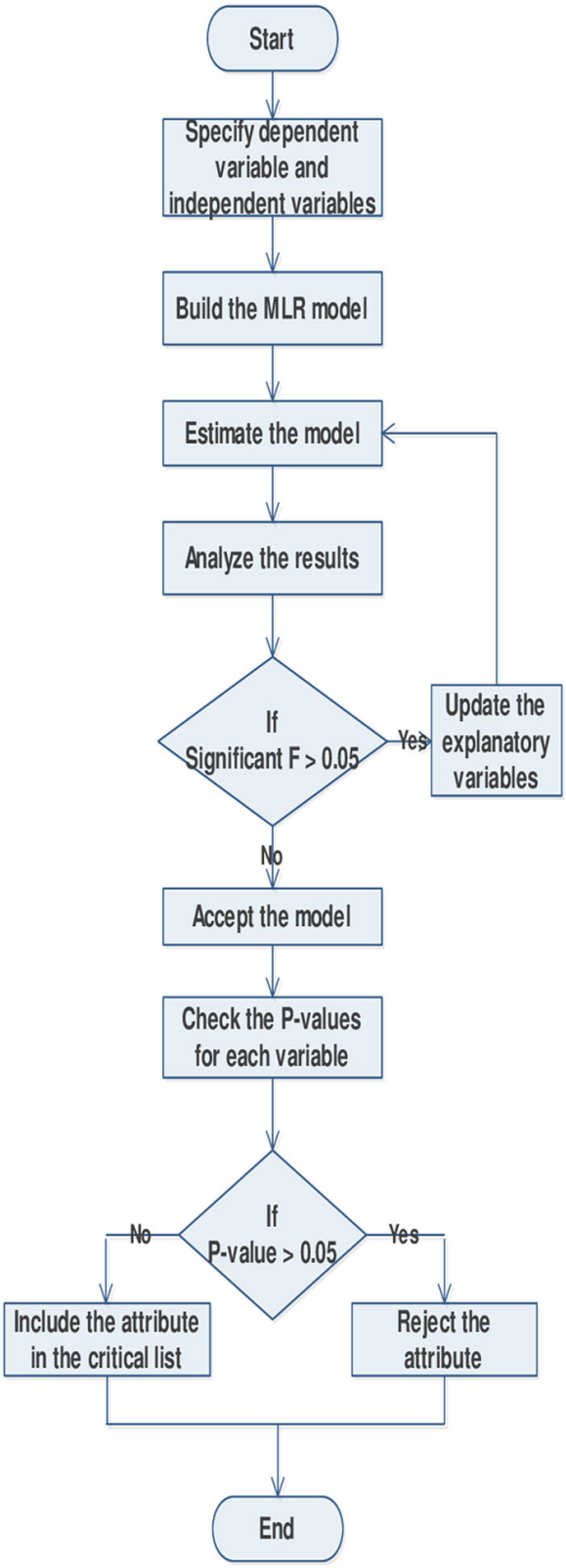


Figure 5.2: *Regression Process Flow*

EDA and statistics for Regression

Exploratory data analysis is an important step to take to explore the data and decide on the hypothesis. EDA's major objective is to validate the assumptions and insights into data. In the EDA phase, a data sanity check is also performed, involving handling of missing data and making transformations of variables as needed. In a nutshell, EDA assists in determining the narrative that the data conveys.

The goal of exploratory data analysis is to do the following:

- Check for any missing information, outliers and other errors in data.
- Understand the data set and ts inherent structure to the fullest extent possible..
- Discover a parsimonious model, that is, a model that describes the data with the smallest number of predictor variables possible.
- Before performing any model fitting or hypothesis testing, double-check all the assumptions made about the data structure.
- Make a list of outliers or other anomalies.
- Find parameter estimates as well as the associated confidence ranges or margins of error.
- Determine which factors or variable have the strongest effect on the outcome.

Following are the primary types of EDA techniques:

- **Univariate Non-graphical**: There is just one variable in a univariate non-graphic data analysis, making it the simplest kind of data analysis. Univariate non-graphical EDA's primary objective is to learn about the distribution of the sample data and draw inferences about the population. The study also includes the identification of outliers. Distributional features include the following:
    - **Central tendency**: It's all about the middle values when it comes to the distribution's central tendency. Statistics like mean, median, and mode are often used to assess central tendency, with mean being the most used. The median may be preferable when the distribution is skewed or there is a risk about outliers skewing the analysis.
    - **Spread**: Spread is a measure of how far from the centre we are while trying to locate information values. Standard deviation and variance

are two important metrics for measuring the spread. Standard deviation is the square root of the variance, while the variance is the mean of the square of the individual deviations or errors.

    - **Skewness and Kurtosis**: Another helpful univariate feature is the distribution's skewness and kurtosis. Skewness is a measure of asymmetry, while kurtosis is a more precise indicator of asymmetry when compared to a normal distribution.

- **Multivariate Non-graphical**: Typically, non-graphical multivariate EDA approaches are used to illustrate the association between two or more variables using cross-tabulation (contingency table).

    - For categorical data, cross-tabulation is a valuable tool and is the method of choice when dealing with more than two variables. A two-way table with column headings for one variable and row headings for the amount of the other two variables is created, and then the counts are filled with all subjects who share a corresponding pair of levels in the first variable. Here, correlation coefficients play a vital role. Correlation coefficients helps understand the type of relationship that exists between two variables. Pearson correlation coefficient tells the relative strength of the linear relationship between two variables. It ranges between –1 and 1. Values closer to –1 indicate a negative relationship, while those close to +1 indicate a positive relationship. Zero indicates that there is no significant linear relationship. To learn more about correlation coefficients, you can refer to any basic statistics contents and refer to the "mathisfun" website for a detailed explanation of various graphs.

    - To compare statistics across the quantity of categorical data, we establish statistics for quantitative variables individually for every level. Mean and median may be compared; mean comparison is an impromptu method, while median is a more reliable method as most real-time data will be skewed.

- **Univariate Graphical plots**: Although non-graphical approaches are quantitative and objective, they do not provide a complete picture of the facts; as a result, graphical methods demand a greater degree of subjective interpretation and are therefore, more time-consuming. Examples of univariate graphics include the following:

- **Graph with a histogram**: An essential graph is a histogram, which is a bar plot with each bar representing the frequency (count) or percentage (count/total count) of instances for a range of values. Histograms are one of the most straightforward ways to quickly discover a great deal about your data, including its central tendency, spread, modality, form, and outliers, in a short amount of time.
- **Plots with stems and leaves**: Stem-and-leaf plots are a simple alternative to histograms that can be used in many situations. They display all the data values and, as a result, the form of the data distribution.
- **Boxplots**: The boxplot is another univariate graphical approach that is quite useful. It is good when it comes to presenting information about central tendency and displaying robust measures of location and spread. However, it can be misleading when it comes to elements like multimodality, so be cautious when using them. Boxplots may be used in various ways, but side-by-side boxplots are one of the most commonly used.
- **Plots of quantile-normal distribution**: Graphical EDA's ultimate univariate approach is plotting the quantile-normal distribution, making it the most difficult. **Quantile-normal plot** (**QN plot**) and more generally, the **quantile-quantile plot** (**QQ plot**), are the two terms used to describe this type of plot. It is customary to evaluate how closely a certain sample follows a specific theoretical distribution. Non-normality can be detected, and skewness and kurtosis can be diagnosed using this method.

- **Multivariate Graphical plots**: Multivariate graphical data is a visual representation of correlations among two or more sets of information. A grouped bar plot can be used. Each grouped indicates a certain level of one variable, and each bar inside that group denotes how much is being measured for the other variable.

  Other types of multivariate graphics that are commonly used include the following:

  - When dealing with two quantitative variables, the **scatter plot** is the most important graphical EDA approach since it has one variable on the x-axis and one on the y-axis and hence, a point for every example in your dataset.

- A **time series graph or a run chart** shows data as a line over a certain period.
- A **heat map** is a graphical depiction of data in which values are represented by various colours.
- A **bubble chart** is a type of data visualisation in which several circles (bubbles) are displayed in a two-dimensional plot.

Figure 5.3 summarizes the various pre-processing steps involved in regression modelling, like renaming variables, EDA, and missing value imputation.

Figure 5.3: *Pre-processing Steps for Regression*

The hypothesis that each of the slopes has a zero value may be tested using regression to see whether there is a functional relation between the variables. Let's take an example; consider two variables:

- Here, y is the dependent variable, and it is a linear function of x. As a result, the independent variable y is reliant on x.
- Once the (x,y) pairs have been gathered, plot them on the x and y axes.
- The objective here is to create an equation for the straight line that is the closest to each point.
- Let the line drawn through the sample be denoted as $y=a+bx$.

- In most real-world situations, the line drawn will not pass through every point, and there will always be an error, referred to as the error term, indicating that many members of the population of pairs will not have the exact predicted value since many points do not lie exactly on the regression line. The error term is usually denoted as ***ε***, or **epsilon**, and you would see the regression equation denoted as follows:

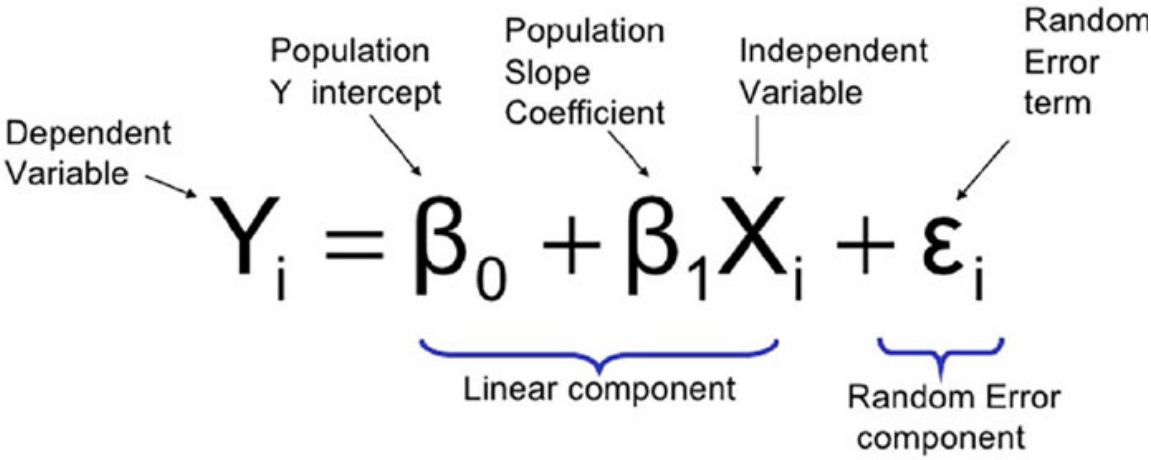


Figure 5.4: *Linear Regression Equation*

Summary of Regression and useful handouts

When given a set of independent variables, regression can be used to predict the value of a continuous (dependent) variable. It is parametric in nature since it is based on a collection of data and makes certain assumptions about that data. Given the dependent variable as follows, regression employs a linear function to estimate (predict) it mathematically:

$Y = \beta o + \beta 1X + \in$

Where,

Y = Dependent variable

X = Independent variable

β_o = Intercept

β_1 = Slope

$\in$ = Error

β_o and β_1 are coefficients

This is the linear regression equation in its simplest form. The term ‘linear’ refers to the fact that there is only one independent variable (X) involved. Multiple regression involves a large number of independent variables (Xs).

Y is the target variable that will be predicted.

X is the variable that will be used to make a prediction.

β_o is the intercept term; it is the prediction value when X = 0.

β_1 is the slope term that explains the change in Y when X changes by 1 unit.

$\in$ represents the residual value, i.e., the difference between the actual and predicted values.

The possibility of human mistake and bias of any form is an unavoidable component of the prediction process. It doesn’t matter how sophisticated the algorithm is; there will always be an (irreducible) error that serves as a constant reminder that “the future is uncertain.” Because regression is a parametric technique, it makes assumptions about the data. Any model whose predictions can be fully described by a fixed set of parameters is said to be parametric. Here are some of the assumptions of the regression model:

- In the presence of dependent and independent variables, a linear and additive connection exists. The term “linear” refers to the fact that the change in dependent variable caused by a unit change in independent variable is constant. The term “additive” refers to the fact that the influence of X on Y is independent of the other variables.
- It is assumed that the variables being used for the use case are not substantially correlated. In other words, there is little to no multicollinearity. Independent or feature variables are correlated with one other if they exhibit multicollinearity. If the features are highly correlated, it becomes exceedingly difficult to interpret the model and its coefficients.
- The error terms are assumed to have the same variance. This condition is called homoscedasticity. On the contrary, if the variance is unequal and scattered, it is called heteroscedasticity, which is mainly the result of the presence of outliers. Outliers can be detected during EDA phase, for example, using box plots.
- Error terms should be uncorrelated, such that the errors at one point in time are not indicative of errors at another point in time. Autocorrelation is a concept used to describe the correlation between error terms. When the assumption of uncorrelated error terms is violated, it impacts the efficiency

of the estimators or coefficients. A normal distribution is required or expected for its error components.

There are various uses for regression analysis, including the following:

- Multiple independent variables may be used to predict a certain outcome or dependent variable.
- Interaction terms can be examined to see whether one independent variable's influence is reliant on the value of another variable. When the impact of one variable relies on the value of another, it is termed as interaction effect. Assess interaction terms to determine whether the effect of one independent variable depends on the value of another variable. The interaction term may be thought of as a direct multiplication of the various combination of predictors in question. Say there are three predictors (y1, y2, y3) or independent variables, then interaction terms can be y1*y2, y1*y3, y2*y3, y1*y2*y3, and so on.

The business problems where regression can be used are as follows:

- **Prediction / Forecasting**: Among the most prominent applications of regression analysis in business is the prediction of future opportunities and risks. Customers' likely purchases are predicted using demand analysis techniques, such as those used in retail. However, when it comes to business, demand is not the only dependent variable to consider. Regressive analysis can predict much more than simply direct income in a given period. By projecting the number of consumers who will pass in front of a specific billboard, we could, for example, anticipate the highest bid for a particular advertising campaign. Insurance companies rely heavily on regression analysis to anticipate policyholder creditworthiness and the number of claims that may be submitted in a given period.
- **Capital Asset Pricing Model**: To determine the link between an asset's expected return and its market risk premium, the **Capital Asset Pricing Model** (**CAPM**) employs a linear regression model. Additionally, financial analysts use it to anticipate corporate returns and operational performance as well as in economics.
- **Competition Comparison**: The phrase is often used to refer to a comparison of a company's financial performance to that of a competitor. Also, it may be used to determine the correlation between two distinct firms' stock prices. Using this method, a company can figure out which

components of their business are influencing their sales in comparison to the competitive firm. Small businesses can benefit from these strategies since they can achieve rapid success in a short period.

- **Providing factual evidence for management decision-making**: Regression is valuable not just for giving factual information to support management decisions but also for identifying faults in judgement. A retail store management, for example, may believe that increasing the number of hours that customers can shop will greatly increase sales. However, Regression Analysis may indicate that the increase in income is insufficient to offset the rise in operating costs as a result of the increased number of working hours per week (such as additional employee labour charges).

Regression analysis (and other forms of statistical analysis) are being used by many companies and their chief executives to make better business choices and minimise their dependence on gut feeling and guesstimates. Businesses can adopt a more analytical management model, thanks to the usage of regression. Information overload is a persistent problem for both big and small businesses. In order to make the best judgments possible, managers can use regression analysis to filter through enormous volumes of data and uncover the most important elements.

Regression analysis has been used widely by organisations for a long time to transform data into meaningful information, and it continues to be an important asset for many leading sectors. Regression analysis is critical in today's data-driven world. Statistics and statistics that are unique to your company are referred to as "data" in this context. The advantages of regression analysis are that it helps you crunch the data to assist you in making better business decisions now and in the future, which is really beneficial.

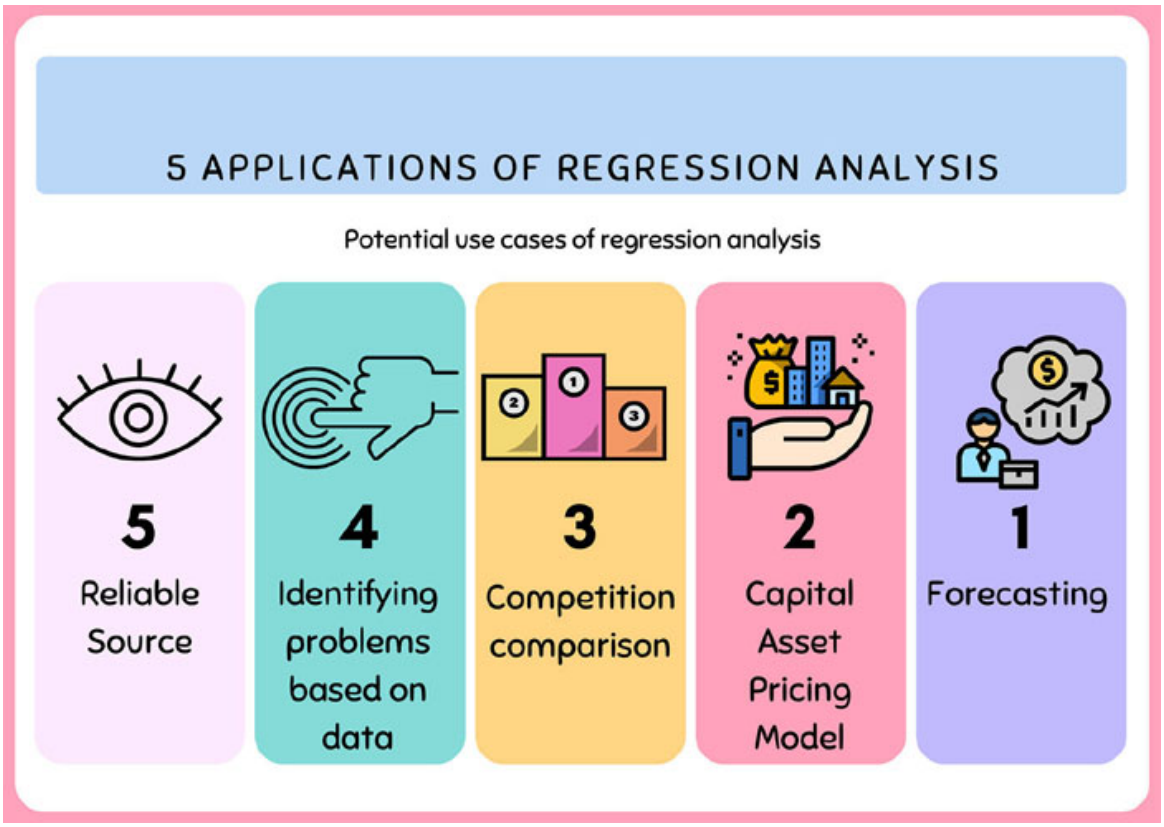


Figure 5.5: *Application of Regression Analysis*

Linear Regression using Orange – No Code

To practise linear regression on a real-world dataset, we will use an original dataset and apply the ideas we have learnt. We'll utilise the Housing dataset, which provides data on various dwellings in Boston. This data was included in the UCI Machine Learning Repository but has been withdrawn now. This dataset has 506 samples and 13 feature variables, and its purpose is to forecast the house's value using the provided characteristics.

To begin, let us download and install the open-source machine learning platform Orange. The following steps are based out of the standard documentation published for Orange v3. For additional details and tutorials, you can refer to the documentation section of the orangedatamining website.

Inputs:

Dataset: Housing data

Pre-processor: Selection of 70% of relevant features or independent variables using univariate regression analysis

Outputs:

Learner: Linear Regression, Random Forest, SVM and Artificial Neural Network

Coefficients: Regression coefficients

The Linear Regression widget creates a learner/predictor that uses input data to train a linear function. Predictor-response relationships may be discovered using the model. Next, you can train a regression model to predict house price in just five simple steps:

1. Select or load data
2. Select Test and Score widget
3. Connect Pre-process to Test and Score
4. Choose a model to train
5. Press apply to commit changes

When no extra pre-processors are provided by user, regression models use the default pre-processing technique. It does the following:

- Eliminates instances with unknown target values
- Continuizes categorical variables (a process of transforming variables in Orange version 3 into continuous ones with one hot encoding)
- Removes empty columns
- Use the arithmetic mean to impute missing values

However, you can use the pre-process widget for customised pre-processing methods. It can perform many actions for you in just a click, including but not limited to the following:

- Discretization of continuous values
- Continuization of discrete values
- Imputing missing values
- Selecting relevant features
- Selecting random features
- Normalize adjusts values to a common scale
- Randomization of instances

To learn more, refer to the Orange web page for easy-to-understand tutorials. The fun part here is that you can create models without writing a single line of code.

The following is a simple workflow with housing dataset.

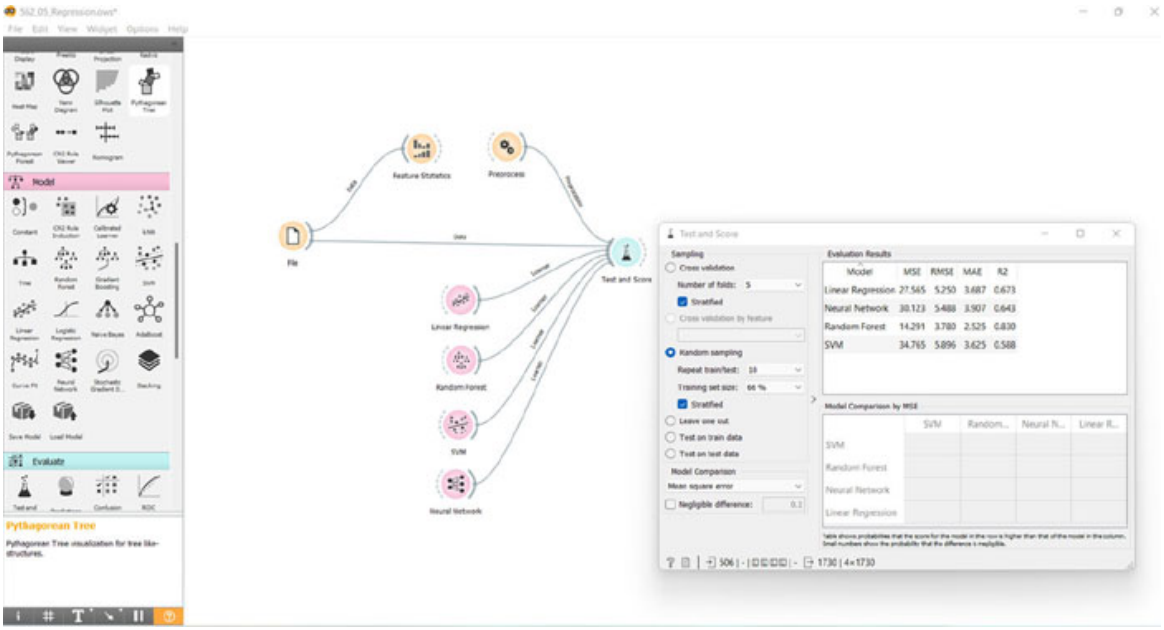


Figure 5.6: *Linear Regression Workflow Using Orange*

As a pre-processing technique, you can select relevant features either as a finite number of features or a percentage of the total, as shown in the following image. Multiple regression models, such as Linear Regression, Random Forest, SVM, and Artificial Neural Network, can be trained and their performance evaluated using Test & Score.

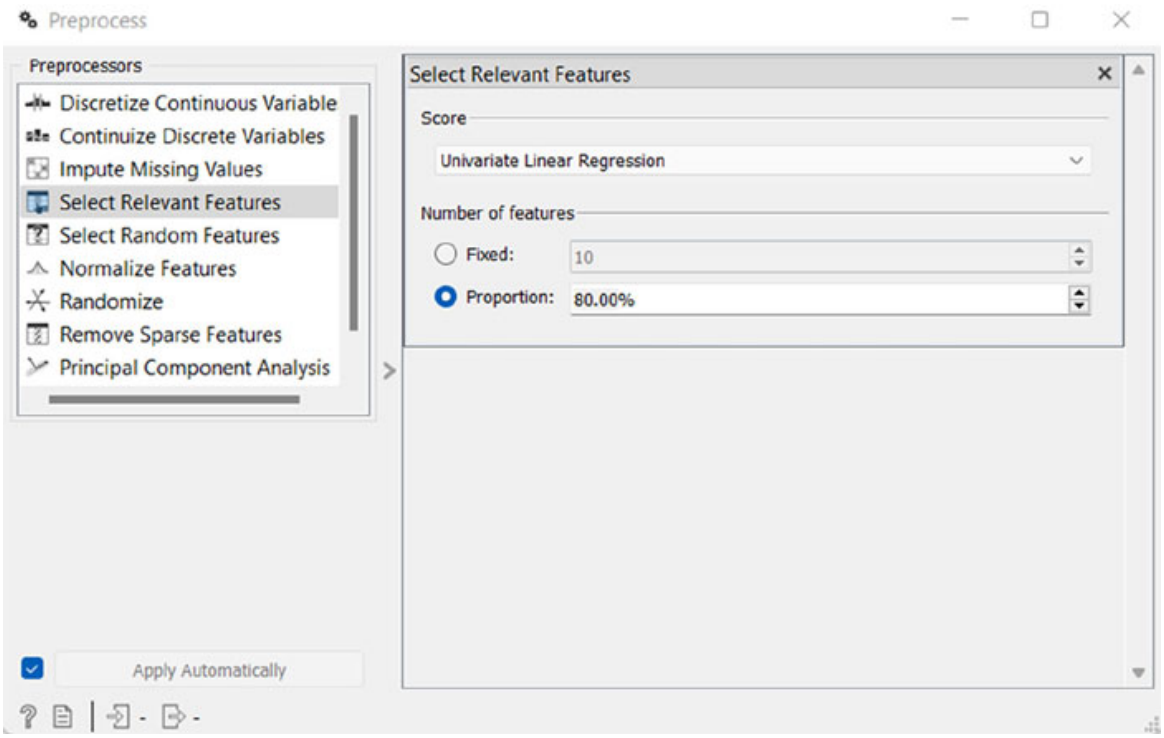


Figure 5.7: *Pre-processing Step Using Orange*

Conclusion

Regression is used to study the relationship between multiple variables. There's a good chance that linear regression would be the first algorithm you'd learn if you were beginning a profession in machine or deep learning. This technique is commonly used in data science and statistical domains to describe the link between a target and one or more predictor variables. Depending on the data being utilized, various regression approaches are available. Although linear regression is based on basic mathematical reasoning, its applications are used in real time across a wide range of industries to gain insights, including customer behaviour, business, and the elements that influence profitability. There is a slew of other uses for it, such as trend analysis and forecasting. Some of our day-to-day challenges may also be solved using linear regression, like decision-making, enhancing operational efficiency, reducing errors, and developing predictive analytics solution.

In the next chapter, we will learn about the next common machine learning algorithms, which is associated with categorical data to classify the data, for example, classifying a email as spam or not spam.

Points to remember

- Regression, in a nutshell, is a technique to find how well a set of data points fits a straight line. The best fit line is the one for which the total prediction error is as small as possible.
- Regression models are quite helpful when it comes to finding the correlation between two or more variables, drawing conclusions, making predictions, and so on.
- Key Assumptions of Linear Regression
    - **Normality**: The error terms are expected to be normally distributed.
    - **Low or no Multicollinearity**: Variables must be independent of each other.
    - **Homoscedasticity**: The residual terms are believed to have the same variance.
    - **No Autocorrelation**: There is no autocorrelation between the residuals.

Multiple choice questions

1. When attempting to describe the relation between two quantitative variables, a _____ is preferable.
    a. Scatterplot
    b. Density curve
    c. Boxplot
    d. Histogram
2. What does the regression coefficient's (β1) negativity imply?
    a. The data is not linear
    b. Most Y values lie under the fitted line
    c. Increase in the value of X will result in a change of Y in the opposite direction
    d. The correlation coefficient is as little as possible
3. Predicting whether a customer will churn is a regression use case.
    a. True

b. False

4. If the correlation coefficient value for two variables is 0, then there is no relationship between the two variables.

   a. True

   b. False

Answers

Question Number	Answer
1	A
2	C
3	B
4	B

CHAPTER 6

Classification

No classification is complete as you can see ML algorithms evolving each day, and it is now almost impossible to say which one is better.

Classification can simply be compared to how a human categorizes and arranges a wardrobe based on garment type or colour. The variable to predict may be as simple as “yes” or “no” and may fall into one of the categories. This category (yes or no), which is predicted, is also called class or class label. For example, understanding the number of workout hours and calories consumed determine whether a person will lose weight. Decision boundaries is defined as the surface that can separate the data points into distinct categories or classes. Finding the decision boundary is the overarching purpose of classification models. Comparing classification with regression, it is important to note that classification is focused on class labels, whereas regression is focused on quantities. Some of the common classification algorithms are classified into three types: linear models, tree models, and artificial neural nets.

Structure

In this chapter, we will cover the following topics:

- Basics of classification
- Classification process flow (identify problem statement, check data availability, choose right pre-processing tool, and define end goal)
- EDA & Statistics for Classification
- Classification using Orange

Objectives

You’ll learn how to specify the input and output of a classification model and how to solve both binary and multiclass classification problems after going through this chapter. You’ll also understand how a logistic regression model and a non-linear decision tree model can be implemented. Finally, you’ll learn how to evaluate your models using several assessment criteria.

Get started with classification

In general, the act of classifying objects into respective categories is the goal of classification. Consider the possibility that your mobile phone or computer may discern the difference between you and another person. Does it ring a bell? Yes, facial feature comparison and classification: the one feature most of us use daily in our mobile phones or laptops.

Guess what? It might sound interesting to you now!

In machine learning, classification is the process by which fresh observations may be assigned to one or more categories based on pre-labelled observations commonly termed as training data. Here, the machine learns to categorise an observation into a specific category. For example, yes or no, cat or dog, male or female, and so on. It uses pre-labelled data to learn and detect the patterns. Fraud detection, credit approval, image classification, customer churn prediction, and text classification are some of the applications of classification. Let us see some of the common types of classification and classification algorithms.

There are three forms of classification based on the number and level of classes in the dataset. They are as follows:

- Binary classification
- Multi-class classification
- Multi-label classification

Based on the algorithms used, classification algorithms can be classified as follows:

- Linear model
- Tree-based model
- Neural network

Based on the learning approach, classification algorithms can be classified as follows:

- Lazy or slow learning
- Eager learning

We'll get into the specifics of these groupings later, but for now, let's establish a few key terminologies. There are other similar terms in machine learning. Whether you're working on a project that involves machine learning or you just

want to learn more about this field of data science, you can use the following words to describe it:

- **Model**: It is the mathematical depiction of a real-world process. A predictive model predicts future outcomes using data from the past. Going forward, the word model will be used in reference to a predictive model.
- **Target**: An example of a target might be a variable or a class name in a model that you want to forecast, or the output of a model.
- **Feature**: Predictor variable, or independent variable, is another term for feature. A feature is something that can be observed and measured by a model. Additionally, features can be engineered in various ways, including merging or adding additional information.
- **Training**: It is the process of feeding use-case relevant data to a machine learning algorithm so that it can learn the patterns and more.
- **Testing**: It is the process of assessing a model based on the results of a test. The performance of various models is measured and benchmarked using a test dataset, which is independent from the training set but has the same structure.
- **Algorithm**: It is a procedure used to solve an issue or make a computation.

Binary classification

There are just two categories in this form of classification. True or false and high or low are the norm. Examples of where this categorization might be utilised are tumour detection, fraud detection and email spam detection.

Multi-class classification

Classification task with more than two labels or categories is multi-class classification. Here, each input will have only one output class.

Multi-label classification

Observations that can have several labels at the same time are classified this way. This find its use especially in object detection where there is more than one thing, like a cat, dog, car and a human, present in a single image.

Linear model

As the name suggests, this type of analysis relies on a simple formula to identify the “best fit” line between two sets of data. In comparison to more recent algorithms, linear models are considered “old school.” However, they may be trained rapidly and are, typically, easier to interpret.

Tree-based model

The term “tree-based” relates to decision trees, i.e., a series of branching processes. There are many ways to represent a choice in a decision tree. Although a decision tree is easy to understand, it’s a poor predictor. Random forest and gradient boosting are two of the most common tree-based models. They are also called ensemble learning methods for generating stronger predictions from many trees. Ensemble learning is a generic term that refers to a model that produces predictions using a collection of various models. By merging many models, the ensemble model becomes more flexible and less sensitive to data. All tree-based models can be utilised for regression or classification. In general, they can also deal with non-linear correlations rather effectively.

Neural network

Neural network is a biological structure through which neurons communicate with one another. In the realm of machine learning, ANN stands for Artificial Neural Networks, a computational model inspired by the functioning of human or biological neurons. These are the best-in-class algorithms available today but are slow to train as they are data thirsty.

Lazy learning

Lazy learning, also known as instance-based learning, is a machine learning method where the training data is not explicitly used to build a model or generalize to new data. Instead, the training data is simply stored, and when a new testing instance is presented, the algorithm retrieves the most similar instances from the stored training data and uses them to make predictions or classify the new instance. It’s named lazy since it waits until it’s essential to construct a model. Lazy learning doesn’t learn a discriminative function from the training dataset. Instead, it “memorises” what it learned.

Eager learning

Eager learning occurs when a machine learning algorithm constructs a model shortly after obtaining a training dataset. It begins building the model as soon as it receives the dataset.

Process flow of classification

As discussed, classification is a technique used for categorizing data into distinct and unique classes, and then labels are assigned to each class. The primary goal of classification is to determine where fresh data should be added by analysing the training set and drawing clear boundaries between classes. In simple words, it is a

method to predict the target class. Some of the common classification algorithms are as follows; each of them has a standard way of implementation:

- **K-Nearest Neighbours (KNN)**: Similar to tree-based models, KNN can be used in both classification and regression prediction problems. It is easy to interpret the predictions of KNN algorithm, and the calculation time is also low. KNN is a non-parametric method for lazy learning as it doesn't perform training when the training dataset is provided. It just holds data and does not do any computations throughout the training period. It constructs a model solely in response to a query on the dataset. It is a non-parametric technique since no assumptions regarding the distribution of the underlying data are made. The benefits of KNN include the following:
    - It is simple to use and implement
    - It can also be used to address classification and regression problems

      For non-linear data, it's a great fit since there are no assumptions about the data
    - It can deal with multi-class instances
    - It does not make any assumptions about the data

The following image shows how KNN uses the stored training data (which is a set of glasses containing water, wine, beer and so on) and look up patterns matching the new data point.

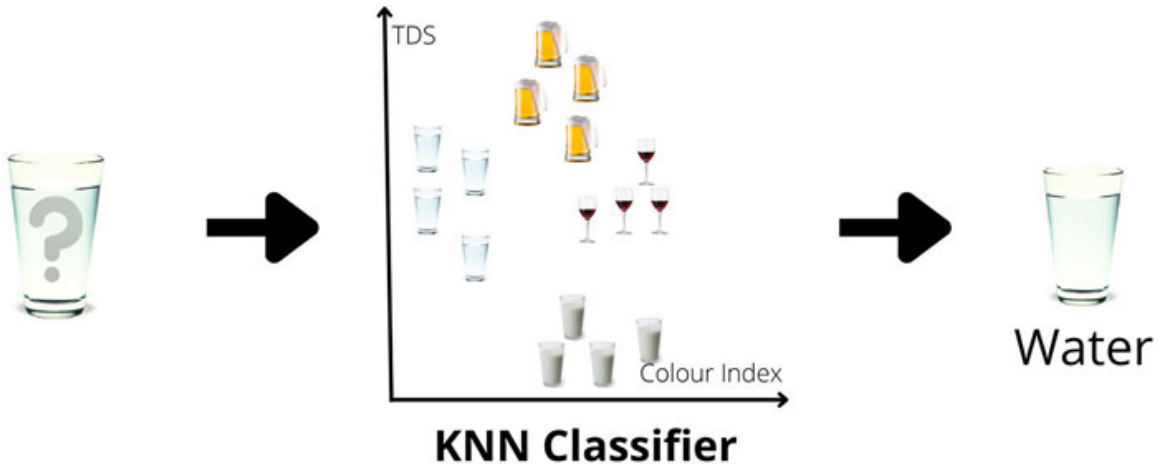


Figure 6.1: *KNN Classifier Application*

- **Naive Bayes Classifier**: A collection of Bayes' Theorem-based Bayes classifiers are referred to as Naive Bayes classifiers. In fact, it's not just one algorithm but a family of algorithms, all of which have a fundamental

principle: each pair of attributes being classified is independent of the other. All features are treated as separate and do not interact with each other, so the probability of a sample being classified into a certain class is determined by each characteristic individually and equally. An easy-to-use and quick Naive Bayes classifier can handle huge datasets with high dimensionality and perform well. Real-time applications benefit from the Naive Bayes classifier's ability to withstand high levels of background noise. It is best to eliminate all correlated features from the model to avoid overemphasizing the impact of correlated features in the Naive Bayes classifier.

Some of the key facts about Naive Bayes algorithm are as follows:

- Text categorization using a large, multi-dimensional training set uses this technique extensively.
- Fast machine learning models may be built using the Naive Bayes Classifier, one of the most straightforward and efficient classification methods.
- A probabilistic classifier makes predictions based on the likelihood of a given object occurring.
- It is used in the following applications: spam detection, sentiment analysis, and article categorization.

Let's take an example of using Naive Bayes classifier on a synthetic cargo ship arrival status dataset where the target variable is "cargo ship arrival status". This is a basic example of deciding whether the ship will arrive on time based on particular weather conditions.

In this dataset, there are four attributes = [Season, Size, Wind Condition, Wave Height] and 4 classes = [On Time, Late, Very Late, Did Not Arrive]. The attributes from the data and available classes allow us to make an informed guess at a more probable categorization for any unknown data; for example, given the season, size, wind condition and wave height, what could be the possible arrival status of the cargo ship?

Season	Size	Wind Condition	Wave Height	Cargo Ship Status
Spring	Small	Calm	Moderate	On Time
Spring	Ultra Large	Calm	Rough	On Time
Spring	Ultra Large	Calm	Moderate	On Time
Winter	Ultra Large	Storm	Rough	Late

Summer	Large	Hurricane	Moderate	On Time
Spring	Medium	Hurricane	Moderate	Very Late
Winter	Large	Storm	Rough	On Time
Autumn	Large	Hurricane	Moderate	On Time
Spring	Ultra Large	Storm	High	Very Late
Spring	Large	Calm	Rough	On Time
Summer	Small	Storm	High	Did Not Arrive
Spring	Large	Storm	Rough	On Time
Spring	Ultra Large	Hurricane	Moderate	Late
Spring	Large	Storm	Moderate	On Time
Spring	Ultra Large	Hurricane	High	Very Late
Summer	Medium	Storm	Rough	On Time
Spring	Medium	Calm	High	On Time
Winter	Small	Hurricane	Rough	On Time
Spring	Small	Hurricane	Moderate	On Time
Spring	Small	Hurricane	High	On Time

Table 6.1: *Cargo Ship Arrival Status Synthetic Dataset*

As Bayes theorem is helpful in determining the probability of a hypothesis with prior knowledge, we will see how that works in this example:

P(A|B) = P(B|A) * P(A)] / P(B)

Where,

- P(A) and P(B) are called prior probabilities, for example, P (On Time) = 14/20
- P(A|B), P(B|A) are called posterior probabilities, for example, P (On Time in Spring) = 9/20

Notes:

- P(B) is called the evidence (also the total probability), and it is a constant.
- The probability P(A|B) (also called class conditional probability) is, therefore, proportional to P(B|A) * P(A).
- Thus, P(A|B) can be taken as a measure of A given B.

- P(A|B) ≈ P(A|B) * P(A)

Consider this example data collection that contains information about weather conditions and develops a frequency table from it. Following that, let's build a likelihood table by determining the probabilities of the given characteristics and eventually computing the posterior probability using Bayes theorem. With reference to the previous cargo ship arrival dataset, let's tabulate all posterior and prior probabilities. The following table summarizes the posterior probability of attribute season and size against each class or arrival status:

		Class (Cargo Ship Status)			
Attribute		On Time	Late	Very Late	Did not Arrive
Season	Autumn	1/14=0.071	0/2=0	0/3=0	0/1=0
	Spring	9/14=0.643	1/2=0.5	3/3=1	0/1=0
	Summer	2/14=0.143	0/2=0	0/3=0	1/1=1
	Winter	2/14=0.143	1/2=0.5	0/3=0	0/1=0
Size	Large	6/14=0.429	0/2=0	0/3=0	0/1=0
	Medium	2/14=0.143	0/2=0	1/3=0.333	0/1=0
	Small	4/14=0.286	0/2=0	0/3=0	1/1=1
	Ultra Large	2/14=0.143	2/2=1	2/3=0.667	0/1=0

Table 6.2: *Cargo Ship Arrival Status Dataset – posterior and prior probabilities*

The following table summarizes the posterior probability of attribute wind and wave conditions against each class or arrival status.

		Class (Cargo Ship Status)			
Attribute		On Time	Late	Very Late	Cancelled
Wind	Calm	5/14=0.357	0/2=0	0/3=0	0/1=0
	Hurricane	5/14=0.357	1/2=0.5	2/3=0.667	0/1=0
	Storm	4/14=0.286	1/2=0.5	1/3=0.333	1/1=1
Wave Height	High	2/14=0.143	0/2=0	2/3=0.667	1/1=1
	Moderate	6/14=0.429	1/2=0.5	1/3=0.333	0/1=0
	Rough	6/14=0.429	1/2=0.5	0/3=0	0/1=0
Prior Probability		14/20=0.70	2/20=0.10	3/20=0.15	1/20=0.05

Table 6.3: *Air Traffic Dataset – posterior and prior probabilities continued*

For instance, if a large cargo ship is sailing in summer with calm sea (wind) and moderate wave conditions, what would be the cargo ship arrival status?

- On Time: 0.7 X 0.143 X 0.429 X 0.357 X 0.429 = 0.0066
- Late: 0.1 X 0 X 0 X 0 X 0.5 = 0
- Very Late: 0.15 X 0 X 0 X 0 X 0.333 = 0
- Did Not Arrive: 0.05 X 1 X 0 X 0 X 0 = 0

The first case seems to be the strongest, with the highest value; hence, the classification would be “On Time”.

Some of the advantages of the Naive Bayes Classifier are as follows:

- Using Naive Bayes, which is quick and simple to implement.
- For binary and multi-class classifications, it can be employed.
- It is a favoured technique for text categorization.

As Naive Bayes assumes that all the features are independent or unrelated, it cannot learn the relationship between features.

- **Decision Trees**: In machine learning classification, the decision tree is, by far, the most successful and widely utilised technique currently accessible. It resembles a flowchart, with internal nodes representing attribute tests, branches representing the outcomes of the tests, and leaf nodes (terminal nodes) representing the target or class name for each attribute tested. Decision trees categorise instances by sorting them along the tree from root to leaf node, at which point the instance’s categorization is determined. As seen in the preceding diagram, an instance (say a decision tree on whether or not to play cricket) is classified by starting at the root node of the tree, verifying the attribute supplied by this node, and then proceeding along the tree branch according to the attribute’s value. After that, the same thing happens for the subtree that starts at the new node. Each internal node in a decision tree represents an attribute test, and each branch represents an attribute test outcome. During learning, a tree is built by partitioning the source data set based on the outcome of an attribute value test. This approach is followed on each derived subset recursively. Recursion is the process of defining a function or calculating a number by the repeated application of an algorithm. It ends when each subset at a node has the same value of the target variable or when further splitting does not improve the predictions. This can increase the depth of the trees. Pruning is a technique

that prevents a decision tree from expanding to its maximum depth. Pruning is crucial because a decision tree trained to its maximum depth would likely overfit the training data.

The following are some of the advantages of decision tree methods:

- Decision trees can generate rules that are easy to comprehend.
- Decision trees are used to conduct categorization without the need for a lot of computing.
- Decision trees are resistant to outliers and reduce the time required for pre-processing the data.

There are certain limitations to decision tree methods, such as the following:

- Decision trees are less suitable for estimating jobs where the goal is to estimate the value of a continuous characteristic. The use of decision trees in classification tasks with many classes and a small number of training instances increases the likelihood of making mistakes.
- The training of a decision tree can be time-consuming and computationally expensive. Computing resources are required to complete the process of building a decision tree. Each potential splitting field must first be sorted at each node to identify the best split for that field, which makes the decision tree greedy.
- Pruning algorithms can be time-consuming due to the large number of candidate subtrees that must be produced and compared.

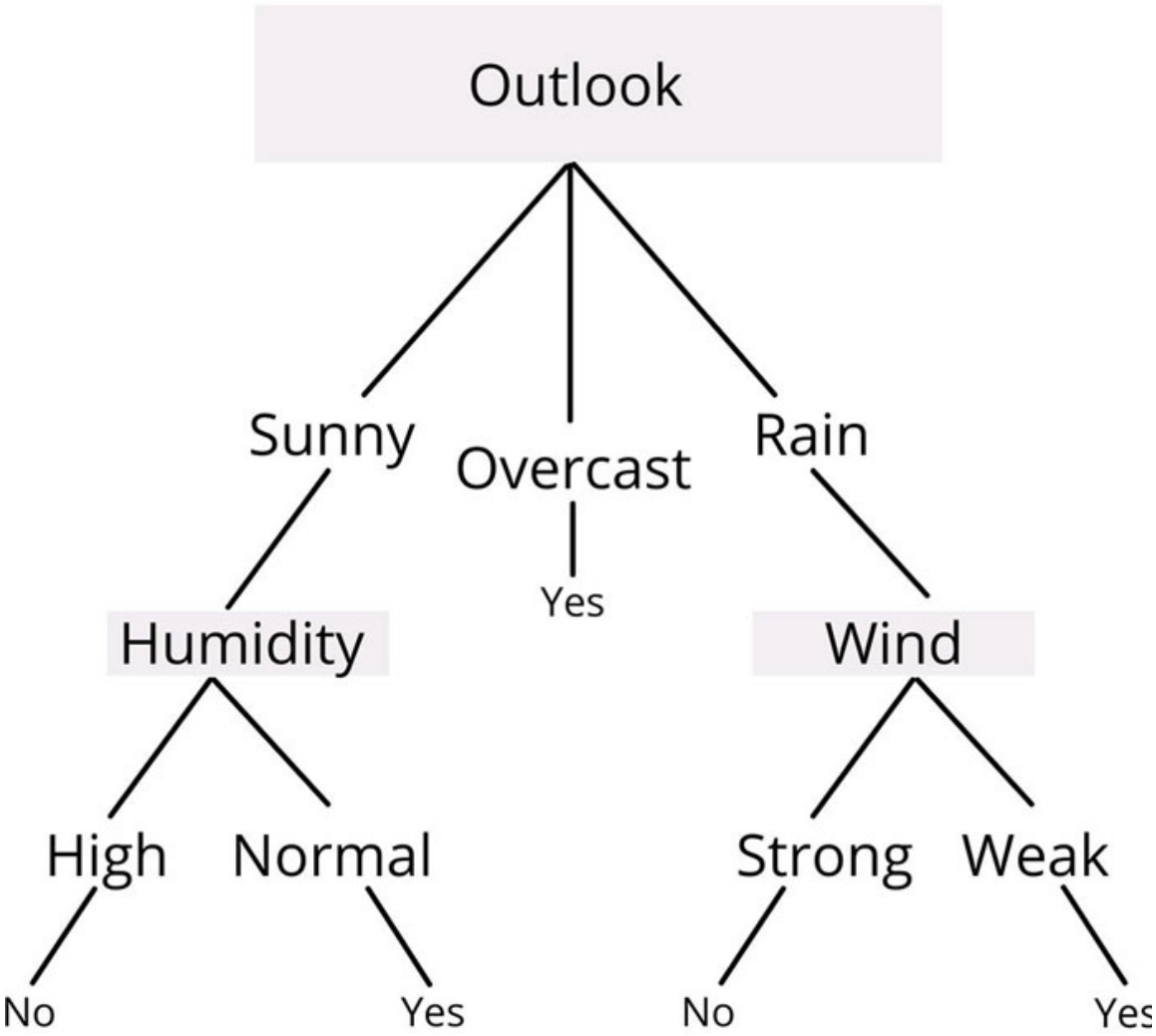


Figure 6.2: *Sample Decision Tree Structure (Reference: geeksforgeeks)*

- **Random Forest**: The Random Forest method is an effective and flexible supervised machine learning technique that builds a “forest” of decision trees from individual ones. It is applicable to classification and regression use cases. Random Forest is based on the idea that a collection of independent models (the trees in the decision-making process) would be more effective than any of them would be alone. Each tree “votes” with its own label, and whichever categorization receives the most “votes” is the one selected by the forest.

  The ‘forest’ formed by the random forest method is trained using bagging or bootstrap aggregation. To improve the accuracy of machine learning algorithms, a meta-algorithm known as “bagging” aggregates all the predictions across multiple algorithms on different subsets of data. Random forest can be considered as an extension of bagging as the model (random

forest) uses multiple decision trees to make predictions about the result. Later, it averages the output from various trees to make final decision. Bagging is the practice of employing multiple samples of data (training data) rather than a single sample of data. A training dataset comprises of data that is used to create future predictions. Precision and tree count are directly related, which means that as the tree count grows, precision improves as well. The limitations (chances of overfitting) of a decision tree algorithm may be circumvented using a random forest method as it uses multiple trees. At each node's splitting point in a random forest tree, a subset of features is chosen at random.

The decision trees generate various distinct outputs depending on the subset of training data that is input into the random forest algorithm. These outputs are aggregated by taking the majority votes, and the output with the highest vote is chosen as the final output. Rather than having a single decision tree, the random forest contains a large number of decision trees. The following figure will give an idea of how decision tree operates.

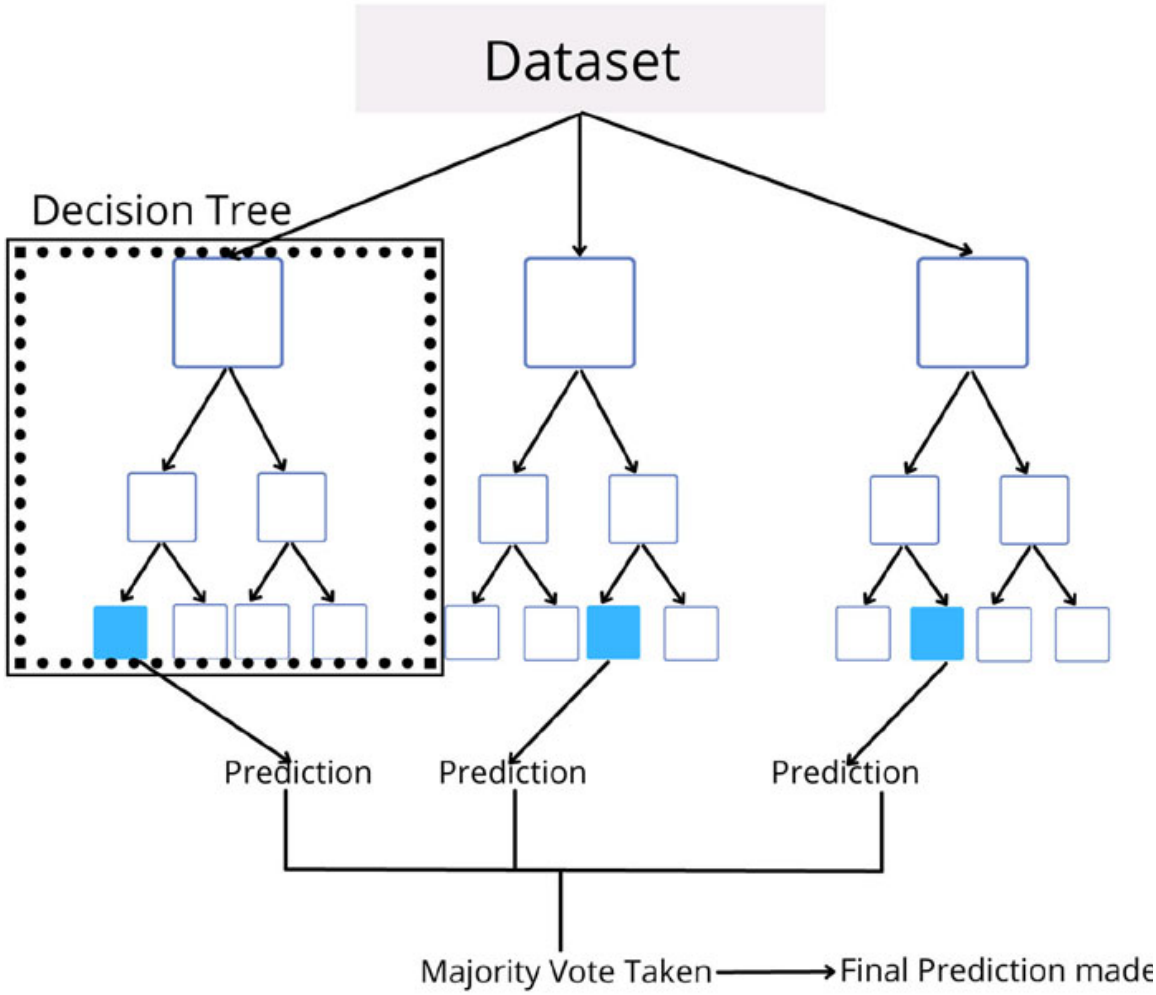


Figure 6.3: *Operation of the Decision Tree*

For example, consider a training dataset that contains numerous fruits, such as bananas, apples, pineapples, and mangoes, and it also has other vegetables and fruits. Based on its classification results, the random forest classifier uses subsets of this dataset for each tree. Each decision tree delivers an output. For instance, it is predicted that trees 1 and 2 would bear apples. The inference would be a banana if a different decision tree (n) was used. To create the final prediction, the random forest classifier combines the results of majority voting. The apple has been selected as the forecast by most decision trees. As a result, the classifier selects apple as its final prediction.

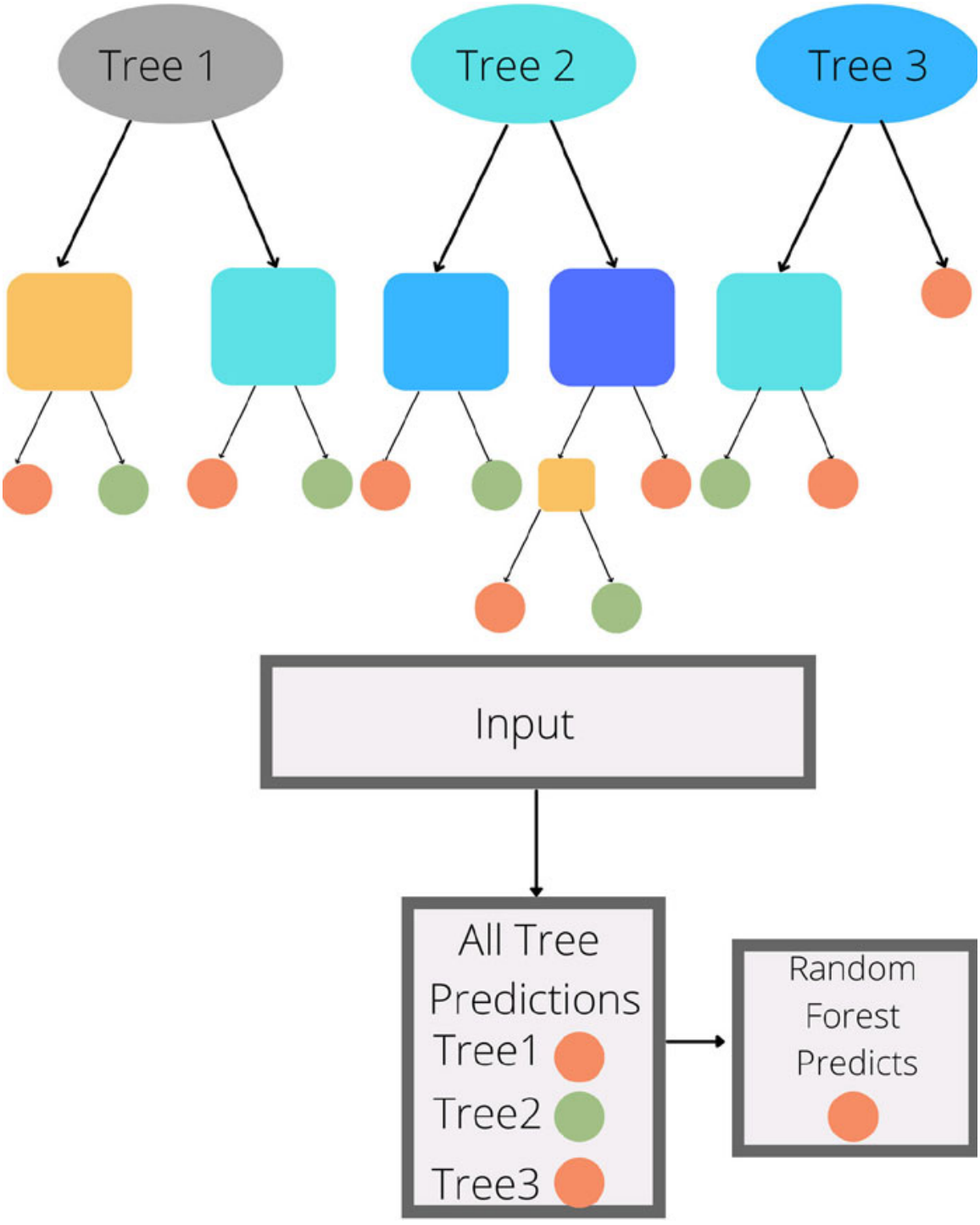


Figure 6.4: *Decision Tree Algorithm*

The advantages of using a random forest are as follows:

- A random forest generates more accurate predictions than a decision tree.

    - It can handle missing data efficiently.
    - The random forest method outperforms the decision tree algorithm in terms of overfitting.

  Random forest has several disadvantages as well:

    - It consumes more computational resources compared to a decision tree.
    - When compared to a decision tree algorithm, it is complex to visualize the results.

- **Support Vector Machine**: For both classification and regression tasks, Support Vector Machines (or simply, SVMs) are one of the most popular Supervised Learning techniques. To enable us to easily categorise new data points, SVM's goal is to determine the optimum line or decision boundary to split n-dimensional space into classes. A hyperplane is a boundary that represents the border of the best decision-making boundary. SVM selects the two points/vectors that will aid in the creation of the hyperplane in the first place. Detecting support vectors and the Support Vector Machine are two terms used to describe this kind of scenario. *Figure 6.6* illustrates how a hyperplane is used to categorise two distinct groups using a decision boundary:

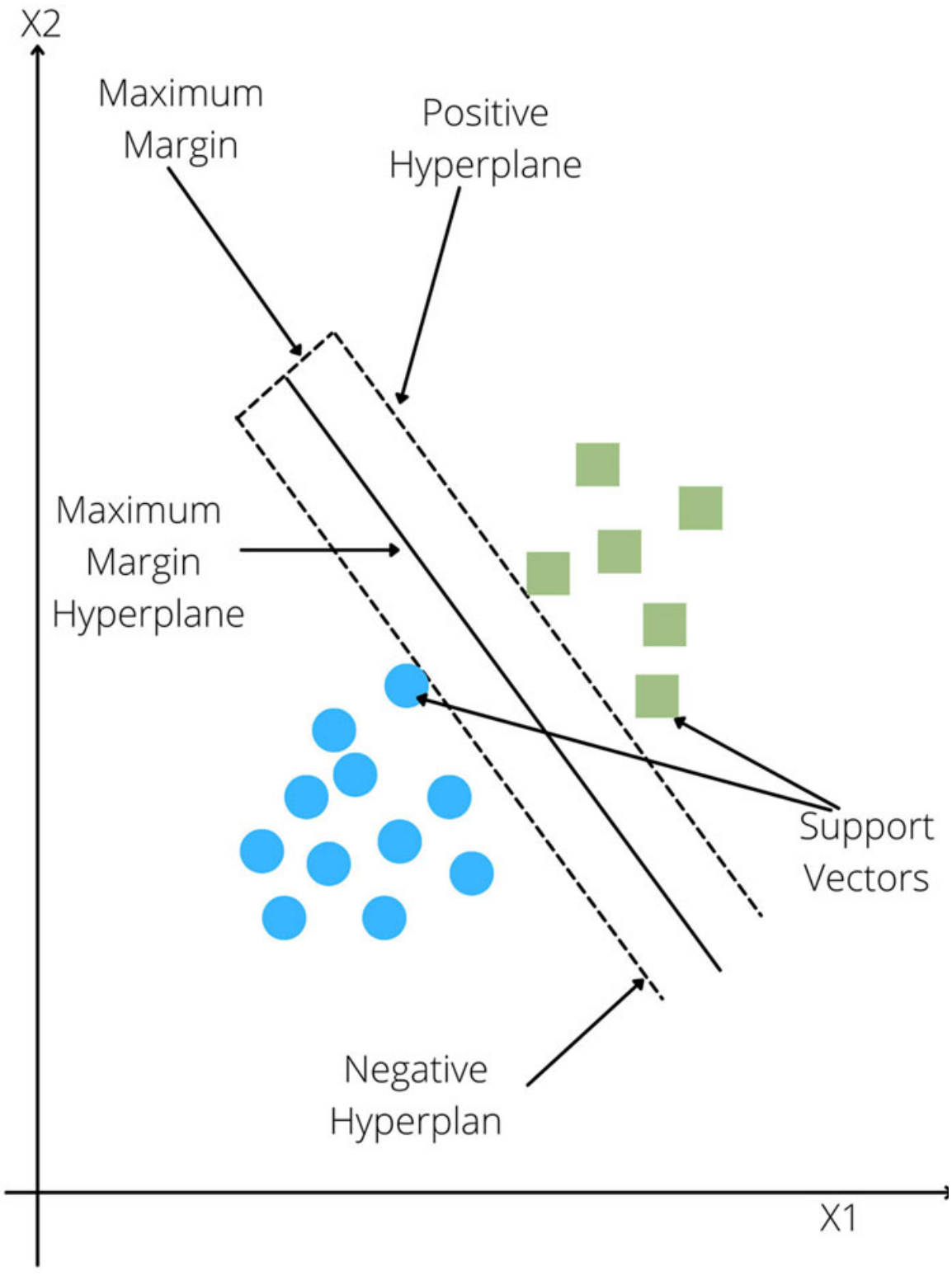


Figure 6.5: *SVM Hyperplane*

Consider this scenario: The SVM approach may be used to create a model that can accurately assess whether a given animal is a cat or a dog when we

come across a strange cat with certain features shared by both. In order to get our model used to the many different traits of cats and dogs, we'll first train it using a huge number of images (i.e., features extracted from images like colour, intensity, and pixel) of these animals. As a result, support vectors will notice the instances of cat and dog because they will establish a decision boundary between these two variables (cat and dog) and choose only extreme occurrences (support vectors). The support vectors that were utilised to identify a cat will be used to determine its identity.

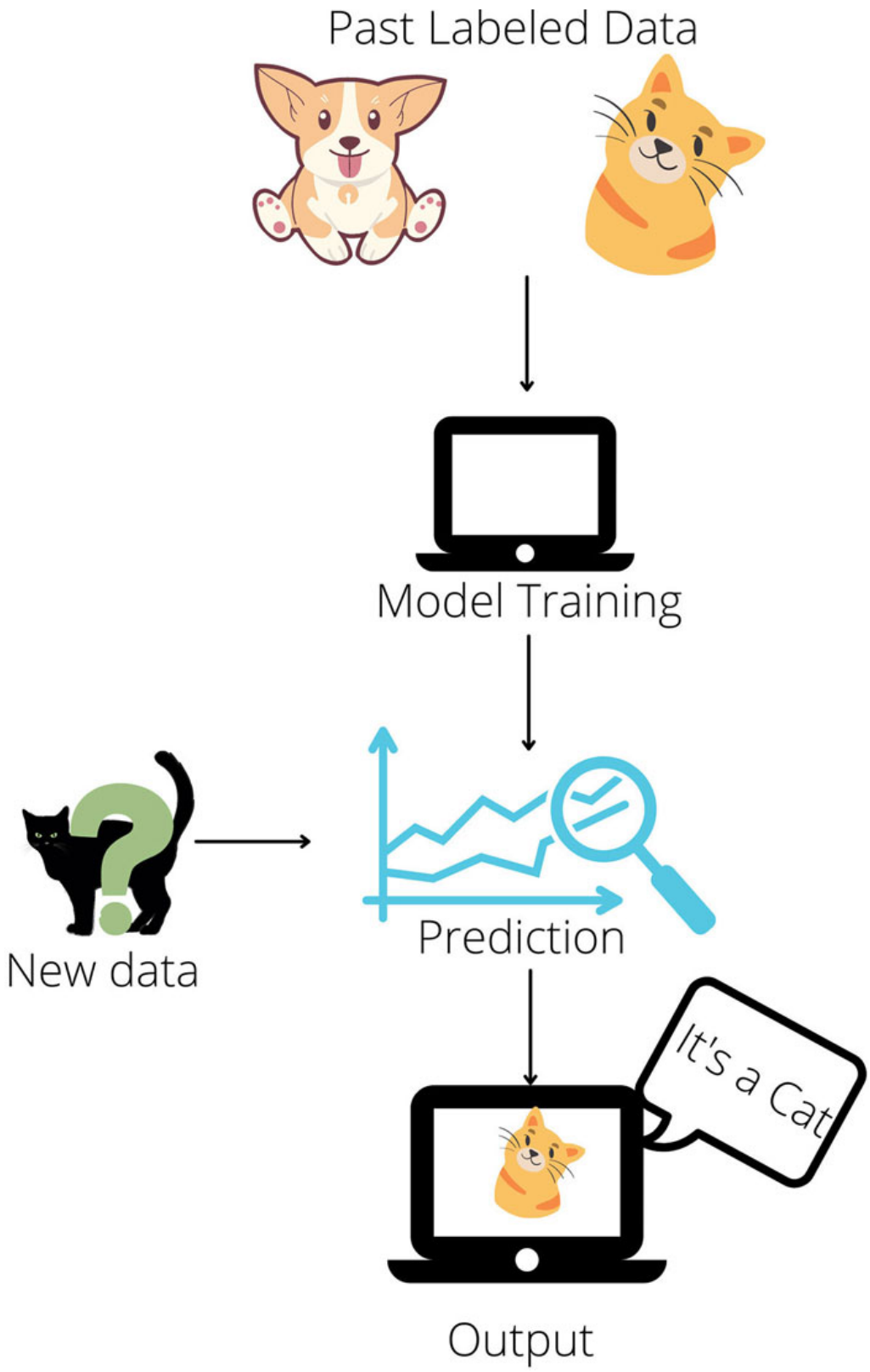


Figure 6.6: *Example of a Classifier*

The SVM approach may be used for face identification, image classification, text categorization, and more. SVMs can be divided into two main categories:

- Linear SVM is an SVM classifier used to classify linearly separable data, which suggests that a dataset can be separated into multiple groups using a single straight line.
- Non-linear data can't be categorised using a straight line, so we need a classifier, called a non-linear SVM classifier, to deal with it.

EDA and Statistics of Classification

Exploratory data analysis (**EDA**) is a technique used by data scientists to study and investigate large datasets to summarise their essential properties. Data visualisation techniques are frequently employed in this process. EDA assists data scientists in determining the most effective way to alter data sources to obtain the answers they require, making it easier for them to detect patterns, identify anomalies, test hypotheses, and verify assumptions. EDA is primarily used to investigate what data can disclose beyond the formal modelling or hypothesis testing work and to gain a deeper knowledge of the variables in the data collection and the relationships between those variables. You will be able to determine whether the statistical procedures you are evaluating for data analysis are suitable for your scenario as a result of this information.

Exploratory Data Analysis is a method of conducting data analysis that uses a range of approaches in order to develop a sense of intuition regarding the data.

- Make sure you're not losing your mind (in order to ensure that the insights we are making are derived from the correct dataset)
- Determine where information is lacking
- Examine the data to see if there are any outliers
- Summarise the information

The following are examples of statistical functions and approaches that can be performed with EDA tools:

- In order to construct graphical representations of high-dimensional data, including numerous variables, techniques like clustering and dimension reduction can be used.

- Each field in the raw dataset is shown as a single-variable visualisation, along with summary statistics.
- Using bivariate visualisations and summary statistics, you may determine the link between each variable in the dataset and the target variable that you're interested in learning more about. Multivariate visualisations are used to map and comprehend the interactions between distinct fields in the data and between different fields in the data.

For further details on EDA, refer to the details in the previous chapters, as the methodology remains same.

Classification using Orange

Classification tree (decision tree; refer to the Orange documentation for details) is a straightforward classification system that divides data into nodes based on the purity of each class. It is considered a forerunner of the Random Forest algorithm. Its classification tree in Orange software was developed in-house and is capable of dealing with both discrete and continuous datasets.

- When the learner is created, it may be given a name that will be used to distinguish it in other widgets when it is later utilised in them. The default name for this structure is "Classification Tree."
- The tree's parameters are as follows: The binary tree is created by inducing a binary tree (split into two child nodes).
    - **Minimum number of instances in leaves**: If this option is selected, the algorithm will never generate a split that places fewer than the specified number of training examples in any of the branches.
    - **Do not divide subsets smaller than**: This option prevents the algorithm from splitting nodes that have fewer occurrences than the specified number. After a defined majority threshold is reached, the nodes are no longer split.
    - **Limit the maximum tree depth to [number of nodes]**: It restricts the depth of the classification tree to the number of node levels specified.
- Write a report on your findings. After making your changes, you must click on apply, which will include the new learner in the result and, if training examples are provided, will generate a new classifier and include it in the output, along with the new learner. In the alternative, check the box on the left, and changes will be communicated to you without your intervention.

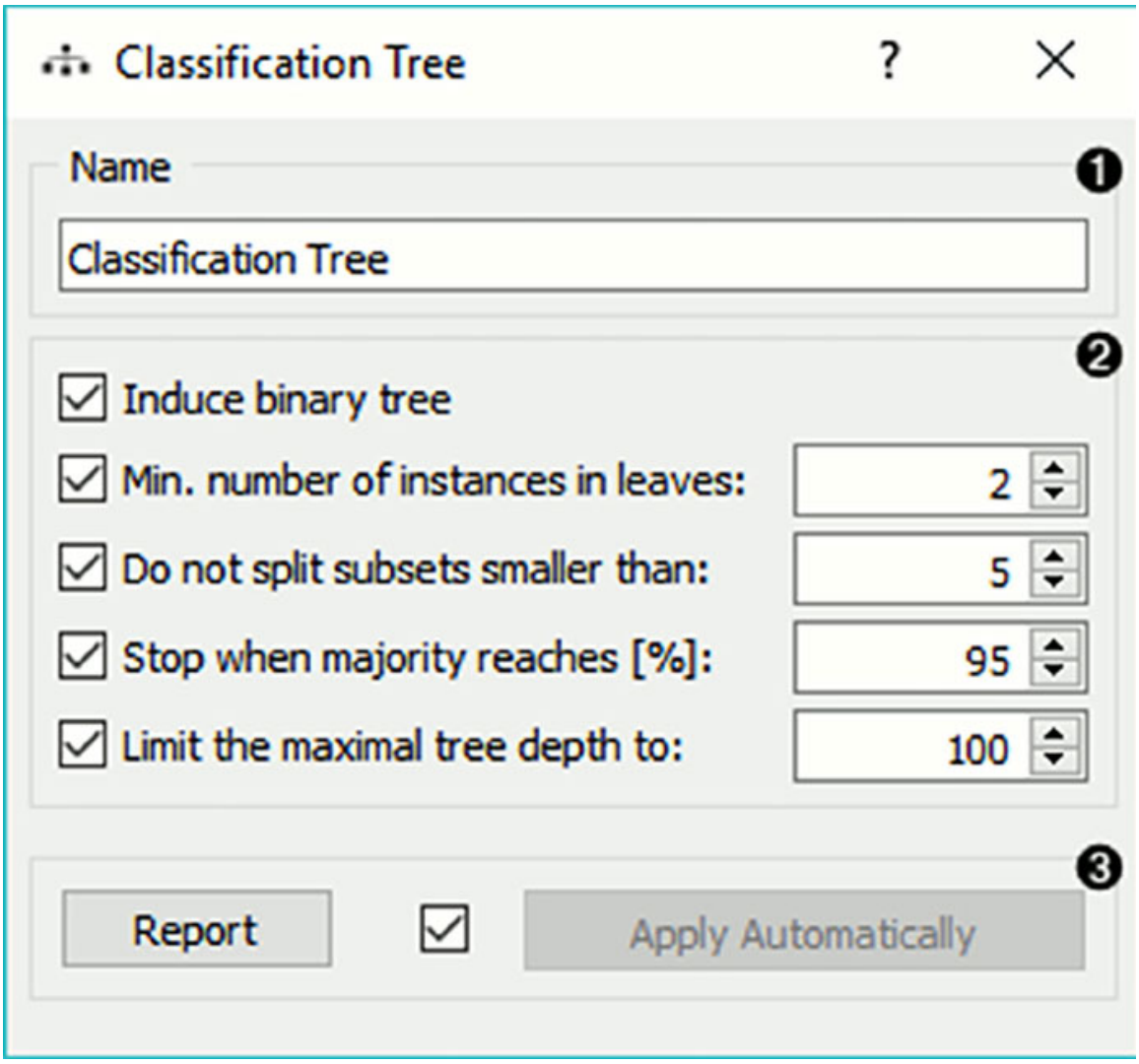


Figure 6.7: *Orange Classification Tree*

- Specifics about the input data.
- Display options include the following:
    - Zoom in and zoom out.
    - Select the width of the tree. When you move your cursor on the screen over the nodes, information bubbles emerge to provide further information.
    - Select the depth of your tree's roots from the drop-down menu.
    - Decide and select the width of the edge. The edges of the tree graph are constructed in accordance with the width of the edges that have been specified.

        - In this case, all the edges will have the same width if the fixed option is chosen.
        - Each node's edge width will be inversely proportional to the number of occurrences in the training set when relative to root is chosen. When going toward the bottom of the tree with this setting, the edge will become thinner and narrower as it gets closer to the ground.
    - The option "Relative to parent" will cause the width of edges to correspond to the proportion of instances in a node in relation to the instances in its parent node.Select the target class, which you may change depending on the data's classes, and then click on **`OK`**.
    - Click on **`Save picture`** to save the classification tree graph to your computer in the form of an SVG or PNG file.

- Analyse your results and produce a report. The analysis screen will have features and options, as shown in the following figure:

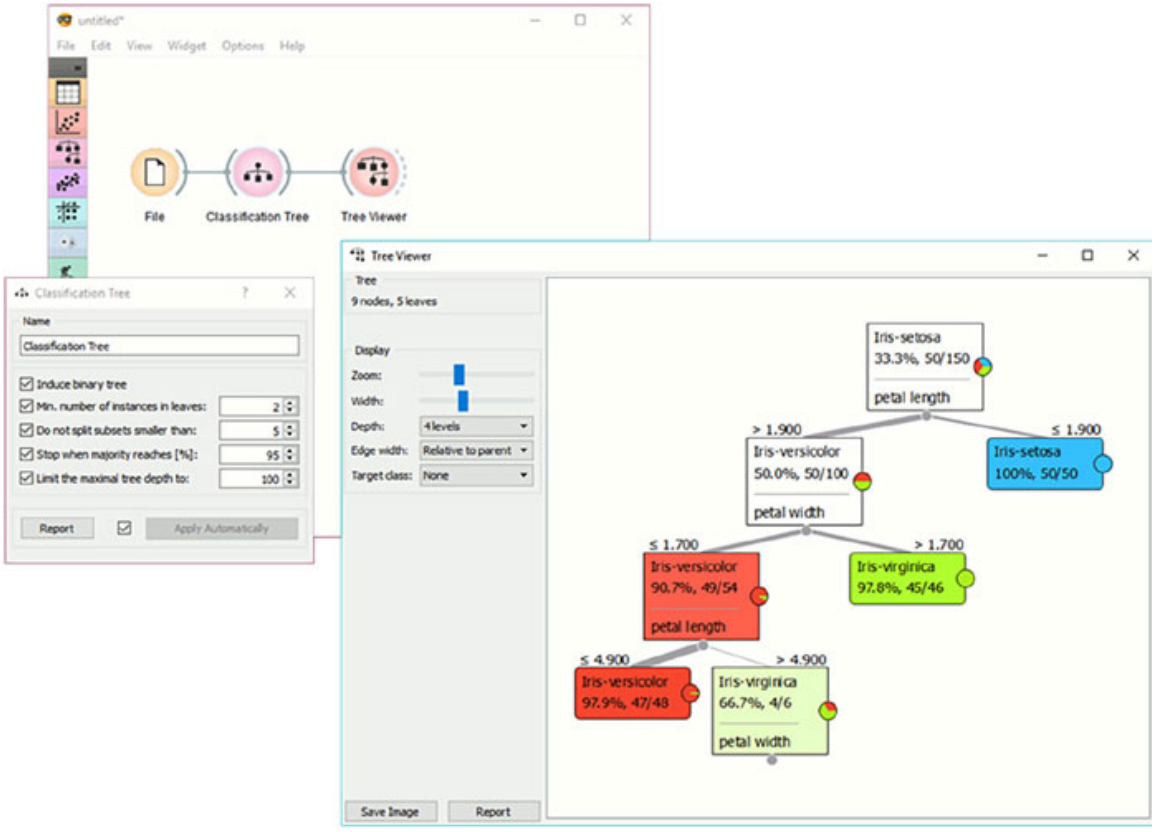


Figure 6.8: *Orange Classification Tree Viewer*

Conclusion

The use of supervised learning may be seen in practically every area or sector, including healthcare. Here's a real-life application: a hospital's special unit tests newly admitted patients suspected of being infected with COVID-19 for 15 distinct factors (such as blood oxygen levels, temperature, and age). It is necessary to make a decision on whether to admit a new patient to an **intensive care unit** (**ICU**). A high demand for ICU beds, combined with a scarcity of available beds, means that patients who need long-term care are given greater priority. In this case, the challenge is to identify high-risk individuals and distinguish them from low-risk ones.

We looked at five different classification techniques, but there are several alternative approaches as well, such as Bayesian networks, for example, logistic regression, neural networks, and stochastic gradient descent. This high number of approaches demonstrates the significance of classification and the broad range of applications for it. It continues to be a hotbed of scientific activity.

Points to remember

- Classification is the process of dividing a given collection of data into groups having similar characteristics. It's a helpful method for figuring out whether an instance fits in a certain category.
- Multi-label classification allows each input to have several output class labels, unlike multi-class classification.
- Ensemble learning can enhance prediction accuracy. Random forest classifiers are ensembles of decision trees that learn from each other to increase model accuracy.
- We covered only a handful of algorithms; there are many more, like logistic regression, neural networks, etc., that are also commonly used in industrial use cases.
- The ideal approach is to explore multiple algorithms and choose one based on evaluation metrics. Stanford University's free learning resources explain it in detail (just Google it).

Multiple choice questions

1. What is conditional probability?

a. It measures the possibility of a single occurrence, provided that a single event has already occurred.
b. It is the likelihood of an event occurring prior to the collection of additional data.
c. It is the concept that quantifies the chance of two occurrences happening concurrently.
d. None of the above

2. KNN classifier is a _____ technique.

a. Parametric
b. Linear
c. Non-parametric
d. All of the above

3. Which of the following statements is/are related to Bayes theorem?

a. It is helpful in determining the probability of a hypothesis with prior knowledge.
b. P(A|B) = [P(B|A) * P(A)] / P(B)
c. It enables you to adjust the estimated probabilities of an event by integrating additional information.
d. All of the above

4. A hyperplane is a boundary that represents the border of the best decision-making boundary.

a. True
b. False

Answers

Question Number	Answer
1	A
2	C
3	D
4	A

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 7

Clustering and Association

Clustering, in the simplest words, is the art of grouping together a set of objects with similar features. And when the different groups have common characteristics, they become associated with one another.

Two of the most commonly used data mining techniques are clustering and association rule mining, which are very useful in marketing, merchandise, and campaign operations. In big commercial databases, the association rule is a valuable tool for discovering intriguing relationships between things that are not obvious to the naked eye. When you imagine your data as being in a huge dimensional space, you want to locate the areas of the space where the data is most dense. This is where cluster analysis has its roots. Clusters are formed when two or more regions are distinct or nearly so.

Structure

In this chapter, we will cover the following topics:

- Basics of Clustering and Association
- Clustering/Association process flow (Identify problem statement, check data availability, choose right pre-processing tool, and define end goal)
- EDA and Evaluation Metric for Clustering
- Clustering and Association using Orange
- Summary with tables, cheat sheets and handouts for easy reference

Objectives

You'll learn how clustering and association are some of the most widely used techniques to discover unknown relationships in data. For an individual who has minimal or negligible understanding of what the data is all about, clustering and association can serve as a starting point to understand the data and how it can be used. With clustering, you can identify comparable patterns and correlations among data points, just like those observed in product recommendations based on purchase history on e-commerce websites.

Get started with Clustering and Association

The objective of clustering is to segregate similar observations in the data into clusters, such that every datum in the cluster is more similar to other datums in the same cluster than it is to the datums in other clusters. It is like grouping or washing all white clothes together. Here, the cluster is based on the color white.

Another use case would be in case of companies that provide guided tours; they may want to group their customers based on their activities and preferences, such as where they want to travel, whether they prefer adventure or luxury tours, and what locations they want to see. The corporation can use this information to create appealing trip packages and target specific sectors of their customer base.

When it comes to machine learning, clustering is an unsupervised machine learning techniques that is meant for identifying and grouping similar data points in larger groups without much focus on the outcome. In order to make data more understandable, clustering (also known as cluster analysis) is frequently employed. Even though it's a common method, clustering isn't a one-size-fits-all concept, as different algorithms employ cluster analysis in different ways.

When you have a large, unstructured data set, clustering can help you reduce a lot of manual work of data cleanup, for example, clustering documents based on topic and file name. Clustering can quickly organise big datasets into something more accessible without any direction. If you don't want to spend a lot of time analysing your data, clustering is the way to go.

Even if you begin with a dataset that is well-structured and well-labelled, there is a possibility that there will still be a lack of detail and differentiation. For data preparation, clustering is a useful first step because it helps shed light on the most important aspects of your dataset. As an example, you might realise that what you believed were two main subgroups are actually four or understand what categories you weren't aware of their own classes. This has been clearly visualized in the following image:

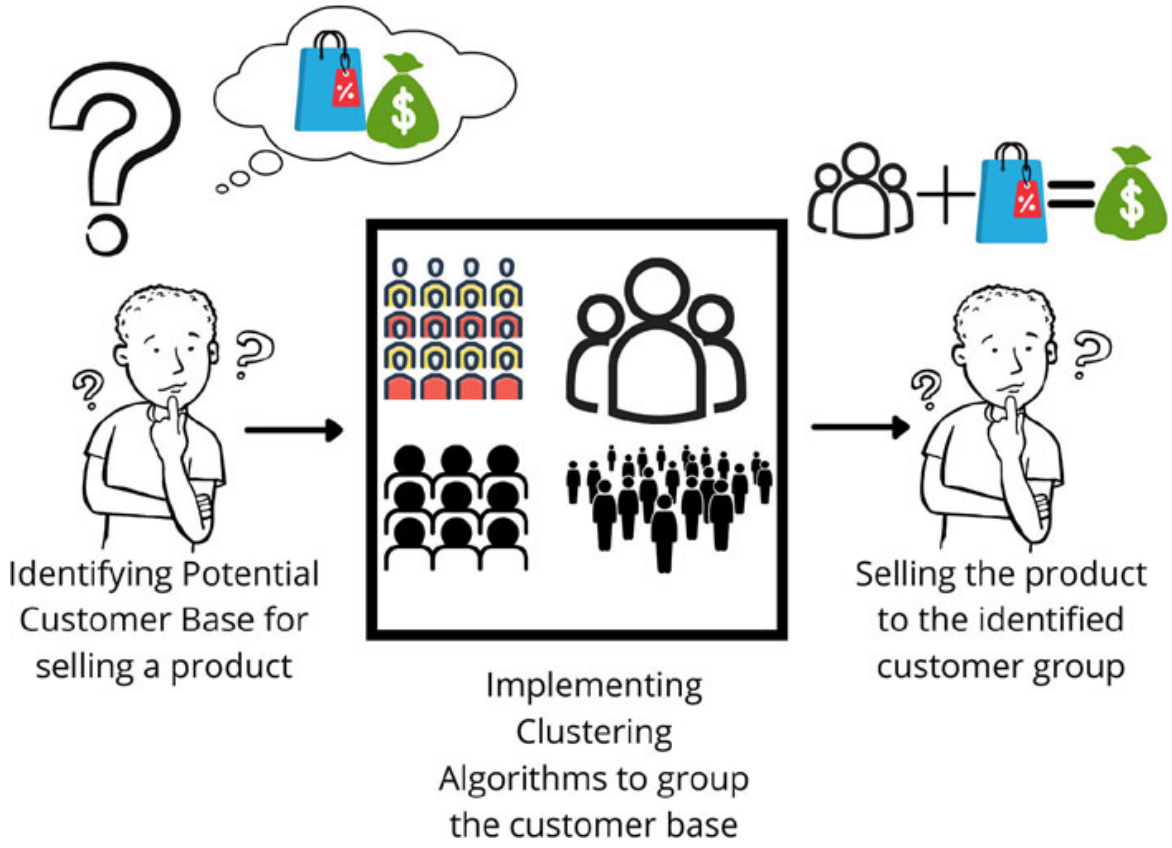


Figure 7.1: *Clustering Application in Real Use Case*

Hard methods and soft methods are the two broad categories into which clustering techniques can be divided:

1. **Hard methods**: Each data point or observation is assigned to a single cluster in the hard clustering method. There are ways to structure clusters in a dataset so that each dataset is placed in only one of the total number of specified clusters, and this is doable for data researchers. In order to properly organise data, a simple classification of datasets is necessary. For example, a clustering algorithm groups data points together based on how closely they resemble each other. Datapoints belonging to different clusters, on the other hand, share no additional similarities.
2. **Soft methods**: As a result of the soft clustering strategy, any data point might be a member of multiple clusters at once. This model comprises a set of membership coefficients that reflect how likely it is that an individual will be included in any given cluster. Soft clustering differs from hard clustering in that it does not require a data point to be part of more than one cluster at a time.

A data point can be included in multiple clusters at the same time when using soft clustering. In other words, soft clustering is characterised by a fuzzy classification of datasets. In machine learning, the Fuzzy Clustering Approach is a well-known approach for unsupervised data processing into soft clusters.

We'll discuss the details of these classifications later; for now, let's define a few crucial terms. Machine learning has several terms of this kind. You may use these definitions regardless of whether you are working on a machine learning project or are just interested in knowing what is happening in this field of data science.

- **Model**: It is the mathematical depiction of a real-world process; a predictive model predicts future outcomes using data from the past.
- **Feature**: Predictor variable, or independent variable, is another term for feature. A feature is something that can be observed and measured by a model. Additionally, features can be engineered in various ways, including merging or adding additional information.
- **Training**: It is the process of constructing a model from the acquired data.
- **Testing**: It is the process of assessing a model based on the results of a test. The performance of various models is measured and benchmarked using a dataset that is independent from the training set but has the same structure.
- **Algorithm**: It is a collection of rules used to solve an issue or make a computation.

Density- based clustering

Based on density, points are clustered together and separated from one other by spaces that are sparse or empty. Density refers to the number of data points in a region. The term "outlier" refers to points that are not associated with a cluster. When using time to locate clusters of points, the time of the points can also be used. This clustering technique is based on the concentration of instances inside a specified region. As long as dense regions can be linked, arbitrary-shaped distributions are possible. As long as dense areas can be connected, it allows for arbitrary-shaped distributions; for example, consider a dataset of customer information with attributes like age and income. For the sake of simplicity, let us consider only age and income. The density will be high in regions where there are more customers with similar age and income. This situation of data with changing density and high dimensions is a problem for these methods. Outliers aren't assigned to clusters because the algorithms aren't set up to achieve that.

To assess the density of a dataset, this algorithm will often begin by defining a neighbourhood around each data point. High-density areas are recognised as clusters, and their limits are established relative to the neighbouring regions. These approaches are referred to be unsupervised since they do not need any training on what constitutes a cluster.

A useful use case can be in the study of a particular pest-borne disease. There are some families in your study region that are plagued with pests, while others are not, and you have a point dataset that represents each of these households. Infested houses can be pinpointed using the Density-based Clustering technique, which helps identify the greatest clusters of pest-infested homes.

Following natural disasters or terrorist attacks, geo-located tweets can be clustered so that rescue and evacuation needs can be informed about the size and position of the clusters found.

Density- Based Spatial Clustering of Applications with Noise (DBSCAN)

Clusters are characterised in the DBSCAN framework as regions of high density separated from areas of low density (Density-Based Clustering). In contrast to an approach like K-Means, which optimises the within-cluster sum-of-squares and hence, works best for convex forms, DBSCAN's generic perspective allows it to detect clusters of any shape. In addition, DBSCAN automatically calculates the number of clusters. DBSCAN requires only a modest amount of domain expertise to establish the parameters of the algorithm. It does not require the user to specify the number of clusters in advance, which can make it more flexible and applicable to a wider range of datasets. It only requires the user to specify two key parameters: the radius of the neighbourhood (Eps) and the minimum number of data points (MinPts) required to form a cluster. By adjusting these parameters, the user can control the shape and size of the clusters that are formed.

Eps: If the distance between two points is less than or equal to 'eps', they are regarded to be in the same neighbourhood. A considerable portion of the data will be deemed outliers if the eps value is set too low, making it difficult to analyse. The eps value can be calculated using a k-distance graph. A minimum of one data point must be located inside an eps radius of another. The k-distance graph is a graphical representation of the density of the data. Here, each data point is associated with a k-distance, which is defined as the distance to the k-th nearest neighbour.

MinPts: It must be set to a higher value when working with a huge dataset. In general, the minimal MinPts may be calculated by multiplying the number of dimensions D in the data by D+1. MinPts must have a minimum value of 3.

This technique uses three distinct types of data pieces:

- If a point has more than MinPts points within the eps, it is termed as a core point.
- A boundary point is a location inside eps that contains less than MinPts but is beside a core point.
- A noise point, also known as outlier, is a point that is not a core or boundary point.

Figure 7.2 visualises the mentioned core point, boundary and outliers.

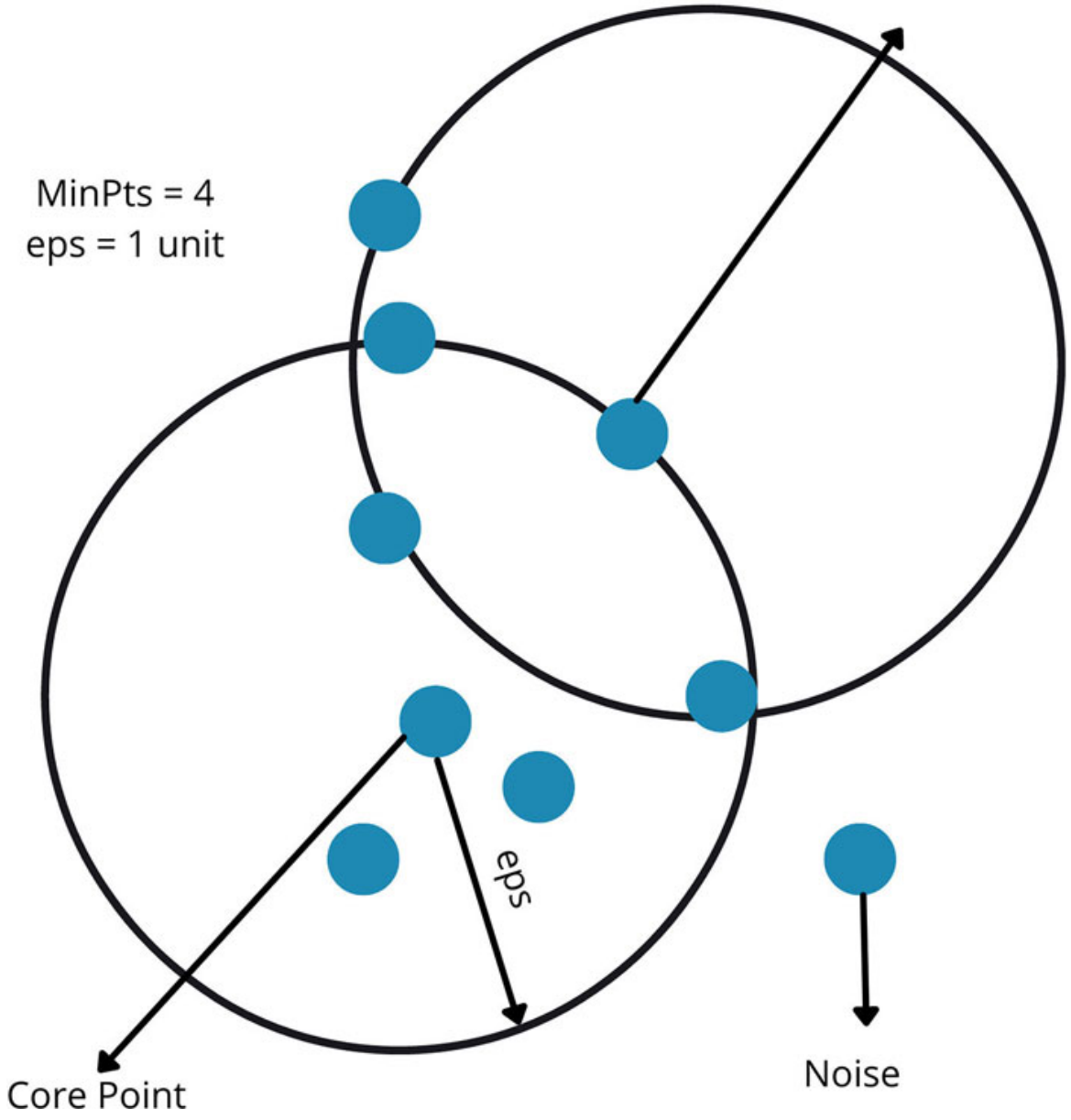


Figure 7.2: *DBSCAN Clustering Principle*

Ordering Points to Identify Clustering Structure

OPTICS is also a spatial data grouping algorithm based on density. Its fundamental technique is similar to that of DBSCAN; that said, DBSCAN does not perform well with clusters of varying densities, but OPTICS does.

It adds two new terms to the DBSCAN clustering vocabulary. It goes this way: In order to be classified as a “core point,” a point needs to have a radius less than the Core Distance. In this case, if the given point isn’t a Core point, it doesn’t know how far away it is from the given point. A data point called q is used to figure out

what this thing is. (Let). If you want to know how far two points are from each other, you add up the Core Distance and the Euclidean Distance between them. Take note that if q is not a Core point, the Reachability Distance is not defined. Euclidean distance is a measure of the straight-line distance between two points in a multidimensional space, calculated as the square root of the sum of the squared differences between the coordinates of the points. The core distance is the minimum distance between a data point and its neighbours within a given radius. It is used to determine the density of a region and to determine whether a data point is a core point.

This clustering technique is distinct from others in that it does not deliberately partition the data into groups. Rather, it creates a visualisation of reachability distances and then clusters the data using this visualisation.

Hierarchical density- Based spatial clustering applications with Noise

Density-Based Spatial Clustering of Applications with Noise or DBSCAN is performed with varying epsilon values, and the result is integrated to discover the clustering that provides the best stability over epsilon. This enables HDBSCAN to detect clusters with varied densities (unlike DBSCAN) and to be more parameter selection-resistant. In real life, this means HDBSCAN automatically creates a good clustering with little or no tweaking of parameters. The main parameter, the least cluster size, is logical and easy to choose. HDBSCAN is an excellent tool for exploratory data analysis; it is a fast and dependable method that consistently returns significant clusters (if there are any).

Hierarchical clustering

Hierarchical clustering is a method for clustering unlabelled datasets into groups. It is also known as hierarchical cluster analysis or HCA. This approach creates a tree-like hierarchy of clusters, which is called a dendrogram. Although the results of K-means clustering (which will be explained later) and hierarchical clustering may appear similar, they are fundamentally different. Unlike K-means clustering, hierarchical clustering does not require the user to specify the number of clusters beforehand. There are two types of hierarchical clustering:

- **Agglomerative**: Agglomerative clustering is a bottom-up strategy in which the algorithm begins by clustering all data points and merging them until only one cluster remains.

- **Divisive**: Divisive algorithms are the inverse of agglomerative algorithms in that they take a top-down approach.

Fuzzy clustering

Fuzzy clustering permits ambiguity and overlap while classifying data points into groups. Instead of categorising each data point as a member of one of two groups, as is done in other clustering algorithms, fuzzy clustering gives a degree of membership to each cluster rather than assigning a hard, binary membership to a single cluster. Clustering organises data points into groups based on their resemblance to one another and searches for patterns or similarities between items in a collection; objects in clusters should be as similar to one another as feasible, while being as distinct to other groupings as possible. It is considerably easier to build fuzzy borders.

Clustering that is “soft” or “fuzzy” allows for the possibility of data points belonging to many groups. The method uses least-squares solutions to find the optimal position for each data point, which can be located in the probability space between two or more clusters. This concept is similar to the behaviour of electrons in atomic orbitals, where electrons are not fixed in one location but have a probability of being in a particular orbital shell. If you consider orbital shells to be “clusters” and electrons to be “data points” (each data point is assigned a probability of being placed in a certain cluster), you have a rudimentary understanding of fuzzy clustering’s fundamentals.

Partitioning clustering

Iterative relocation methods are the most often used clustering algorithms. Iteratively shifting data across clusters until a (locally) optimal partition is achieved is the goal of these techniques. Several partitioning algorithms are available, including K-Means, PAM (k-Medoid), and the **Clustering Large Applications** (**CLARA**) algorithm, as depicted in *Figure 7.3*:

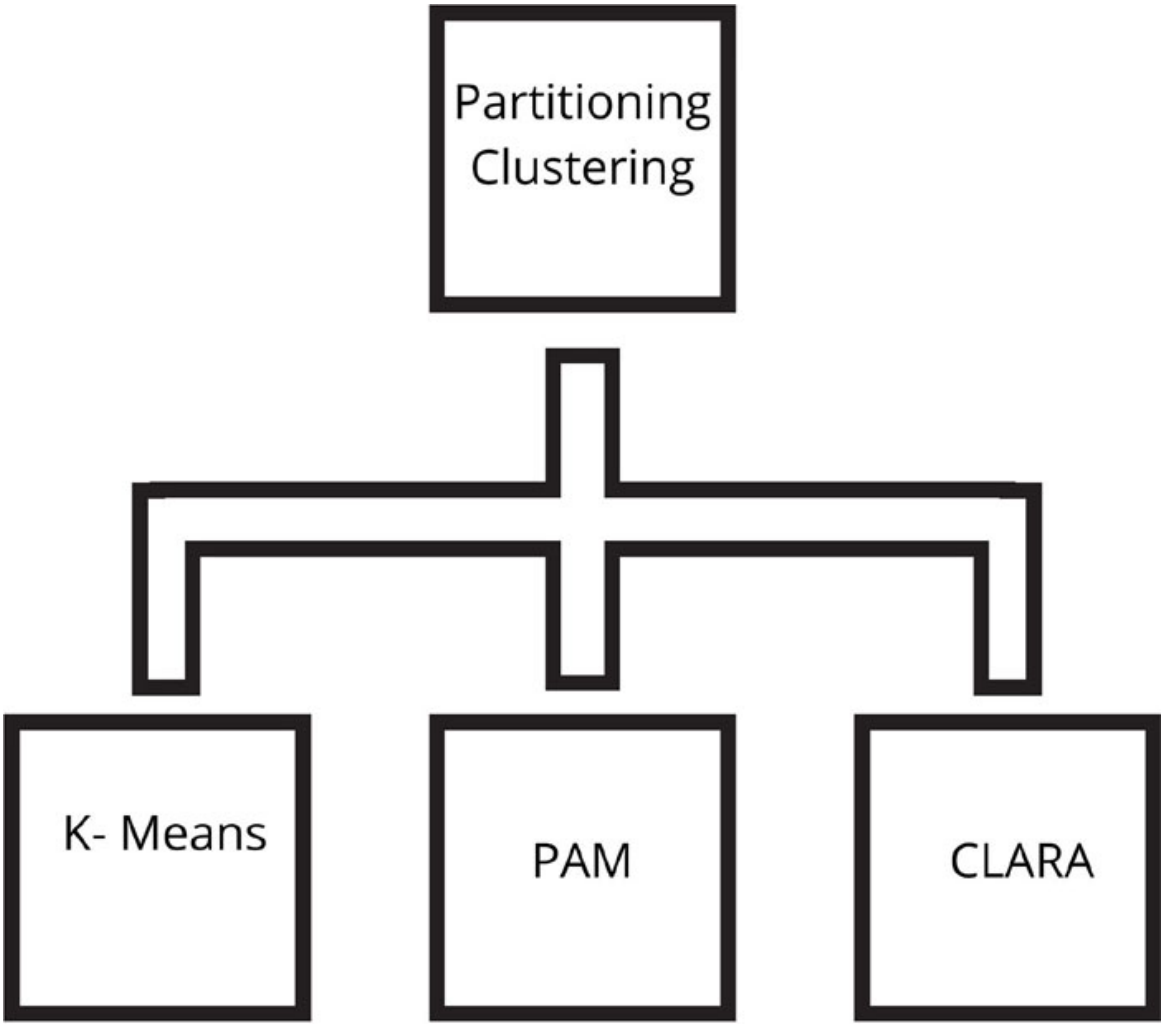


Figure 7.3: *Various Partitioning Clustering Methods*

K-Means algorithm (A Centroids-Based Technique): It is a frequently used technique for dividing a data set into k groups (i.e., k clusters), where k is the number of clusters. It clusters things into several groups (i.e., clusters), with the goal of keeping objects within the same cluster as similar as feasible (i.e., high intra-class similarity), while keeping objects from different clusters as distinct as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its centre (i.e., centroid), which is the mean of the all data points in the cluster. The essential notion underpinning k-means clustering is to construct clusters in a manner that the total intra-cluster variance is minimised (also known as total within-cluster variance).

K-Medoids algorithm (Partitioning Around the Medoid): A medoid is a point in a cluster that is similar to many other points in the cluster. Each cluster in k-medoids clustering is represented by one of the cluster's data points. Cluster

medoids are the names given to these sites. The word medoid refers to a data point inside a cluster that is most similar to its neighbours. A medoid is selected as the representative of the cluster because it is the data point that is closest to all other points in the cluster. It is a node that corresponds to the cluster's core node. The core nodes are the medoids selected to represent each of the k clusters. Remember that the centre of a cluster is calculated using k-means clustering as the average value of all the data points inside the cluster. K-medoid is a more robust approach for clustering data than k-means. This implies that, in contrast to k-means, the method is less susceptible to noise and outliers since it employs medoids rather than means as cluster centres (used in k-means).

Clustering Large Applications (**CLARA**) is a modification of the k-medoids (PAM) approach for dealing with data containing a high number of objects (over few thousand occurrences) in order to reduce computational effort and RAM storage needs. This is performed by the application of sampling. Rather than generating medoids for the complete data set, CLARA evaluates a fixed-size subset of the data (sampsize) and uses the PAM algorithm to build an optimal set of medoids for the subset. The quality of the generated medoids is determined by the average dissimilarity between each item in the total data set and its cluster's medoid, defined as the cost function. CLARA iteratively repeats the sampling and clustering operations in order to decrease sampling bias. The ultimate clustering results match the collection of medoids at the lowest possible cost.

Grid-based clustering

Grid-based clustering approaches use a data structure that is divided into a grid of cells, with each cell representing a cluster of data points. The data structure has several resolutions, meaning that the size of the cells can be adjusted to change the granularity of the clustering. The basic idea behind grid-based clustering is to divide the data space into a grid of cells and then assign each data point to the cell that contains it. After assigning data points to cells, clusters can be formed by merging cells with high data point densities. This may be done by iteratively improving the grid resolution and combining cells with comparable data distributions. The objective is to identify an ideal grid resolution that produces meaningful clusters that capture the data's structure. The method's advantage is its speed, which is generally independent of the quantity of data objects and is still dependent on the numerous cells in each dimension of the quantized space. STING, Wave Cluster, and CLIQUE are all examples of the grid-based approach.

Association

Association Rules are used to discover intriguing relationships and patterns buried in large datasets using rule-based machine learning. Using these rules, you may see how often a particular set of things appears in your datasets or transactions.

They are most suited for analysing data from relational databases and transactional databases to identify trends. Association rule mining or “mining associations” is a term used to describe the practise of employing association rules.

In association rules mining, new and fascinating insights between various items in a set, regular patterns in transactional data, or any other type of relational database are discovered through the analysis of association rules. In addition, they are used for Market Basket Analysis, Assortment Decisions, Product-bundle Pricing, Cross-Selling, and many other things. These are just a few of the many things they can be used for. This can be seen as a more intricate variation of the “what if” scenario, in which a hypothetical situation is imagined and its potential outcomes are explored.. The Apriori method generates association rules based on frequent item sets and is intended to be used with databases that include transactions. We will look at how association rules mining works in detail later.

Process flow of clustering

A clustering algorithm is a way to look at groups of data that have some degree of similarity. To accomplish successful clustering, the method determines the distance between each point and the cluster’s centroid. Clustering can be used to find the inherent grouping of a collection of unlabelled data in order to do this.

- **K-means clustering method**: If k is chosen, the K-means algorithm may be implemented as follows:
    - Partitioning is the process of dividing items into k non-empty subsets. It is necessary to locate and identify the cluster centroids (mean points) of the current division.
    - The initial step is to initialize the centroids to data points.
    - Calculate the distances between each point and assign them to the closest centroid.
    - Repeat the preceding two steps until the centroids no longer change or until a maximum number of iterations is reached.

- Some of the advantages of K-means Clustering are as listed here:
    - It has a time complexity of O (nkt), which means it is relatively efficient. In each iteration of K-Means, the time complexity for computing the distance between each data point and each centroid is O(nk), and the time complexity for updating the centroids is O(k). Thus, the overall time complexity for each iteration is O(nk + k) = O(nk).
    - It converges at the local optimum.
- **Hierarchical clustering**: Instead of the unstructured set of clusters produced by non-hierarchical clustering, hierarchical clustering provides a more informative structure.

  The fundamental algorithm is as follows once the number of clusters is predefined:

    - Calculate the closeness (similarity) matrix between the two data points.
    - Start with each data point as a single cluster.
    - Two closest clusters are merged into a single larger cluster.
    - Repeat the preceding step until the predefined number of clusters are formed.

  Agglomerative hierarchical clustering is a type of hierarchical clustering.
- **Partitional clustering**: Partitional clustering, in contrast to hierarchical clustering approaches, tries to create consecutive groups through the use of iterative procedures. The partitioning of a collection of data points into k-clusters is accomplished using iterative procedures in partitional clustering. These techniques organise n data points into k-clusters based on similarities. The pre-set criteria function J allocates the data to the kth number set in accordance with the outcomes of the kth number set's maximisation and minimization computations.
- **Association**: Consider the following example of how an association rule is written. This structure consists of two components: an antecedent and a consequent. Each of these contains a list of items. It is critical to keep in mind that the sense here is one of co-occurrence rather than causation. The term "itemset" refers to a collection of all the things in the antecedent and consequent of a given rule. You can get an idea of how association works

from *Figure 7.5*, where a customer who buys bread and egg can possibly buy milk as well.

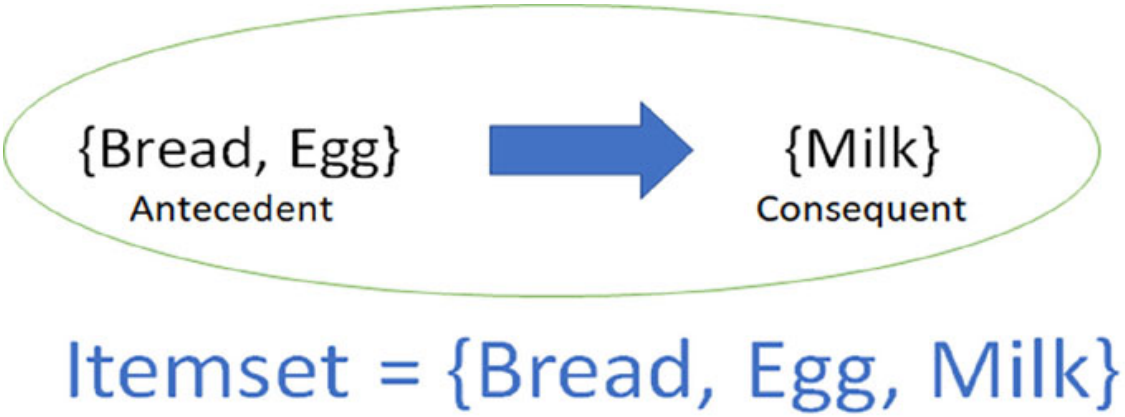


Figure 7.4: *Example of Association*

Numerous indicators are used to ascertain the strength of the association between these two variables.

- **Support**: This metric provides an indication of how frequently a particular itemset appears in the transactions. Consider the following itemsets: itemset1 = "jam" and itemset2 = "soap." Assume that there will be significantly more transactions involving jam than transactions using soap. Consequently, as you correctly predicted, itemset1 will typically have a larger level of support than itemset2. Consider the following itemsets: itemset1 = "jam, butter," and itemset2 = "jam, shampoo." Frequently, transactions will have both jam and butter in the basket, but what about jam and shampoo? That's not the case. As a result, in this instance, itemset1 will typically receive more support than itemset2.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- **Confidence**: In the case of a cart that already has the antecedents, this measure determines the likelihood that the consequent will appear on the cart.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

- **Lift**: When determining the conditional probability of occurrence of 'Y' given 'X,' lift controls are used to support (frequency). When referring to this metric, "lift" is used to describe it in the most literal

sense. Think of it as the *boost* in self-assurance we get from having 'Y' on the cart, thanks to 'X.' When you are aware that a particular item is present in a cart, the probability of the cart containing that item increases compared to when you are not aware of its presence..

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

EDA and evaluation metric for clustering

Exploratory data analysis (**EDA**) is a technique used by data scientists to study and investigate large data sets to summarise their essential properties. Data visualisation techniques are frequently employed in this process. The technique assists data scientists in determining the most effective way to alter data sources to obtain the answers they require, making it easier for them to detect patterns, identify anomalies, test hypotheses, and verify assumptions. EDA is primarily used to investigate what data can disclose beyond the formal modelling or hypothesis testing work and to gain a deeper knowledge of the variables in the data collection and the relationships between those variables. Additionally, it can help you decide whether the statistical approaches for data analysis that you are contemplating are suitable for your circumstance.

Exploratory Data Analysis is a method of conducting data analysis that uses a range of approaches in order to develop a sense of intuition regarding the data.

- Make sure you're not losing your mind (in order to ensure that the insights we are making are derived from the correct dataset).
- Determine where information is lacking.
- Examine the data to see if there are any outliers.
- Make a summary of the information.

The following are examples of statistical functions and approaches that can be performed with EDA tools:

- Techniques like as clustering and dimension reduction can be used to construct graphical representations of high-dimensional data including numerous variables.
- Each field in the raw dataset is shown as a single-variable visualisation, along with summary statistics.
- Using bivariate visualisations and summary statistics, you can determine the link between each variable in the dataset and the target variable that you're

interested in learning more about. Multivariate visualisations are used to map and comprehend the interactions between distinct fields in the data, and between different fields in the data.

You might have a question in mind "*How would EDA assist me in creating effective and appropriate clusters?*". Well, EDA will help you in the following ways. You can evaluate whether it makes sense to do a clustering analysis. It will help figure out the potential number of clusters. Sometimes, clustering will be used as an EDA method for supervised learning algorithms. For more details on EDA, refer to the previous chapters as the methodology remains the same.

You can evaluate the performance of the clustering algorithm using silhouette score. The silhouette score is a measure of the similarity of a data point to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating that a data point is better matched to its own cluster than to other clusters. In simple terms, the silhouette score provides a measure of how well a data point fits in its assigned cluster and how distinct and separate the clusters are.

Clustering using Orange

Interactive K- Means Clustering Widget: The data is clustered using the k-Means technique. Then, a new collection of data is created with the cluster index as a class property. The original class attribute is redirected to the meta-attributes, if any. Additionally, the widget displays results for various k. You can see the widget in *Figure 7.6*:

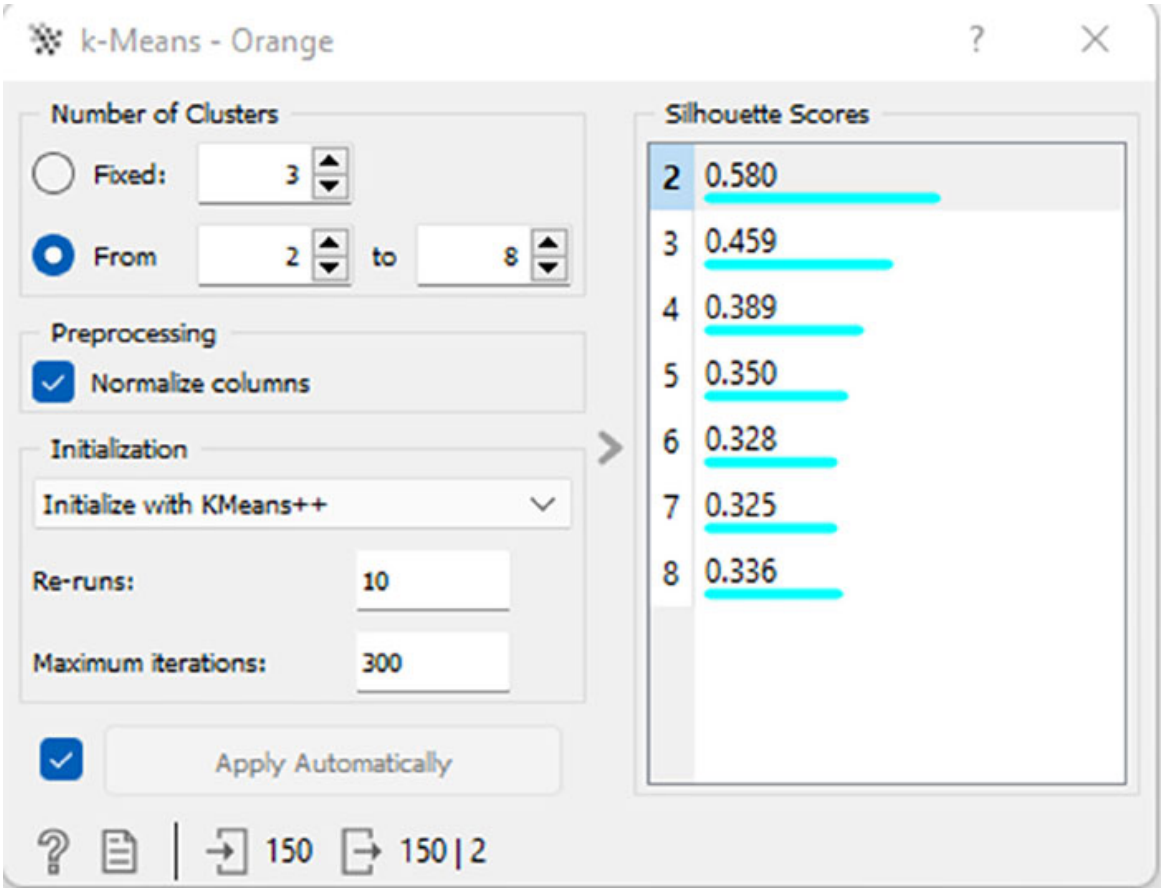


Figure 7.5: *Clustering Widget in Orange*

Step 1: Select the cluster count

- **Fixed**: It clusters data into a set number of clusters.
- **Optimized**: It now shows the clustering scores for the chosen cluster range.
- **Scoring**: Silhouette compares the average distance of each data point in a cluster to the average distance of data points in other clusters, in order to measure how well-separated the clusters are..
- **Inter-cluster distance**: It measures the distances between clusters, usually between the centres.
- **Distance to centroids**: Centroids distance captures the distance between clusters' arithmetic means.

Step 2: Select the initialization method

- **k-Means**: The first centre is picked at random, and subsequent places are chosen with a probability proportional to the square of the distance from the initial center.

- **Randomization**: Clusters are first allocated randomly and then modified with subsequent iterations.
- "Re-runs" (number of times algorithm is executed) and "maximal iterations" (number of iterations each time the algorithm runs) can be manually configured.

Step 3: A new data set with appended cluster information will be outputted by the widget. The column for the appended cluster information can be named and the method for appending it (as class, feature, or meta attribute) can be selected.

Step 4: If "apply automatically" is checked, changes will be automatically committed by the widget. Otherwise, click "apply".

Step 5: Get the results in the form of a report.

Step 6: Check and verify the scores of clustering results for various values of k.

For more details on using Orange, refer to the Orange official documentation or website as all these explanations are based on the information available at the Orange documents store.

Hierarchical Clustering Widget: Hierarchical clustering is performed using a matrix of distances, and the dendrogram is displayed. Refer to *Figure 7.7*:

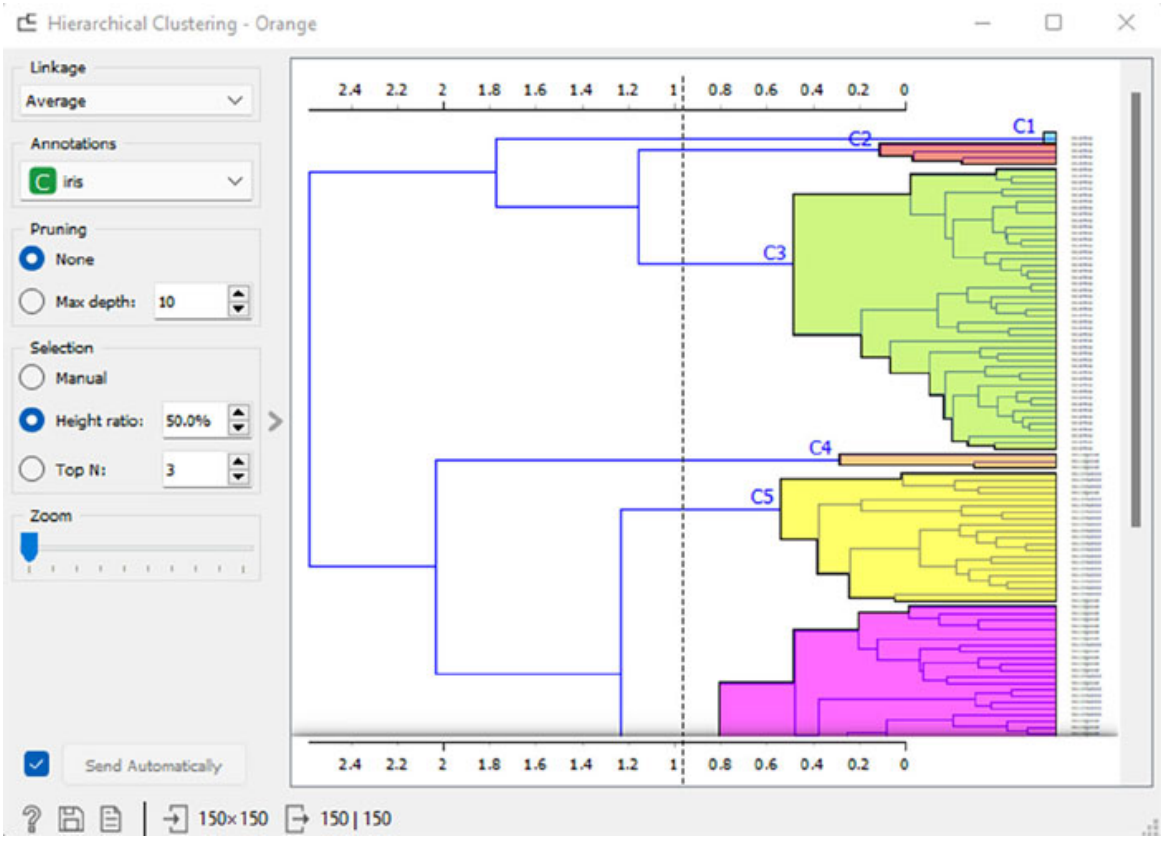


Figure 7.6: *Dendrogram in Orange*

Step 1: Clusters can be measured in four different ways using this widget:

- A single linkage is used to calculate the distance between the two clusters.
- The average linkage of two clusters quantifies the distance between their members.
- PGMA is the method used for weighted linking.
- The distance between the most distant members in a cluster is calculated using full linkage.

Step 2: In the Annotation box, you can choose the names of the nodes in the dendrogram.

Step 3: In the Pruning box, you can prune massive dendrograms by setting the dendrogram's maximum depth to the maximum value. Clustering itself is unaffected by this; only the visual presentation is affected.

Step 4: The widget has three ways to choose from:

- **Manual**: A cluster can be selected by clicking inside the dendrogram. Ctrl/Cmd can be used to select several clusters. Each chosen cluster is highlighted in a distinct colour and is treated as a separate cluster in the result.
- **Height ratio**: In the dendrogram, clicking on the bottom or top ruler draws a cut-offline across the graph. To the right of this line, all the options have been checked.
- Top N sets the maximum number of top nodes.

Step 5: Use the Zoom option to zoom in and zoom out.

Step 6: A cluster index can be added if the items being grouped are instances (Append cluster IDs). Attribute, Class, and Meta attributes can all have the same ID. If the data already has a class attribute, it is appended to the meta-attributes with the original class.

Step 7: In either case, the data can be sent either automatically (if "Auto send" is selected) or manually (if "Auto send" is not selected).

Step 8: This button creates a picture that may be saved by right-clicking on it.

Step 9: A report can be generated to view the results. Refer to *Figure 7.8:*

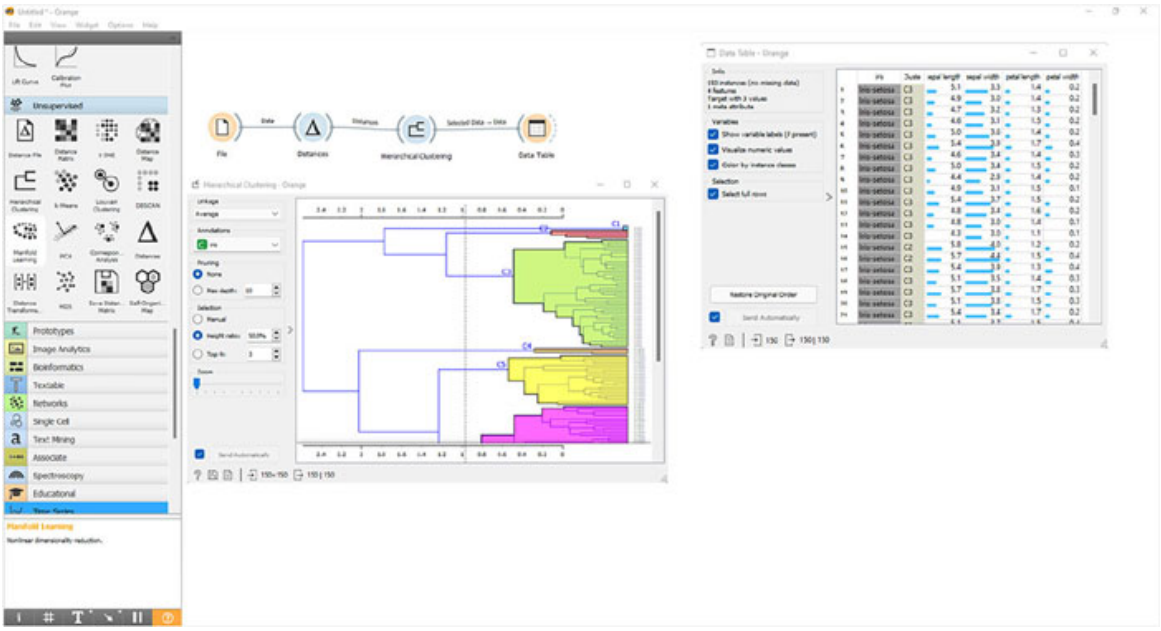


Figure 7.7: *Hierarchical Clustering in Orange*

Clustering cheat sheet

The following image will help you understand the various clustering algorithms and their working principles:

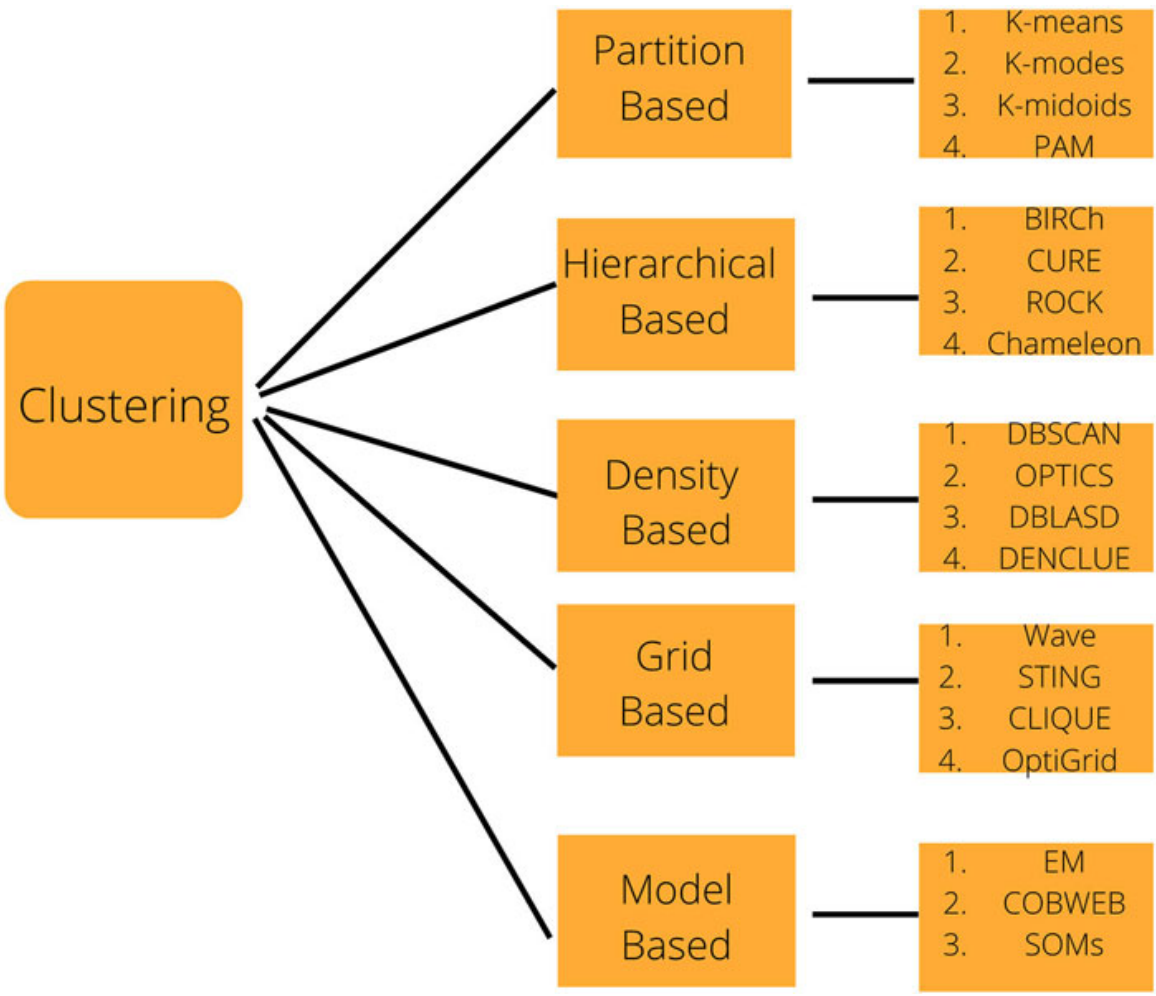


Figure 7.8: *Clustering Cheat Sheet*

Conclusion

Clustering is an unsupervised learning technique that psychologists refer to as sorting and marketers refer to as segmentation. Data is classified into classes with a high degree of intra-class similarity and a low degree of inter-class similarity. In contrast to classification, class labels are determined directly from the data here, along with the number of classes. In a more casual sense, discovering natural groups of items. You learnt about numerous clustering techniques in this chapter.

Clustering is the process of separating a population or set of data points into several groups such that data points belonging to the same group are more comparable to one another than to data points belonging to other groups. In a nutshell, the objective is to classify groups with similar characteristics and assign them to clusters.

In the next chapter, you will look at an interesting machine learning technique to handle time series data.

Points to remember

- Clustering can be subjective to some extent, as the interpretation and evaluation of clustering results depend on the goals and requirements of the specific application. The distance measure determines the similarity between data points, and the choice of distance measure plays a crucial role in defining the clusters. Clustering finds its application in market segmentation, image segmentation and anomaly detection.
- Association rules analysis is a method for determining how objects are related to one another.
- Association may be used to aid decision-making in a broad range of applications, including medical diagnostics, relational and distributed databases.

Multiple choice questions

1. What is the aim of clustering?
    a. It is used to divide data points into clusters.
    b. It is used to sort the data point into distinct categories.
    c. It is used to predict the values of the output data points based on the values of the input data points.
    d. All of the above
2. Clustering requires _____ data.
    a. labelled
    b. unlabelled
    c. only numerical
    d. only categorical
3. Which of the following is an application of clustering?
    a. Medical image analysis to identify relationships
    b. Market segmentation
    c. Social network analysis

d. All of the above

4. Which of the following is an association algorithm?

a. Apriori
b. Clara
c. K-means
d. DBSCAN

Answers

Question Number	Answer
1	A
2	B
3	D
4	A

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 8

Time Series Forecasting

> *Time series forecasting in the simplest terminology is the art of predicting futuristic events by analysing the past trends because the future trends would be like the historical trends.*

Forecasting time series involves making scientific predictions based on time-stamped data from the past. This technique helps draw conclusions and make plans for the future based on the analysis results of past events.

Structure

In this chapter, we will cover the following topics:

- Basics of time series forecasting
- Time series forecasting process flow (identify problem statement, check data availability, choose the right pre-processing tool, and define the end goal)
- EDA & Statistics for Time Series analysis
- Time series forecasting using Orange
- Summary with tables, cheat sheets, and handouts for easy reference

Objectives

You'll learn how time series forecasting is one of the most commonly used techniques to make scientifically backed predictions on a time stamp basis. To put it simply, a time series is a list of events in chronological order. These observations are collected over a fixed interval of time, and the dimension of time adds structure and constraint to the data.

For example, in the field of economy and finance, concepts like exchange rates, interest rates, and financial indices are all time series observations.

When it comes to industrial concepts, factors like electric load, power consumption, and water usage are all time series observations.

In this chapter, we will learn how to predict the future based on past time-stamped observations.

Get started with time series forecasting

Time series forecasting is a method for predicting future events by evaluating past patterns. This strategy analyses previous patterns with the premise that they will remain true in the future. In order to generate predictions about the future, forecasting entails fitting algorithms to data from the past. Effective and timely planning may be achieved using time series forecasting. This method is necessary for solving prediction problems that have a time component.

Guess it might sound interesting to you now!

The applications of time series models are numerous and cover numerous fields, ranging from forecasting sales to weather. Time series models have been found to be among the most effective techniques for predicting, making them useful in situations where there is at least some degree of unpredictability regarding the future. Time series-based predictions are utilized to make a wide range of business decisions.

Here are a few examples:

- Demand projections for the purpose of determining whether to construct an additional power production plant in the next 5 years.
- Making projections about the number of calls that will be placed to a call center in the coming week.
- Estimating the amount of inventory that will be required to meet demand through forecasting, and making accurate projections of supply and demand in order to improve fleet management and other areas of the supply chain.
- Predicting when equipment will break down and what kind of maintenance would be needed in order to reduce downtime and maintain safety standards.

- The practice of predicting frequencies of infection to improve disease prevention and outbreak strategies.

From the anticipation of customer ratings all the way to the projection of product sales, different temporal frames may be included in forecasting, depending on the circumstances and the nature of the information being anticipated. *Figure 8.1* gives a general overview of time series data points:

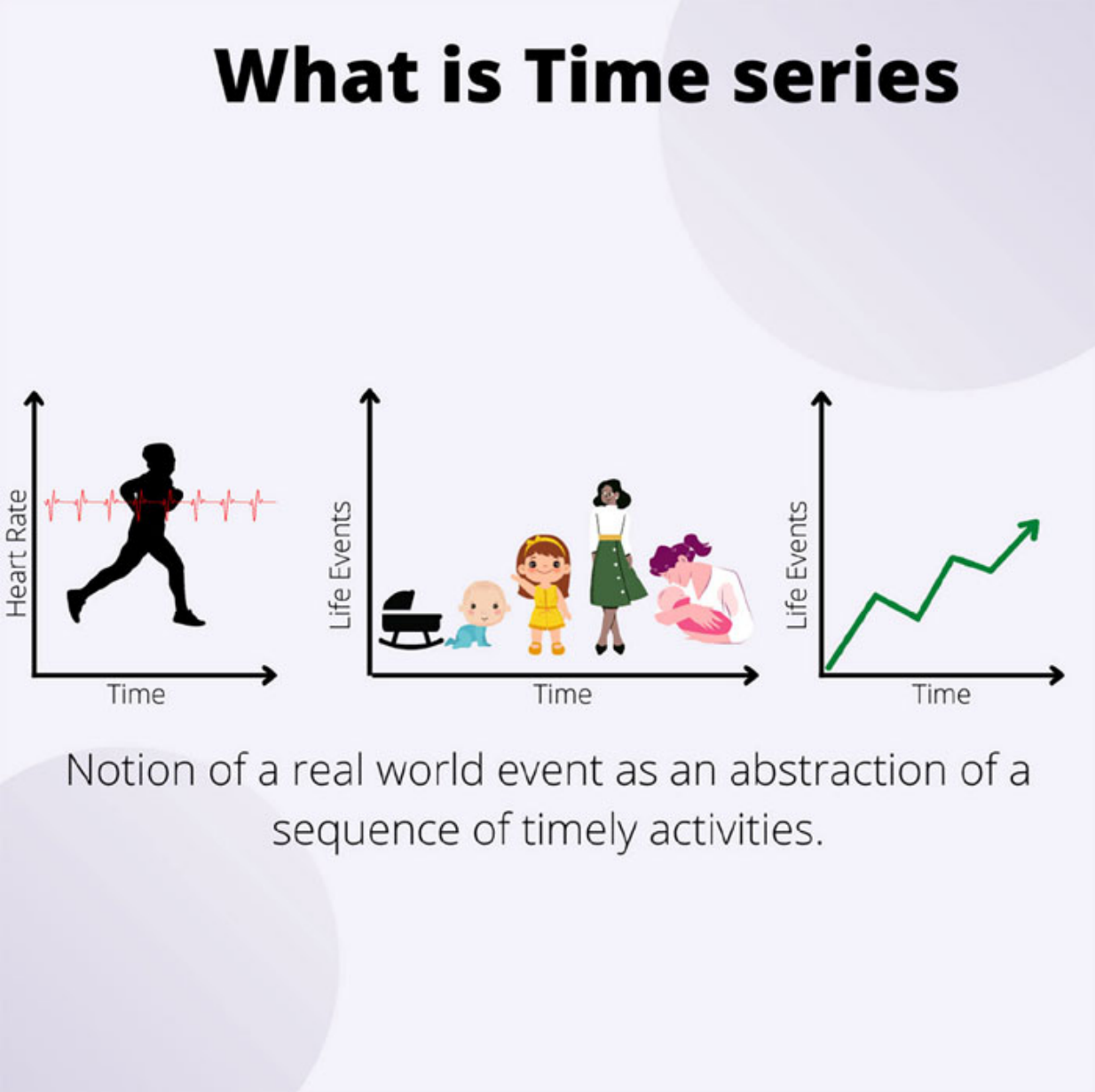


Figure 8.1: *What is Time Series?*

We can perform an in-depth analysis of a time series by first deconstructing it into its component parts. This method is referred to as time series decomposition on our end. If you take a closer look at a time series, you'll notice that it has several different components, including Trend, Seasonality, Cyclic, and Residual. This scan is represented in an equation, as shown in *Figure 8.2*:

Figure 8.2: *Time Series Equation*

We'll get into the specifics of these groupings later; for now, let's establish a few key terminologies. There are a lot of terms like this in machine learning. You may use these definitions whether you're working on a project involving machine learning or just curious about what's going on in this area of data science.

- **Model**: It is the mathematical depiction of a real-world process; a predictive model predicts future outcomes using data from the past.
- **Target**: An example of a target might be a variable in a model that you want to forecast or the output of a model.
- **Feature**: The predictor variable, or independent variable, is another term for a feature. A feature is something that can be observed and measured by a model. Additionally, features can be engineered in various ways, including merging or adding additional information.
- **Training**: It is the process of developing a model using the data that has been collected. Algorithms learn representations of the issue and build models from data put into them.
- **Testing**: It is the process of assessing a model based on the results of a test. The performance of various models is measured and benchmarked using a dataset that is independent of the training set but has the same structure.
- **Algorithm**: It is a collection of rules used to solve an issue or make a computation.

Aspects of time series forecasting

While working on time series forecasting, it is necessary to understand some of the key aspects of the time series equation. Is it stationary? Is the dependent variable associated with itself or autocorrelated? Does it display seasonality?

When analyzing a time series, the first step is to visualize the data trend. It will be helpful to break it down into its constituent parts, such as level, trend, seasonality, and noise. In the context of time series analysis and forecasting, decomposition offers a helpful abstract paradigm for thinking about time series in general and for better comprehending difficulties.

So, what is stationary?

For time series, stationarity is a crucial factor. In terms of making predictions, a time series is considered stationary if its statistical features remain constant, i.e., if its mean, standard deviation, and covariance are all constant and unrelated to the passage of time. Ideally, we will aspire to have a stationary time series model, but it is not always stationary.

To learn whether a time series has a unit root and is thus, non-stationary, a Unit Root Test is often run. The alternative hypothesis defines a series as stationary if there is no unit root in the data, whereas the null hypothesis defines a unit root as the existence of a root in the data. When it comes to seasonality, it is used to describe cyclical patterns of behavior, such as increases and decreases that occur at regular intervals. These can be environment-related factors, such as temperature swings, and are included in the category of “natural causes.”

Operations of a business or administrative nature, such as the beginning or end of a fiscal year.

Cultural practices, such as celebrating holidays or adhering to a certain religion.

Variations in the calendar, such as the number of Mondays in a month or the dates of holidays from one year to the next. To model and predict the values of a single variable, the input for a time series model should only include data for that variable.

Autocorrelation is a measure of the linear relationship between the lagged values of a time series, like how correlation evaluates the strength of a linear link between two variables. The **Autocorrelation function** (**ACF**) considers all prior data, regardless of future or current effects. It calculates the t-1 correlation and contains all delays between t and (t-1). Pearson's formula is used to calculate correlation. *Figure 8.3* represents a simple matrix of the various aspects of the time series:

	Trend	Seasonality	Irregularity
Time	Fixed Time Interval	Fixed Time Interval	Not Fixed Time Interval
Duration	Long and Short Term	Short Term	Regular/Irregular
Nature	Gradual Upward/Downward trend	Pattern repeatable and swings between up or down	Short and Not repeatable errored or high fluctuation
Prediction Capability	Predictable	Predictable	Challenging

Figure 8.3: *Time Series Aspects Matrix*

Types of time series methods

The various techniques for analyzing timed data are known as “times series approaches.” **Autoregressive** (**AR**), **Moving Average** (**MA**), **Autoregressive Moving Average** (**ARMA**), **Autoregressive Integrated Moving Average** (**ARIMA**), and **Seasonal Autoregressive Integrated Moving Average** (**SARIMA**) are all common methods. Let us look at some of the methods.

Autoregressive (AR) model

In an autoregressive model, information from prior time intervals is used as input to a regression equation, which subsequently predicts the value at the following time interval. The number of historical time steps to be employed in the analysis is controlled by a single parameter, p, in the AR model. The order of the autoregressive model may be determined by inspecting the **partial autocorrelation function** (**PACF**). By regressing the time series at all shorter lags, the PACF offers the partial correlation of a stationary series to its own past values. PACF calculates t and (t-1)’s correlation; it doesn’t consider (t-1) time delays.

Moving average model

In the field of time series analysis, the **moving-average model** (**MA model**) is a popular technique for modeling univariate time series. It requires the linear connection between the output variable and the current and prior values of a stochastic (not totally predictable) element. In contrast to the AR model, the finite MA model always stays static.

Autoregressive Moving Average (ARMA) Model

This model is the result of combining elements from the AR model with the MA model. For the purpose of predicting the future values of the time series, it takes into account the influence that earlier lags have had, in addition to the residuals. In ARMA, it is expected that the time series is stationary and that when it does vary, it does so evenly around a certain time. Additionally, it is assumed that when it does fluctuate, it does so around the same time.

Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA models are part of a family of models that explain a particular time series based on its historical values, also known as its own lags, and the lagged prediction errors. This kind of model is known as an explanatory model. The equation may be used to make predictions about the values of the future. ARIMA models can be used for the modeling of any 'non-seasonal' time series that displays patterns and is not made up of random white noise or error.

Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

SARIMA stands for Seasonal-ARIMA, and it incorporates the influence that seasonality has on prediction. ARIMA fails to capture the information implicitly related to seasonality, despite the fact that the significance of seasonality is extremely obvious.

The **autoregressive** (**AR**), **integrated** (**I**), and **moving average** (**MA**) components of the model are all identical to those found in ARIMA. The SARIMA model becomes more robust if the seasonality variable is included. These models may be used to make predictions for both short-term and long-term tendencies shown by data. Other forms of time-series forecasting models, such as ARIMA models, are often regarded to be less accurate than SARIMA models, which are usually believed to be more accurate. In addition to this, interpreting and using SARIMA models is not too difficult.

Vector Autoregressive (VAR) Model

This is a common kind of multivariate time series model that uses historical data on a given variable in conjunction with data on other variables to predict future behavior. This model is known as the **vector autoregressive** (**VAR**) model.

VAR models are distinguished from univariate autoregressive models by the fact that they permit feedback to occur between the parameters that are included in the model.

It is a strategy that is both methodical and flexible, and it is designed to capture complicated behavior in the actual world. It can capture the interconnected dynamics of time series data and provides higher predicting performance overall. VAR models provide a framework for completing crucial modeling objectives, so they have historically seen widespread usage in the fields of finance and econometrics. A VAR model is a kind of mathematical model that consists of a set of equations that reflects the connections between different variables.

Vector Error Correction Model (VECM)

A broad framework that is utilized to represent the dynamic interaction among stationary variables is referred to as the VAR model. Therefore, the first thing that should be done when performing a time-series analysis is to establish whether the data is stationary. If that this is not the case, return to the beginning of the series and try again. In most cases, the initial differences in your time series will remain stationary even if the levels (or log levels) of the series itself are not. For reliable estimation of the relationships between time series, the VAR framework must be modified if the series is not stationary.

There are non-stationary data types that are co-integrated, so VECM puts extra restrictions on what may be done with the data. It incorporates the information pertaining to the co-integration restrictions into its specifications.

When the differences between the variables remain constant throughout time, we can use a special form of the variance analysis of the covariance (VAR) model known as the **vector error correction** (**VEC**) model. In addition to this, the VEC can take into account any cointegrating correlations that exist between the variables. We can comprehend both long-term and short-term equations with the help of VECM. It is necessary for us to count the number of connections that include cointegration.

Process Flow of Time Series Forecasting

A time series forecast process is a set of steps or a recipe that takes you from problem definition to having a time series forecast model or set of

predictions. The following image summarises the steps involved in time series forecasting:

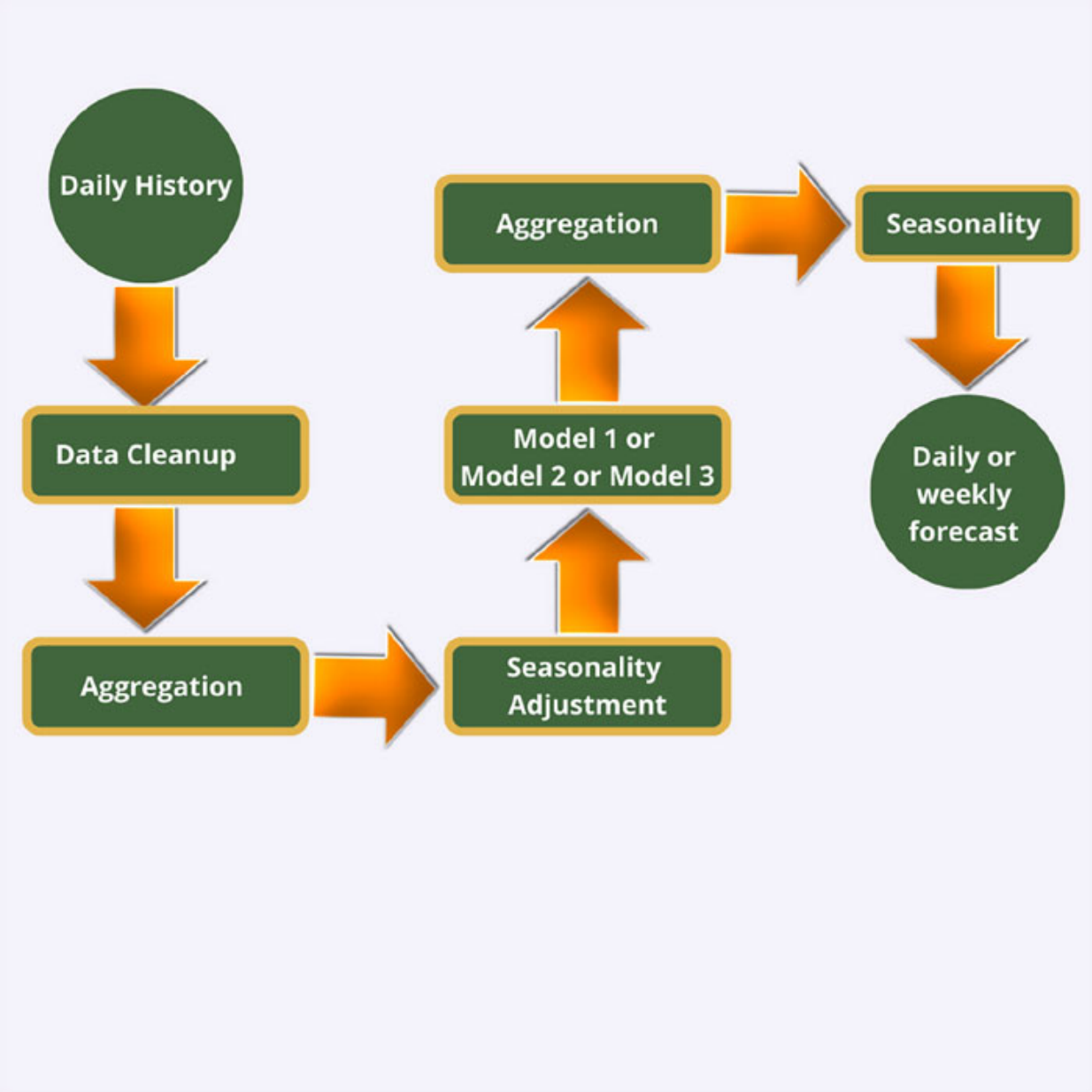


Figure 8.4: *Time Series Forecasting Process Flow*

The five fundamental steps in a forecasting task are summarised below:

- **Problem identification**: The forecast must be carefully considered in terms of who needs it and how it will be used. This is described as the most difficult part of the process, most likely due to the fact that it is completely problem-specific and subjective.

- **Information gathering**: The acquisition of historical information for the purposes of analysis and modeling. This also entails having access to subject matter specialists and collecting data that may assist in the most accurate interpretation of previous data and, eventually, projections.
- **Initial exploratory analysis**: Simple techniques like charting and summary statistics are used so that the data may be comprehended on a deeper level. Examine the plots and summarize and make notes on any noticeable temporal structures, such as trends and seasonality; anomalies like missing data, corruption, and outliers; and any other structures that may affect predicting.
- **Model selection and fitting**: The optimal model choice depends on historical data availability, the forecast variable's relationship with explanatory variables, and the intended use of the forecasts. Typically, two or three candidate models are compared. Each model has assumptions and parameters to be estimated using historical data.
- **Deploying and testing a forecasting model**: The model is put to use or deployed in order to generate predictions, and the accuracy of those forecasts and the model's competence are both evaluated. Back-testing using old data or waiting for fresh observations to become available can be necessary before making a comparison.

EDA and Statistics of time series forecasting

Exploratory data analysis (**EDA**) is a technique used by data scientists to study and investigate large data sets in order to summarise their essential properties. Data visualization techniques are frequently employed in this process. EDA assists data scientists in determining the most effective way to alter data sources to obtain the answers they require, making it easier for them to detect patterns, identify anomalies, test hypotheses, and verify assumptions. EDA is primarily used to investigate what data can disclose beyond the formal modeling or hypothesis testing work and to gain a deeper knowledge of the variables in the data collection and the relationships between those variables. It can also assist you in determining whether the statistical techniques you are considering for data analysis are appropriate for your situation. It can also assist you in determining whether the statistical

techniques you are considering for data analysis are appropriate for your situation.

Exploratory data analysis is a method of conducting data analysis that uses a range of approaches to develop a sense of intuition regarding the data.

- Make sure you're not losing your mind (in order to ensure that the insights we are making are derived from the correct dataset).
- Determine where information is lacking.
- Examine the data to see if there are any outliers.
- Make a summary of the information.

The following are examples of statistical functions and approaches that can be performed with EDA tools:

- In order to construct graphical representations of high-dimensional data, including numerous variables, techniques like clustering and dimension reduction can be used.
- Each field in the raw dataset is shown as a single-variable visualization, along with summary statistics.
- Using bivariate visualizations and summary statistics, you can determine the link between each variable in the dataset and the target variable you're interested in learning more about. Multivariate visualizations are used to map and comprehend the interactions between distinct fields in the data and between different fields in the data.

For further details on EDA, refer to the previous chapters, as the methodology remains the same.

Time series forecasting using Orange

The add-on includes a wide range of features, such as ARIMA and VAR models, model assessment, time series pre-processing, seasonal adjustment, and several different visualizations.

Inputs

- Data: Any type of data table.

Outputs

- Time series: A data table that has been reinterpreted as a time series.

Any data table may be reinterpreted by this widget as a time series, making it possible to utilize it in combination with the other widgets that come included with this add-on. You can define which data property in the widget corresponds to the time variable. Refer to *Figure 8.5*:

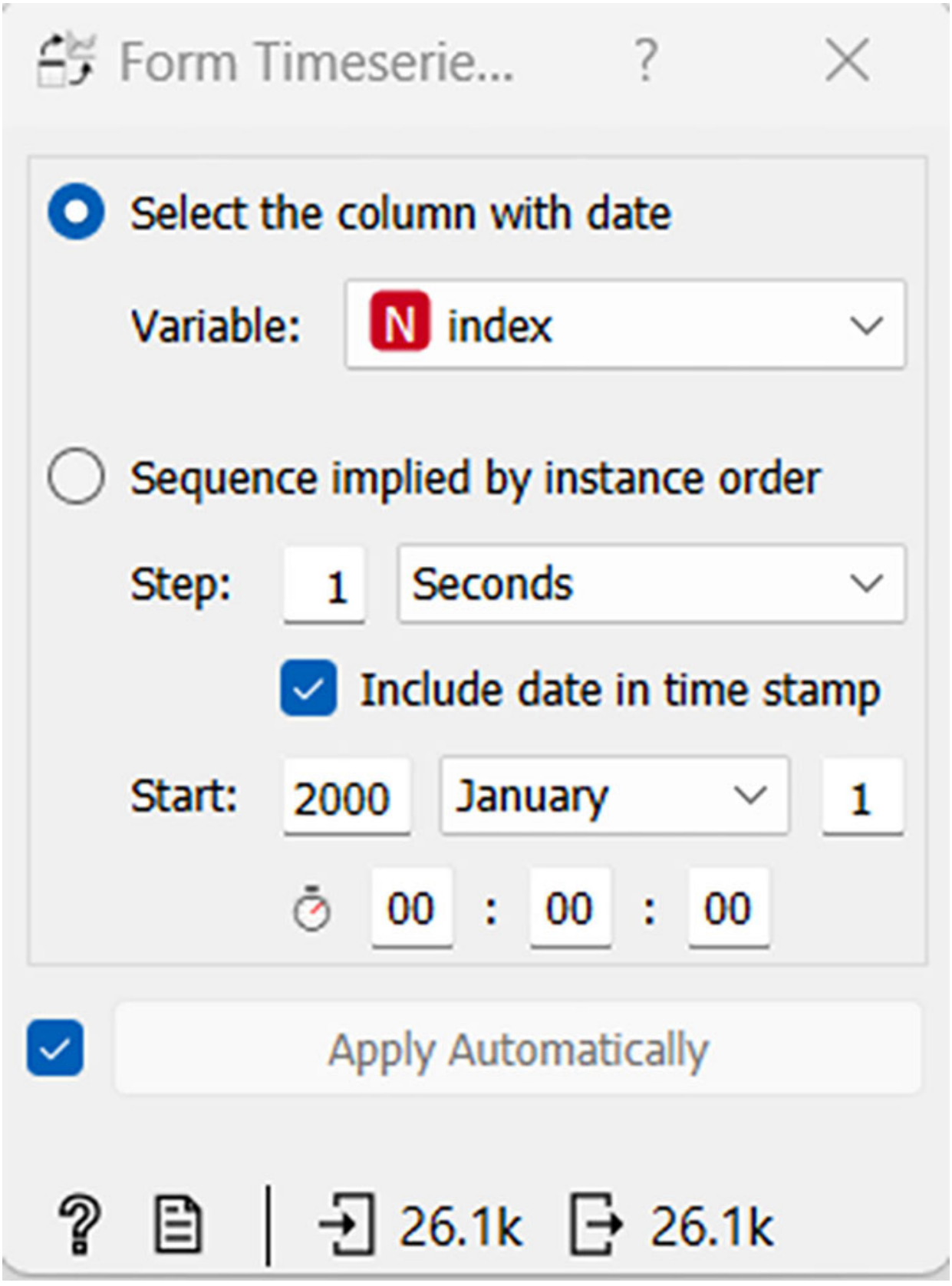

Form Timeserie...
?
Select the column with date
Variable:
N
index
Sequence implied by instance order
Step:
1
Seconds
Include date in time stamp
Start:
2000
January
1
00
:
00
:
00
Apply Automatically
26.1k
26.1k

Figure 8.5: *Orange Time Series Window*

1. The time feature represents the sequence and intervals between measurements, and can take on any continuous value. .
2. Additionally, you have the option of stating that the time series sequence is inferred from the instance order.

Any data-emitting widget, such as the File widget, may send information to this widget to be processed. It is essential to keep in mind that every time you execute processing with Orange core widgets like the select columns widget, you will need to re-apply the conversion into time series. This is one of the most significant aspects of Orange. Refer to *Figure 8.6*:

Figure 8.6: *Orange Time Series Process Flow*

Time series cheat sheet

The following is a simple cheat sheet that will give you an overview of time series forecasting, various methods, and so on. Refer to *Figure 8.7*:

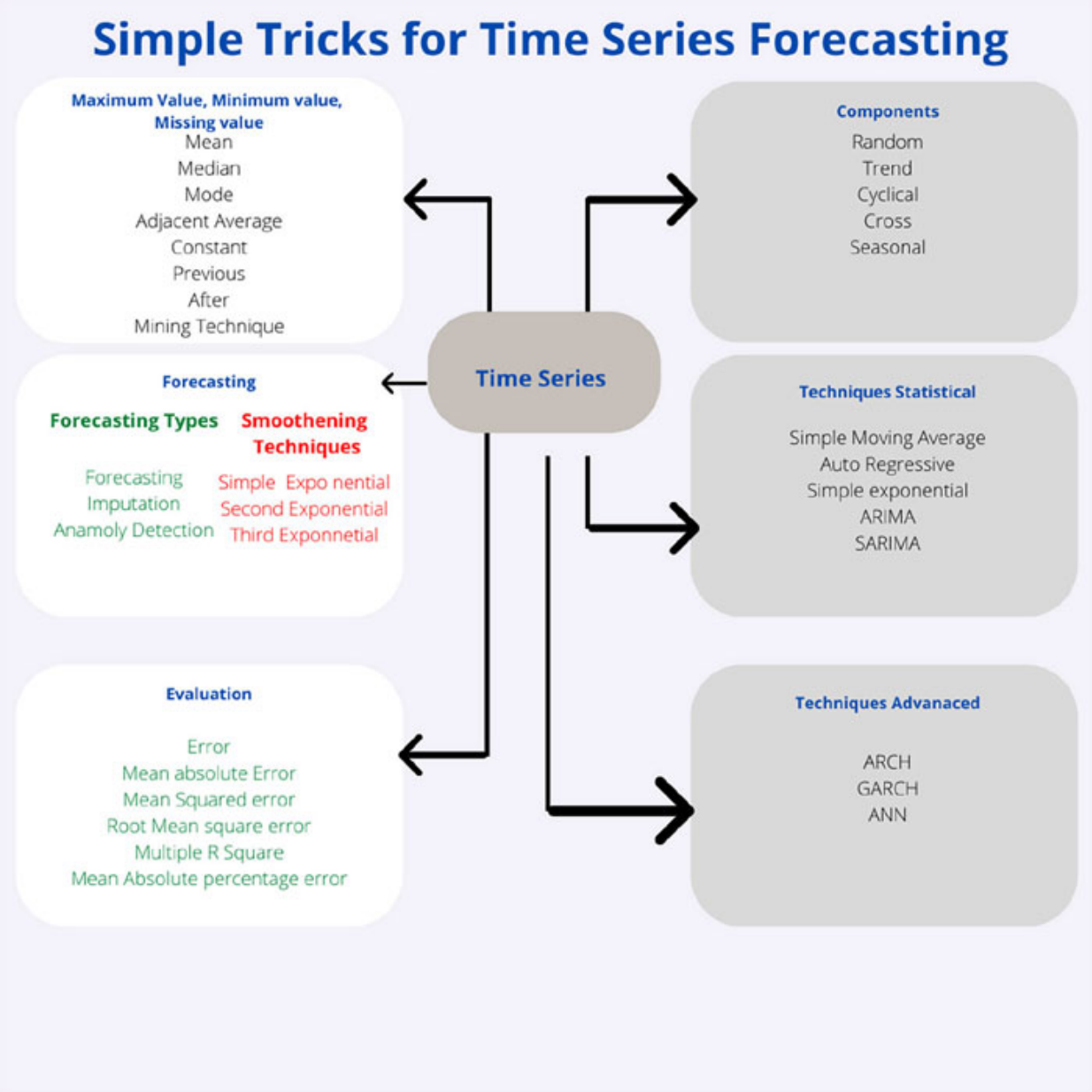


Figure 8.7: *Time Series Cheat Sheet*

Conclusion

Machine Learning has been considered as the future of data analytics. It is a discipline that involves the design, development, deployment, and use of systems and models that can extract knowledge and intelligence from data. Time series data is one of the most common types of data, and it is also quite interesting.

This type of data contains the same information repeatedly in different formats over time. Time series forecasting is something that has been around for a while, and it is used in order to make accurate predictions about the future. For example, we use it when we want to predict the weather or the market.

In the next chapter, we will discuss the basics of image analysis and look at how image data can be processed and utilized for machine learning activities. We will also see how to analyze the image data and cover some of the practical applications.

Points to remember

- What is a Time Series? It consists of a set of measurements obtained at regular intervals throughout time.
- Predicting future values from historical data is its primary use. Trend, seasonality, irregularity, and cyclicity are some of the elements you may find in a time series study.
- Time series analysis methods like Autoregressive, moving average, autoregressive integrated moving average, seasonal autoregressive integrated moving average, vector autoregressive, vector error correction etc.
- Univariate time series are those in which observations are made on just one variable at a given moment in time.
- Multivariate time series are those in which two or more variables are tracked over a given period.

Multiple choice questions

1. Which of the following is a technique for time series forecasting?

    a. Moving Average

    b. Pareto Analysis

    c. Outlier Detection

    d. Bar Chart

2. Which of the following is the first phase of time-series analysis?
    a. Conduct preliminary regression analysis
    b. Visualize the data trend
    c. Compute a moving average
    d. Find the Pearson correlation coefficient
3. Which of the following is not one of the aspects of time-series analysis?
    a. Autocorrelation
    b. Stationarity
    c. Predictable
    d. Seasonality
4. Vector Error Correction (VEC) is a special form of the Variance Analysis of Covariance (VAR) model.
    a. True
    b. False

Answers

Question Number	Answer
1	A
2	B
3	C
4	A

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 9

Image Analysis

> *Image analysis (sometimes referred to as "computer vision" or "picture recognition") is the capacity of a computer to distinguish characteristics inside an image.*

Guess you might be familiar with photo editing apps like Google Photos or Apple Photos. They both employ certain fundamental image analysis characteristics to detect faces and classify them in your images, allowing you to see all your photos of a specific individual. You can type "dog" into the search feature of either app to easily access your collection of puppy images or type "beach" to retrieve your pictures from your tropical trip. The practice of extracting useful information from photographs, often digital photos using digital image processing methods, is known as image analysis. Simple image analysis jobs include scanning barcodes, while more complex ones include facial recognition.

Structure

In this chapter, we will cover the following topics:

- Basics of deep learning
- What is image analysis
- Image analysis process flow (identify problem statement, check data availability, choose the right pre-processing tool, and define the end goal)
- Image analysis forecasting using Orange

Objectives

You'll learn how image analysis permits us to extract useful data from photos via image processing and, more recently, via computer vision. Recent developments in machine learning and deep learning have made it possible for us to provide imaging data in near-real time. The potential benefits of information extraction are much more than most people realize. In addition to improving video surveillance, this has applications in healthcare, manufacturing, safety, and so on.

Medical imaging applications may help us make more precise and timely diagnoses. Using computer vision can assist in hazardous production environments and other settings where human workers are at risk. From elementary procedures to sophisticated uses like pattern recognition in materials science and 3D modelling, the development of image processing techniques has been impressive.

Get started with Deep Learning

Before diving deep into image analysis, let us look at the basics of deep learning. As a kid, someone would have fed you a lemon wedge for the first time and had fun observing your reaction. Prior to this, you might never know what lemon tastes like. Similarly, you understand things around you by learning and recognizing patterns. When a computer does the same (pattern recognition), it is called "Deep Learning," which functions based on neural networks in the human brain. Human-created feature representations are the backbone of traditional machine learning techniques. As a result, ML boils down to a simple process of adjusting weights to provide the most accurate predictions. The machine learning area known as "**deep learning**" (**DL**) employs several layers to study how to best represent input data or information. These layers are formed by neural networks similar to the human brain. Learning patterns is one area where DL shines. Some of the real-time use cases of deep learning are as follows:

- Speech recognition and synthesis
- Image processing
- Language processing

The most popular neural network architectures are **Artificial Neural Networks** (**ANN**), **Convolutional Neural Networks** (**CNN**), and **Recurrent Neural Networks** (**RNN**). The basic difference between a traditional machine learning algorithm and deep learning is shown in *Figure 9.1*:

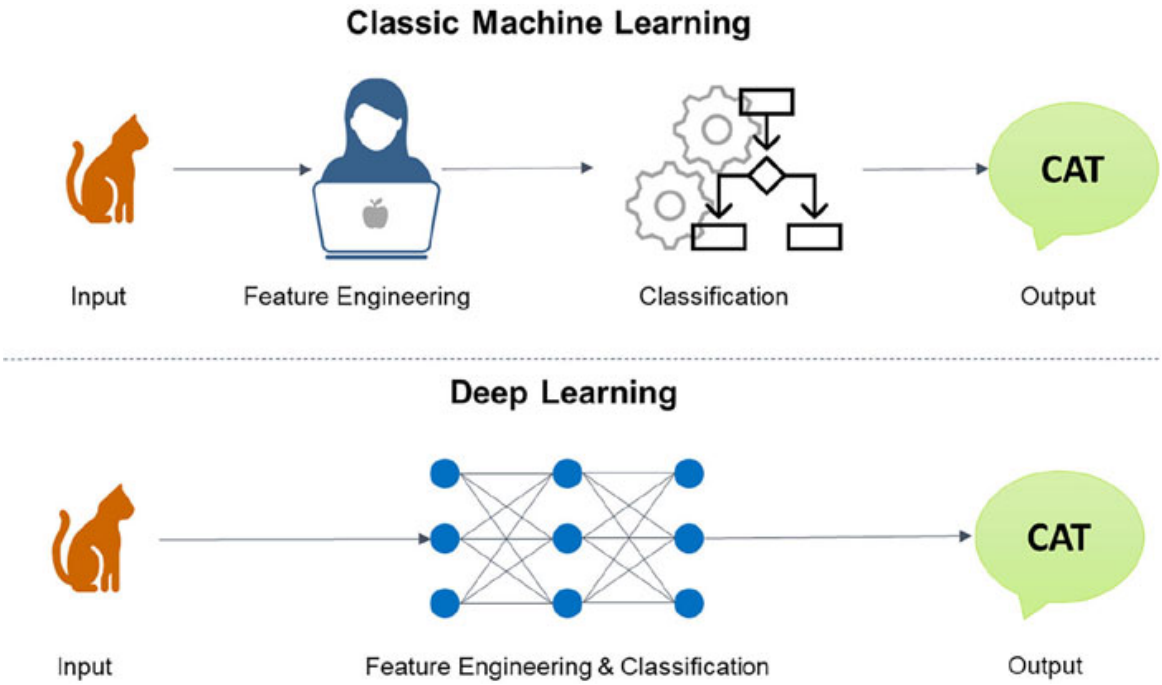


Figure 9.1: *How Deep Learning Works*

The inspiration for ANN is the human brain. Biological neuron works as follows: receives signals *in the form of electrical or chemical signals*, which are processed by the cell body and may be amplified or inhibited, and passes them on to the next neuron. Frank Rosenblatt developed the first practical artificial neuron, the perceptron, in 1958. The perceptron is a one-layer neural network that may be combined with others to create more complex structures. The perceptron is the fundamental neural network used in deep learning. The perceptron is a kind of simple artificial neural network unit whose purpose it is to do computations to discover and learn features or insights from the input. It can have a single layer or multilayer. The parts of the perceptron can be seen in the following figure: the input layer, weights & bias, net sum, and activation function:

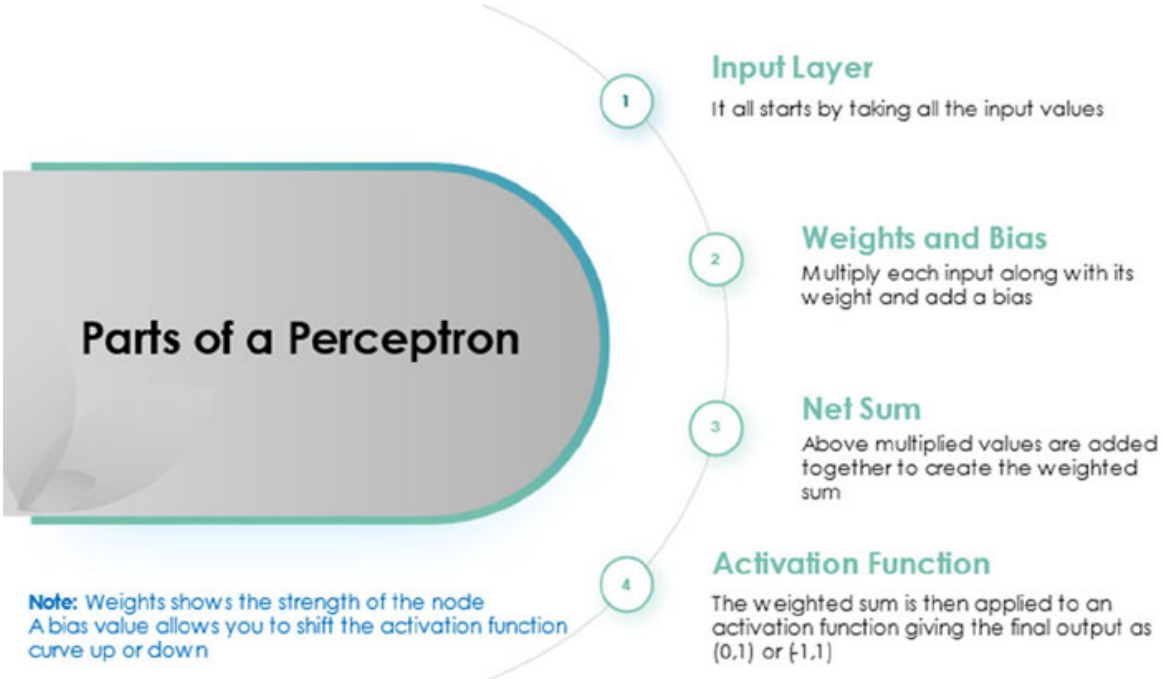


Figure 9.2: *Parts of an ANN or Perceptron*

Without going into the details, a simple overview of the math and working of a perceptron is shown in the following image:

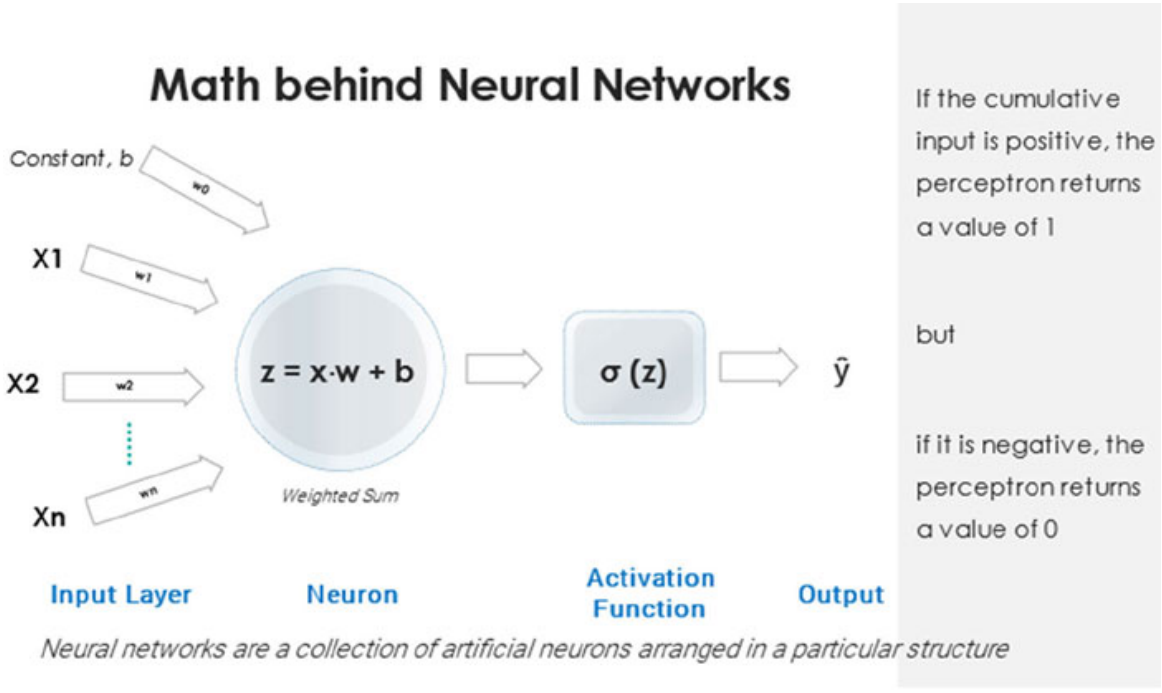


Figure 9.3: *Math Behind ANN*

Image analysis

Understanding, classifying, and quantifying pictures of all kinds require the use of image analysis as a foundational technique. Images might be monochromatic (grayscale) or multi-colored (color), multispectral (many spectral channels or wavebands), or hyperspectral (several contiguous wavebands spanning a single spectral area). Image analysis is an essential tool used to identify, categorize, and quantify pictures of varying sorts and formats.

Manual image analysis is a process in which a human analyst manually examines an image and makes observations and measurements based on the visual information present in the image. Technological advancements allow for faster observation and documentation of results. Classification methods for digital images are used to label parts of an image in line with a taxonomy that makes sense for the intended usage. Two or more similar images are used for pattern recognition, and the results may be further characterized by the kind of change and quantified by the amount or degree of change.

Guess it might sound interesting to you now!

Applications that use digital image processing are rapidly expanding into almost every industry because of the rapid development of related technologies. Let’s take a look at the many applications of digital image processing.

Here are a few examples:

- The term “image sharpening and restoration” refers to the manipulation of pictures acquired by a digital camera to improve the quality of the image or accomplish the desired outcome. It refers to carrying out the tasks that are typical for Photoshop, including zooming in and out, blurring the image, sharpening it, converting the grayscale to color, finding edges in the opposite direction, and image retrieval and identification.
- The term “medical image processing” refers to the use and investigation of 3D image datasets of the human body. These datasets are typically obtained from a **computed tomography** (**CT**) or **magnetic resonance imaging** (**MRI**) scanner. The medical field, in addition, is helpful in the process of diagnosing diseases, directing medical procedures like surgery planning, and doing research. Examples of the uses of medical image processing are gamma-ray imaging, PET scans, X-ray imaging, medical CT scans, and ultraviolet imaging.
- Vision for machines and robots involves employing a mix of camera hardware and computer algorithms to enable robots to analyze visual input

from the outside environment. This ability of a robot to see is referred to as “machine vision.” The creation of digital images is assisted by several automated equipment. Robots can navigate their environments using image processing techniques. The capacity of a robot to observe, process and recognize is crucial to its ability to perform complicated tasks in a dynamic environment. Due to the high level of sophistication inherent in digital camera technology, high resolution pixel arrays may be sent to the robot’s onboard computer. Algorithms from digital image processing are utilized to improve and analyze these pictures.

- **Color processing**: Each pixel’s color information is recorded in a digital color image. Each pixel in a color image has three values that together define the hue, saturation, and brightness of the overall scene. The digital image’s contents are saved, with the brightness information in each dark band containing different pieces of data. Understanding how these color images may be transferred, stored, and encoded is also a part of this process.
- Video processing may be defined as the act of assembling a series of still images in a way as that provides the illusion of continuous motion. Conversion of color spaces, noise reduction, detection of motion, and frame rate conversion are all included in this process.

When it comes to signal processing, in particular image processing, video processing is a special instance that stands out from the rest. Video filters are often used in video processing, with video files or video streams serving as both the input and output signals. Techniques for processing video are often used in electronic devices like televisions, **videocassette recorders** (**VCRs**), DVD players, video codecs, video players, and video scalers.

What is an Image

Before we move on to image analysis, we need to understand what an image is. An image can be defined as a function represented by a function f(x, y) in which x and y are planar values. At any two distinct points of x and y values, the value of the function can be equated to the intensity of the image. If the values of x, y, and f’s amplitude are all finite and discrete, we refer to the resulting picture as a digital image. Each of the picture’s pixels represents a specific location and value in the final digital image. The height and breadth of an image are used to represent it, and these values are determined by the number of pixels. For instance, if the width and height of a picture are 500 pixels and 400 pixels, respectively, then the total number of pixels that make up the image is 200000.

The pixel in question is a location on the picture that has a particular tint, degree of opacity, or color assigned to it. In most instances, it takes the form of one of the following:

- Grayscale images are represented by pixels, each of which may have a value between 0 and 255. (0 equals black, and 255 equals white).
- Using the RGB color model, a pixel is composed of three numbers ranging from 0 to 255. Color saturation is shown by the integer values for red, green, and blue.
- RGBA is an expansion of RGB with an extra alpha field, which depicts the opacity of the picture.

The processing of images involves predetermined sequences of actions to be carried out on each pixel that makes up a picture. The image processor is responsible for carrying out the first series of operations on the picture, which are carried out one pixel at a time. After this step has been completed in its entirety, it will move on to the next operation, and so on. Any pixel in the image's coordinate space may have its output value calculated according to these processes.

Image processing

After an understanding of the image, let us move on to image processing. By digitizing a picture and performing different procedures on it, useful data may be extracted. This is known as image processing. Images are normally interpreted as 2D signals by the image processing system unless a particular set of signal processing methods is being applied. There are five basic elements that describe the purpose of digital image processing:

- **Visualize**: To look for things in the picture that are not visible
- **Recognize**: To identify or recognize things in a picture
- **Sharpen and restore**: To create an upgraded image from the original picture
- **Recognize patterns**: To measure the numerous patterns that surround the items in a picture
- **Retrieval**: To use a large digital picture database, you may look for photos similar to the one you're trying to find

The three fundamental processes in image processing are as follows:

- Importing the picture using image capture software
- Editing and analyzing the picture
- Looking at the result, which might be an altered picture or an analytical report

There are two distinct categories of image processing methods:

- **Analog image processing**: When it comes to prints and photographs, analog image processing might be useful. Image analysts use a wide range of interpretative frameworks with these visual methods.
- **Digital image processing**: Tools for digital image processing make it possible to edit digital photos on a computer. All types of data need three primary procedures before they can be used effectively by digital techniques: pre-processing, augmentation, display, and information extraction.

Sources of digital images

Digital images are two-dimensional images represented numerically. They might be raster or vector type.

Types of digital images

In general, we consider four types of images:

- **Binary photos**: Binary pictures can have two values: 0 and 1 or black and white. Binary pictures use just one binary number to represent each pixel, so they are also known as 1-bit images; for example, **optical character recognition** (**OCR**)./p>

  Binary pictures are formed from grayscale images using the threshold technique. Every pixel over the threshold value is colored white (1), while those in the following threshold are turned black (0).
- **Greyscale pictures**: These are also referred to as monochrome or one-color photos. Gray-level photos include just gray-level information and no color information. The number of grey levels available is governed by the number of bits used for each pixel./p>

  For example, a grayscale picture with 256 distinct grey levels should have 8 bits per pixel.

  Medical imaging and astronomy employ data with 12 or 16 bits per pixel.

- **Color pictures**: Color pictures are formed as three-band monochrome image data, with each band representing a distinct color. There is gray-level information in each spectral band, which is the real information recorded in the digital image./p>

  Color pictures are sometimes known as RGB images since they are represented by the colors red, green, and blue. Color graphics would have 24 bits per pixel if the 8-bit monochrome standard was used as a reference, with 8-bits for each of the three color bands (red, green, and blue).
- **Multiple-spectral photos**: This sort of picture contains information that is outside the typical range of human perception. The information portrayed is not directly visible to the human system, so these are not pictures in the traditional sense. The information is displayed visually by mapping the various spectral bands to RGB components. Ultraviolet, infrared, X-ray, radar data, and acoustic pictures are all part of multispectral imaging.

Levels of digital image processing

Image processing is classified into three levels:

- **Low level**: It includes pre-processing, noise reduction, contrast improvement, picture sharpening, and other similar tasks. Both input and output in low-level processing are pictures. A simple rotation of the image can be an effective augmentation technique that can help improve data size too.
- **Mid-level**: It includes picture segmentation, image description, and object identification. At the mid-level, inputs are often photos, and outputs are usually image properties. Image processing includes the fundamental steps of feature extraction and dimensionality reduction. Since pictures are mapped to lower dimensions feature vectors, these processes are regarded as intermediate-level operations.
- **High level**: Entails “making meaning” of a collection of identified items. This is commonly used in computer vision.

Applications of digital image processing

- **Medical**: Chromosome identification, automatic detection, and categorization of cancers in X-ray pictures, **magnetic resonance imaging** (**MRI**), processing of CAT scan, ultrasound images, and so on

- **Computer vision**: Bin-picking, robotics, teleoperation, autonomous systems, and part identification in manufacturing lines
- **Remote sensing**: Meteorology and climatology, resource tracking, geographic mapping, crop forecasting, urban growth and weather, flood management, and more
- **Radar and sonar**: Target detection and identification, aircraft or missile guidance or movement
- **Image Transmission**: Teleconferencing, satellite and computer network communications, HDTV and 3DTV, military communications, and space missions
- Identification systems, including those based on the face, iris, fingerprints, and security checks at banks and airports

Process flow of image processing

Building an autonomous interactive system that can extract symbolic descriptions from a picture is the overarching objective of image analysis. You may think of pattern recognition as the opposite of computer graphics. It begins with a sight or image and turns it into an abstract description, which is then translated into a series of numbers or symbols.

The following diagram illustrates a whole image processing or computer vision system:

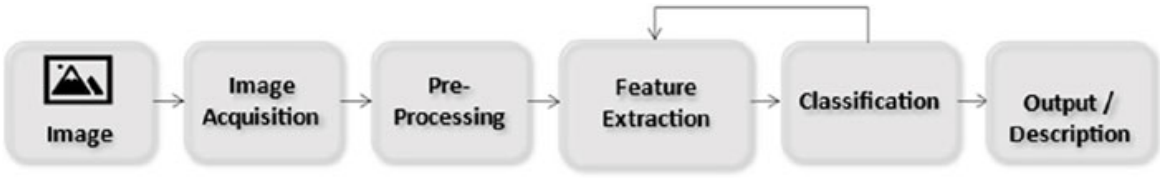


Figure 9.4: *Image Processing System*

The following points discuss the flow of the image processing system:

- Image data is collected by the sensor
- **Pre-processor**: Noise reduction, data compression, and so on
- **Segmentation**: Isolate the items of interest using edge detection and region-growing
- **Extraction of features**: From segmented objects, extract a representative collection of features

- Each item or location is classified by the classifier, which also collects properties from them
- The structural analyzer establishes the connections between the categorized primitives, and the original scene's description appears as the output
- The world model is utilized to direct the analyzing system at each level; each stage's outcomes can be applied to improve the global model, which is built with as much a priori knowledge about the scene as feasible before we begin to study it

Following is an example of a more thorough schematic of an image analysis system:

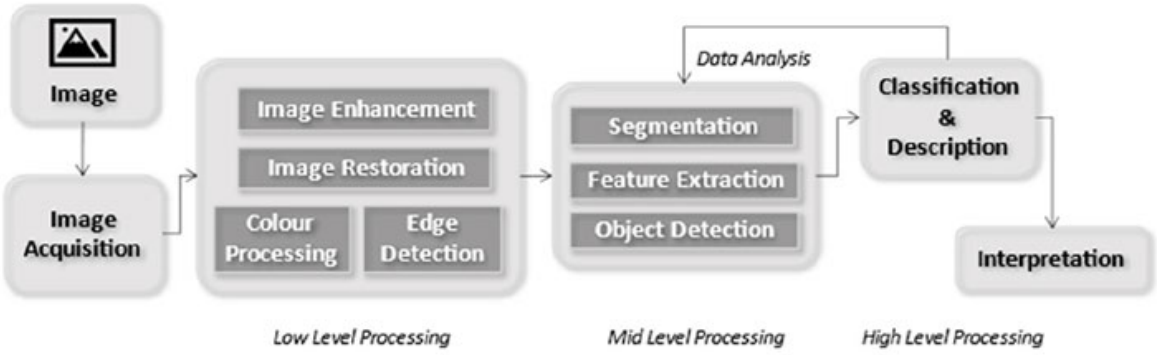


Figure 9.5: *Detailed Image Processing System*

EDA and Statistics of image processing

The first step in many machine learning procedures is data exploration. However, there is no simple method for systematic data exploration on datasets for object recognition and image segmentation. Working with ordinary image datasets differs from object and segmentation datasets in several ways:

- The label and the picture are inextricably linked. Suddenly, whatever you do to your photographs must be carefully considered, as it can ruin the image-label mapping.
- Typically, a picture has many more labels.
- Several other hyperparameters to adjust (especially if you train on your custom datasets).

Evaluation results, investigation, and error analysis become significantly more difficult as a result. Choosing a single performance metric for your system might

also be challenging; in these situations, manual research may still be an important step.

When working on any machine learning topic, including object identification and picture segmentation, you should first evaluate the quality of your data. When developing object detection and image segmentation models, common data issues include the following:

- Dimensions and aspect ratios of images (especially dealing with extreme values)
- Asymmetry in the composition such as aspect ratios, bounding box size, and other factors must be addressed through appropriate dataset preparation.
- Data modeling strategy that is not consistent with the data

In practice, datasets are unlikely to contain photos with identical sizes and aspect ratios. Basic dataset statistics like aspect ratios, picture widths, and heights can help you make key decisions:

- Should you, and can you perform destructive resizing? (Destructive resizing refers to resizing that alters the aspect ratio.)
- What should your ideal output resolution and padding be for non-destructive resizing?

Deep Learning models may require additional parameters to be adjusted based on the aforementioned factors (such as anchor size and ratios), or they may have minimum input image size requirements that must be met. Here anchor size refers to a predefined set of bounding box used to detect objects of different scales and shapes in an image. If most of your photographs have the same proportions, your only option is to choose how much to scale them (if at all). This is mostly determined by the object's area, size, and aspect ratios. The aspect ratio range is mostly within (0.7... 1.5), with a slightly bimodal distribution. It is suspected that other "natural-looking" datasets would exhibit a similar distribution, for which a non-destructive resize -> Pad method should suffice. Padding essentially enlarges the region of an image that a neural network analyzes for various reasons. Padding will be necessary, but only to a tolerable extent that does not significantly increase the dataset's size. Your dataset may contain outliers, such as very broad photos combined with very narrow ones, making it more difficult. More complex solutions are available to avoid unnecessary padding. You may consider sampling batches of photographs based on aspect ratio. However, keep in mind that this might introduce bias into your sampling method, so make sure it is appropriate.

Image analysis using Orange

Orange has several useful add-ons, such as the Image Analytics add-on.

- When you click on the options menu, a dropdown list will be displayed, as shown in the following figure:

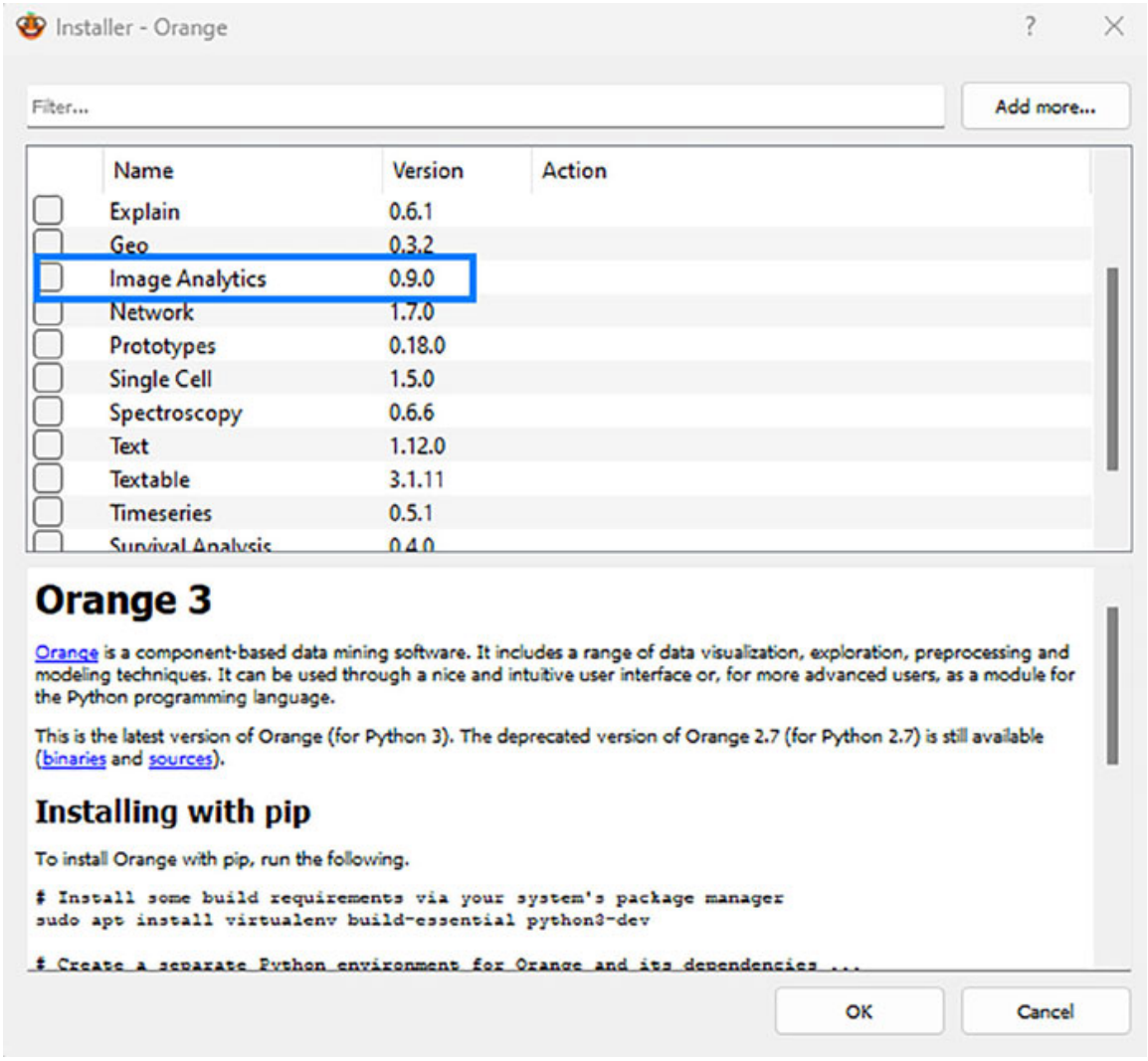


Figure 9.6: *Orange Image Analytics Add-on*

- Select add-ons and press the button to launch the add-ons interface.
- Check the box labeled "**`Image Analytics`**," and then press the "**`OK`**" button. Do not go forward until the installation is finished.

Orange may need to be restarted for it to operate. Look at the following for an illustration of where to locate the image analytics widget:

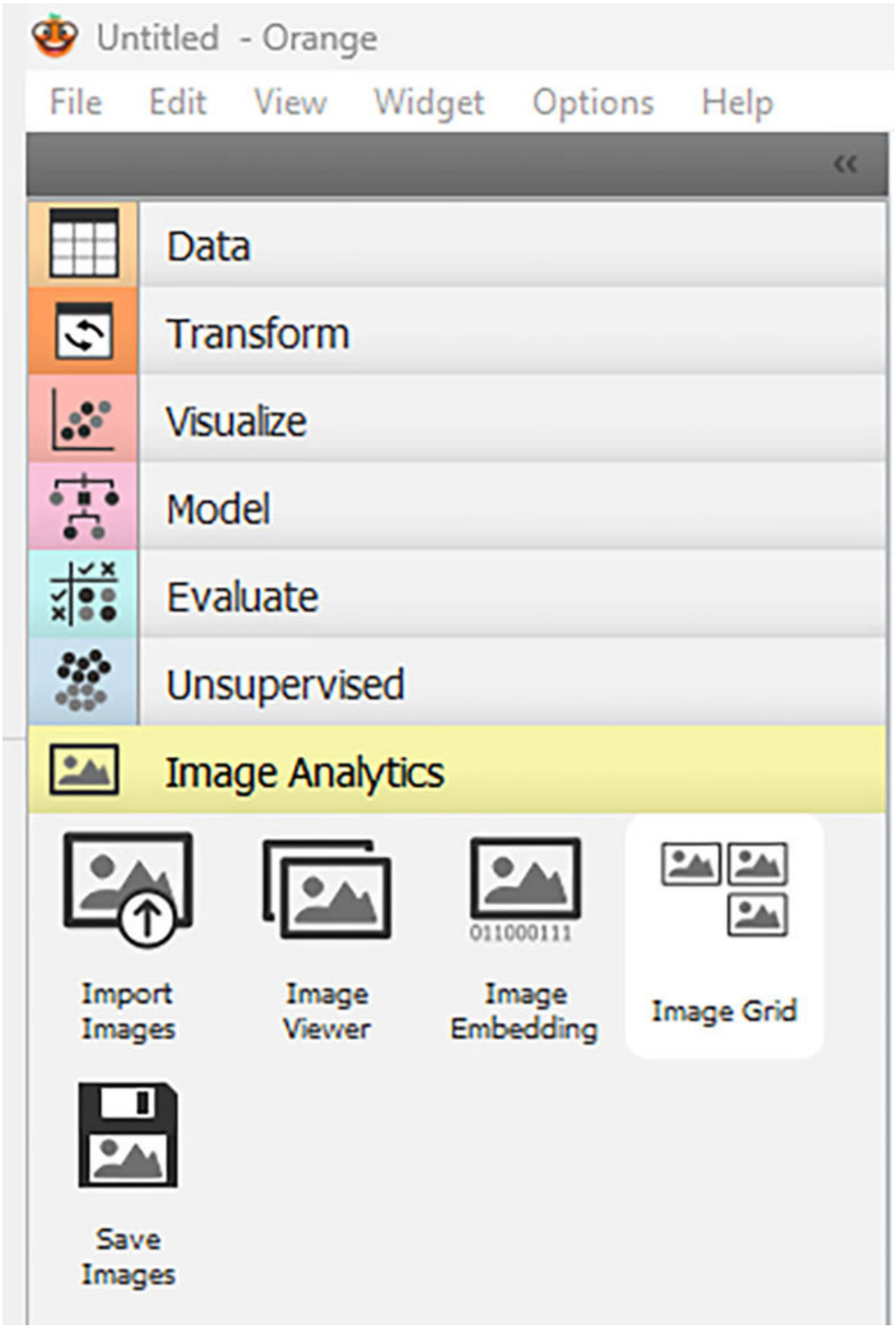


Figure 9.7: *Orange Image Analytics Add-on*

Import Images: The first step is to use the import images widget to import the picture. Consider this widget to be the file widget for images. The Import Images widget, on the other hand, takes a directory rather than a file.

- Insert the **`import images`** widget into the canvas.
- To open the interface, double-click on it.

The dropdown list displays the previous directory, while the center button opens any new directory. If you have added or deleted photos from the directory, you may refresh the material by clicking on the "**`reload`**" button. There is additional information text that shows the number of photographs in the directory.

Image Viewer: The image viewer widget will then be used to inspect the directory's content. This widget will show all the photos that have been loaded. This is beneficial since the full procedure can be completed without having to open Explorer.

- To the canvas, add an image viewer widget.
- Connect the import images and image viewer widgets.
- To access the interface, double-click on the image viewer widget.

Image Embedding: This is the most crucial widget in the image analytics package since it is where the magic happens. To your knowledge, classification and regression tasks require numerical data, and there is no acceptable method to conduct such tasks with photos unless we express them numerically. This is where the image embedding widget comes into play by transforming it to a vector of integers. The Image Embedding widget reads pictures and either uploads or evaluates them locally.

- Add an image embedding widget to the canvas.
- Connect the import images and image embedding widgets.
- To access the interface, double-click on the image embedding widget.

The embedder is one of the most significant factors for the image embedding interface. There are several embedders available for usage. Except for Squeeze Net, most of them require an internet connection to do the conversion. The following is a list based on official documentation:

- SqueezeNet is a small and fast image recognition model trained on ImageNet.
- Google's Inception v3 model was trained using ImageNet.

- VGG-16: ImageNet-trained 16-layer image recognition model.
- VGG-19: ImageNet-trained 19-layer image recognition model.
- **Painters**: A model that can identify artists from digital images of their work.
- DeepLoc is a model that has been trained to evaluate yeast cell pictures.

The following is an image of a typical image processing flow in Orange:

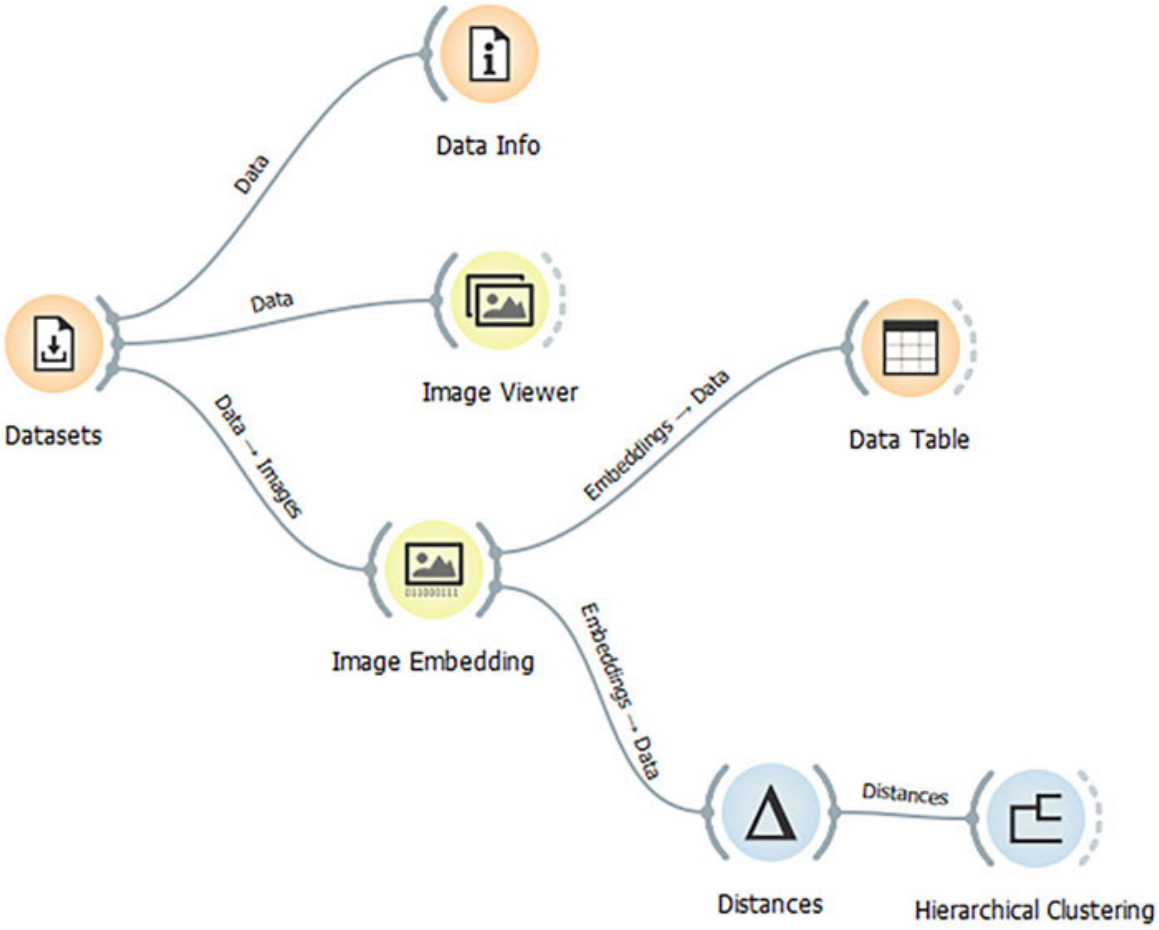


Figure 9.8: *Orange Image Processing Flow*

To see the results, link it to a data table widget. You should get something like the following image:

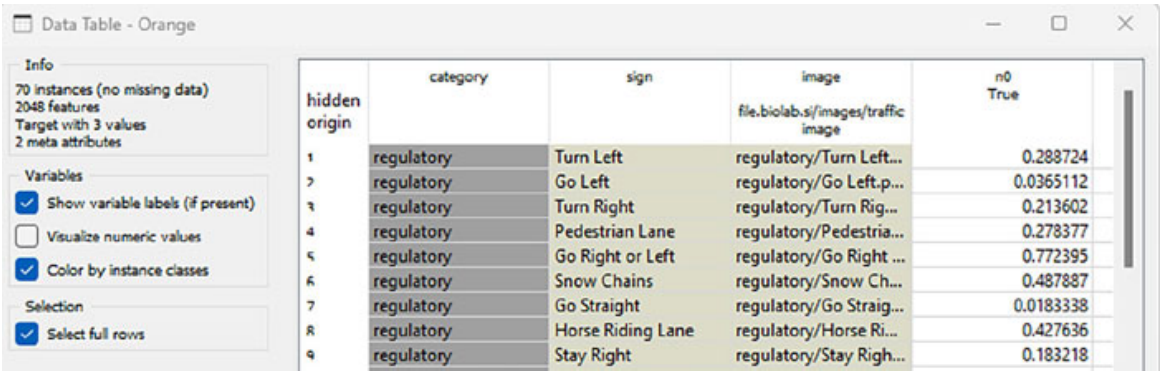


Figure 9.9: *Orange Data Table Widget*

You may further investigate it by feeding the output of the image embedding widget into the distances widget and the hierarchical clustering widget. A dendrogram, which is a tree diagram commonly used to depict the layout of clusters formed by hierarchical clustering, will be visible. Follow the given steps to create a dendrogram in Orange:

- Add a **`distances`** widget to the canvas.
- Connect the **`image embedding`** and **`distances`** widgets.
- Add a **`hierarchical clustering`** widget to the canvas.
- hierarchical clustering is used with the connect distances widget.
- To access the interface, double-click on the **`hierarchical clustering`** widget.

A typical dendrogram output after performing the preceding steps will look like the following image:

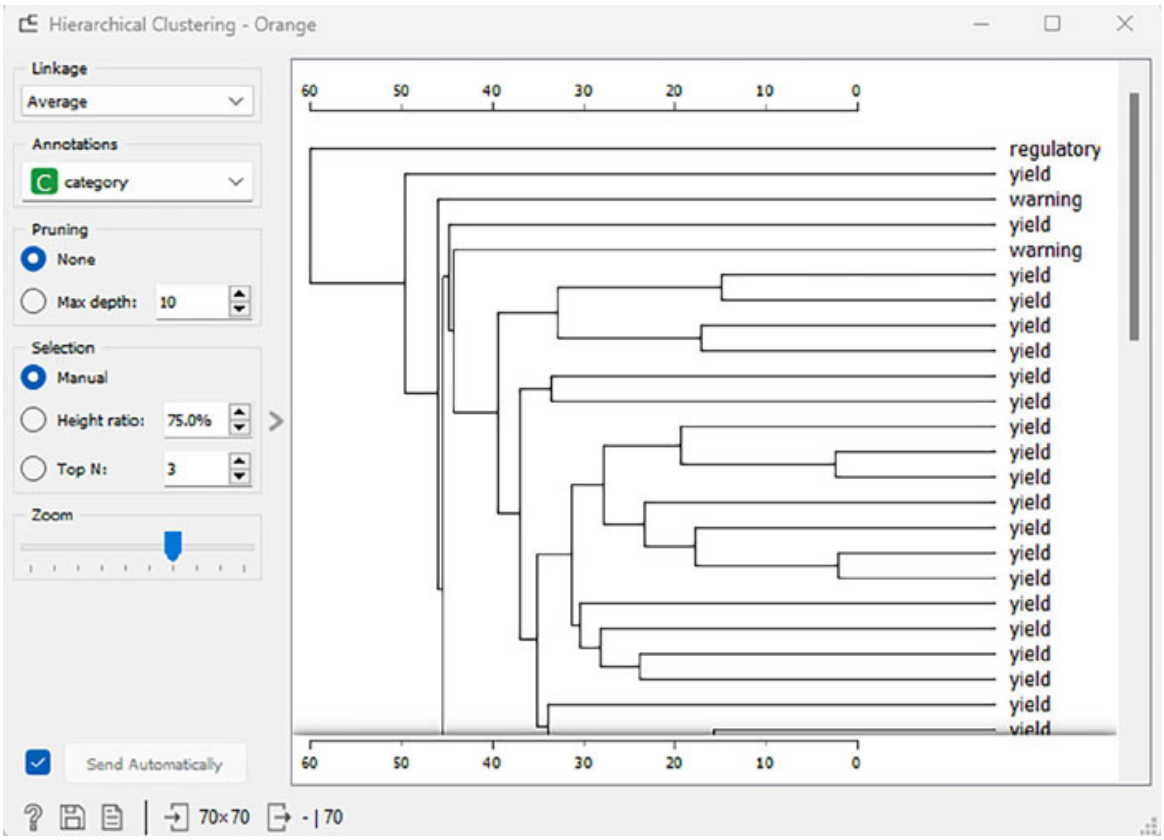


Figure 9.10: *Orange Hierarchical Clustering Dendrogram*

If you want to see things visually, you may use the image grid widget instead. The image grid widget, according to official documentation, may show pictures from a dataset in a similarity grid. This indicates that photos with similar topics are grouped together. It may be used to compare images and check for similarities or differences between specified data examples. Follow the given steps to create an image grid view in Orange:

- Insert an image grid widget into the canvas.
- Connect the image embedding and image grid widgets.
- To access the interface, double-click on the image grid widget.

After performing all the preceding steps, you should be able to get something like the following figure. The photos are clearly split into several groups depending on form and color.

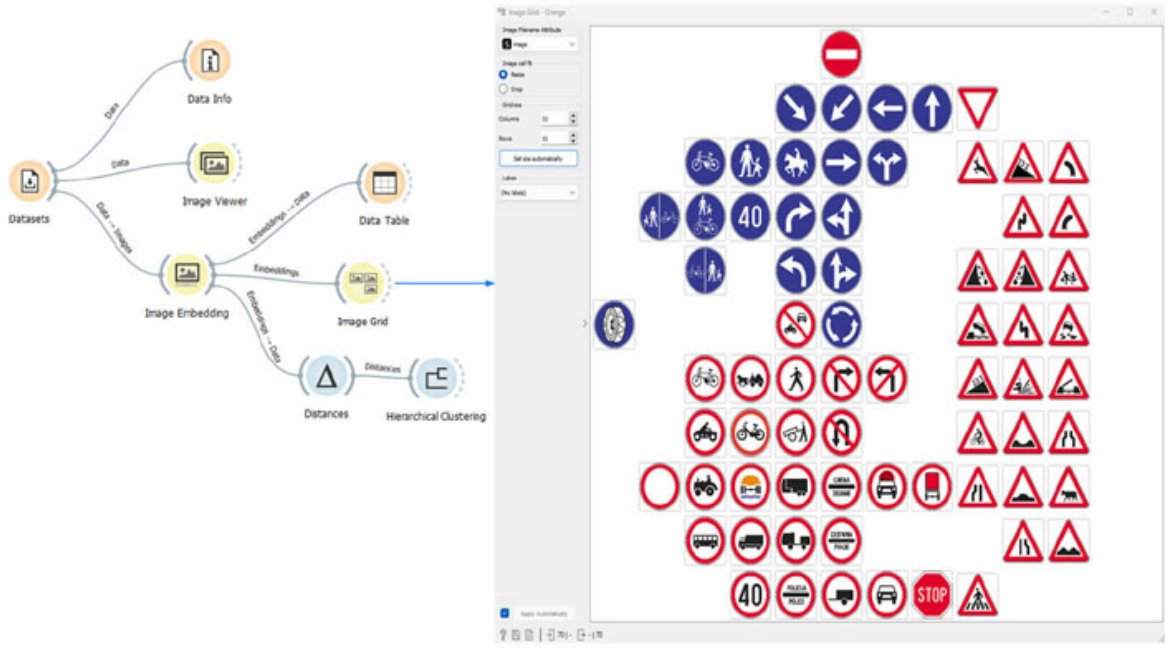


Figure 9.11: *Orange Image Analytics Overview*

Conclusion

In this chapter, we discussed the basics of image analysis and how neural networks play a key role in extracting features of an image. There are several methods to modify your models, and people are always coming up with new ideas. The topic of deep learning is one that is constantly evolving, and currently, there are no foolproof approaches. We need to do a lot of testing, and sufficient amounts of both trial and error may lead to breakthroughs. Auto machine learning tools like Orange make image analysis easy and allow anyone to perform image classification tasks with just a few clicks.

In the next chapter, we will be discussing some tips and tricks on data handling, exploratory data analysis, and how to present the data to tell the data story, along with some cheat sheets.

Points to remember

- In a neural network, neurons are arranged in layers, which, in turn, are arranged sequentially
- Each neuron in the input layer is responsible for one feature. The features can be pixels in an image, or any other input data that the neural network is designed to process.

- The number of neurons in the input layer is determined by the number of features in the input data. Images are represented as 2D grid numbers represented by the pixel values, which vary between 0 and 255

Multiple choice questions

1. What are some of the popular neural network architectures?
    a. ANN
    b. CNN
    c. RNN
    d. All of the above
2. A perceptron cannot have multiple layers.
    a. True
    b. False
3. Image rotation is an augmentation technique.
    a. True
    b. False
4. When developing object detection and image segmentation models, common data issues include which of the following?
    a. Dimensions and aspect ratios of images
    b. Identifying the image quality and edges
    c. Outliers
    d. None of the above

Answers

Question Number	Answer
1	D
2	B
3	A
4	A

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 10

Tips and Tricks

> *Data, although powerful, may also be misleading. You must be wary of data-related misconceptions and pitfalls to tell a compelling and true data story.*

There are several benefits to using big data in business. Some data formats are more convenient for analysis than others. Because of this, it might be easy to ignore facts or force them to conform to our preconceived views. Decisions may be inclined to rely solely on aggregate data analysis as access to more information and rapid advancements in techniques for evaluating large amounts of data increase. But blind trust in statistics might alter one's perception of the world. Although the phrase "ecological fallacy" was first used more than 50 years ago, it is more relevant now than ever before due to the prevalence of large data, aggregate analysis, and rapid decision-making. Ecological fallacy refers to the logical error that occurs when one draws false conclusions about specific individuals based on general trends or patterns.

Structure

In this chapter, we will cover the following topics:

- Data handling tips
- EDA tips
- Data presentation tips
- Cheat sheets to make business problem-solving easier

Objectives

You'll learn why understanding the details of data and data storytelling is important. Some people find it challenging to understand analytics. Unfortunately, analytics is not a strength of the average person. No matter

how much time and effort you put into data analysis, you still won't have the evidence you need to convince others. Keep to the narrative form; everyone can understand a narrative. You may tell a tale as part of a presentation. You may use the narrative to highlight and explain all the key aspects of the analytics you're presenting. The others listening to your presentation will find it fascinating as well.

If you want to tell a narrative using data, you need to follow three stages. The data itself must be comprehended first. The next step is to decide what kind of tale you want to tell with the information. Then, you must relate the tale. Imagine the plethora of uses of this information. A thorough examination of the data might help you grasp its implications for the company. You may even consider discussing the topic in a blog post. A set of slides detailing your investigation might be made. You may make compelling charts displaying the percentage of individuals in your field who have tried X, and your insight might be the basis for a compelling narrative. It's possible to estimate how many customers will buy from you. Before getting into the details, you must understand the basic data science life cycle as this will give you an idea of what areas need attention. Let us term all the topics that will be covered in this chapter as "*Far Side of Data Science*". The reason is that such topics will not be voluntarily discussed by any subject matter expert or courses or tutorial and so, will not be readily available as a curated content in a single place. *Figure 10.1* explains the sequential steps involved in every data science and machine learning project.

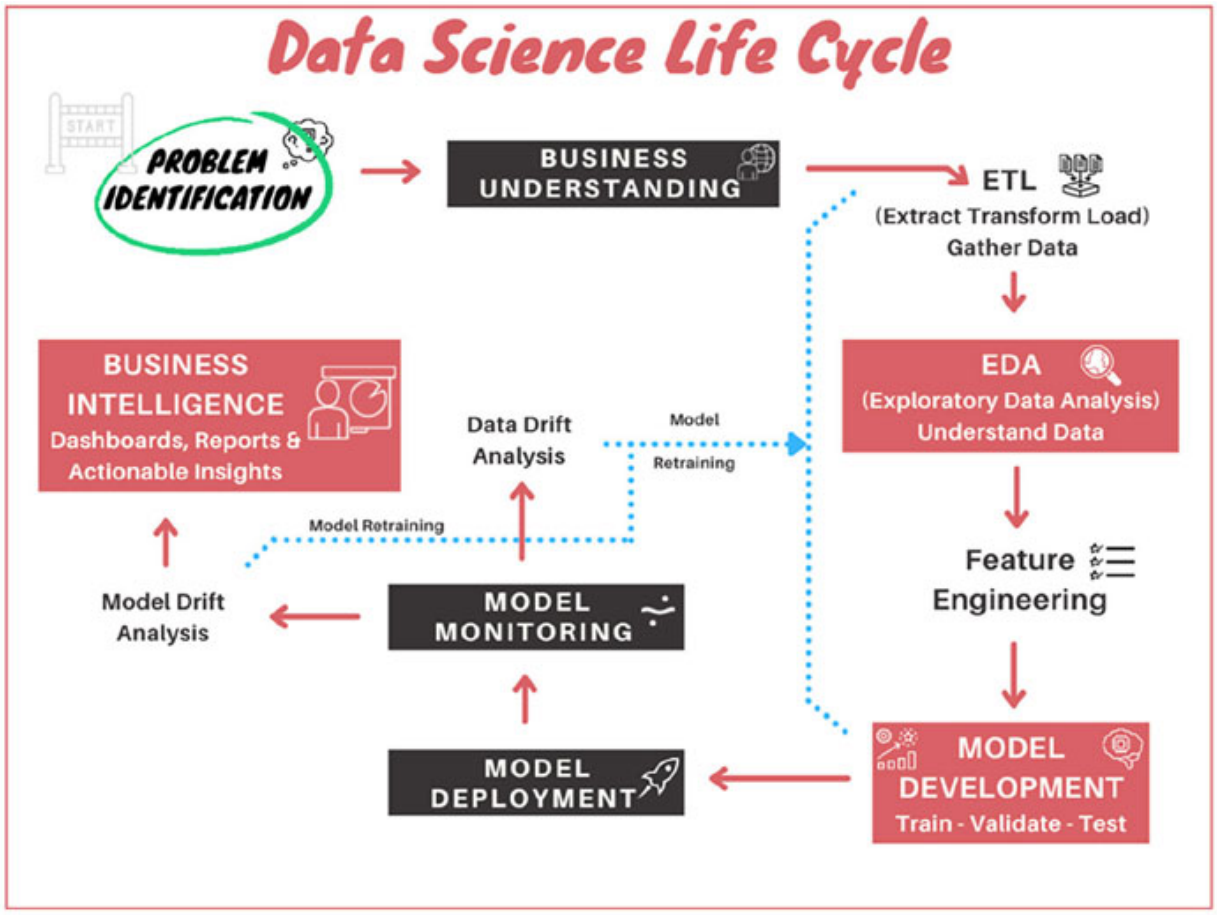


Figure 10.1: *Data Science Lifecycle*

Data management tips

Having access to data is essential for the work of data engineers and data scientists. Data analytics, data mining, building machine learning models, etc., are all tasks that fall within the purview of the various members of the team.

These team members need to collaborate in order to identify pertinent data sources, establish connections between those sources, and ensure that their efforts are aligned with the challenges confronting the organization.. The term "*data management*" or "*data handling*" refers to the process of organizing, deploying, developing, and governing an organization's data. It encompasses the entire data life cycle, from inception to deletion. With respect to machine learning, it can be broadly grouped into following two categories:

- Data Governance
- Data Fallacies

Data Governance

One challenge that you face as you move forward with your data strategy is the fact that there is so much data around. It is very easy to have more data than you know what to do with. Companies are spending millions and billions of dollars a year just to create and manage all the data they have. The issue at hand is that there are limited resources in terms of time and money to process and analyze the data available. It's evident that companies' methods of storing and managing data are evolving, leading to the emergence of new data governance systems. With this new kind of data governance system, there are more ways than ever to have data, and data is everywhere.

Data governance (**DG**) is a process used by organizations to manage and maintain data, and to ensure the availability, integrity, and consistency of data across their organization. It provides organizations with a framework to control and manage data across their organization, helping them protect the data assets; identify data quality issues; and provide accurate, timely and secure access to data. It can be implemented to create an effective strategy for managing data in an organization, ensuring compliance with regulations, and improving the efficiency and accuracy of processes and business activities. *Figure 10.2* gives an overview of various topics that will be covered under data governance: **data integration (DI), data security (DS), data quality (DQ)** and **master data management** (**MDM**). The best practices for success in this context include having a clear and well-defined objective with an end-goal in mind. Additionally, it is important to start with cultural change by bringing people together and defining the appropriate lead and lag measures of success.. Avoid making data governance overly bureaucratic, and ensure constant communication, as it is the key to success. Following this will make sure the machine learning projects does not end up being shelved and will help you focus on the right issues are the right time.

Organizations cannot effectively evaluate data, make decisions, and act quickly without the use of data integration. Using data integration technologies, you should be able to analyse data and create reports. The data

can be accessed in various formats, thanks to modern data integration tools and platforms offered by *Microsoft*, *Amazon*, *Google*, *Oracle*, and so on. Your data can be retrieved from any device with their help. Information from various sources can be combined into a single schema for the convenience of analysis. Moreover, you can merge data from many files, establish relationship between them, and examine similarities and differences between the data. Before proceeding with machine learning use cases, the first logical step would be to establish a data integration platform where all critical data is available for the data scientists and machine learning engineers to play with. *Figure 10.2* will give a rough idea about the high-level architecture behind data integration platforms:

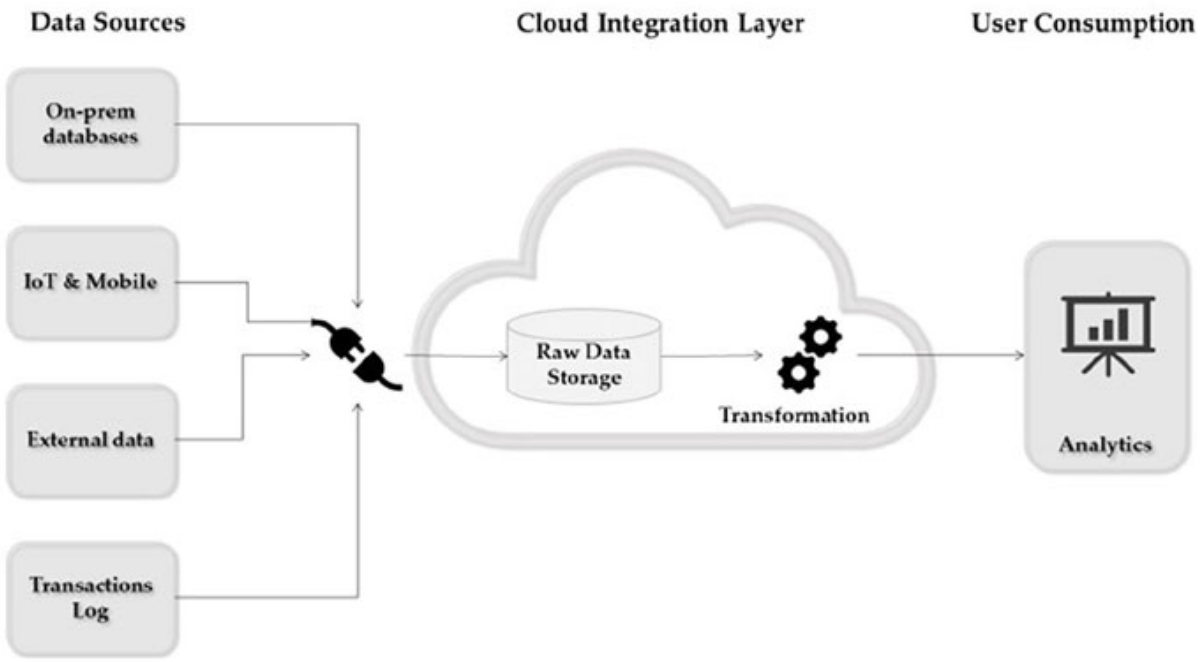


Figure 10.2: *A Simple Data Integration Architecture*

Information confidentiality is maintained using data security measures. Data security encompasses all the measures used to protect sensitive information from unauthorized access or disclosure. A breach in data security is a severe problem that may have far-reaching effects on your business if it occurs. In machine learning, you have to be careful when dealing with personal and sensitive data. There is a thin line when it comes to using the data ethically; always keep this in mind. It is worthwhile to check the data guidelines of the

country or region when you work on projects covering multiple geographical locations.

Companies in today's data-driven era need to prioritize data quality to be successful. Data quality is a metric used to assess how adequate and helpful certain data sets are for various use cases. Organizational data is formed via **master data management** (**MDM**), which includes the development and maintenance of MDM procedures, standards, governance, and tools. MDM is the practice of centralizing an organization's most important data in a single location to prevent inconsistencies and duplications. In machine learning projects, always ensure that the master data is accurate and recent. Among various aspects, data quality has to be measured in terms of consistency, completeness, accuracy, origin and timeliness. *Figure 10.3* will give a perspective on data bare minimum quality measures:

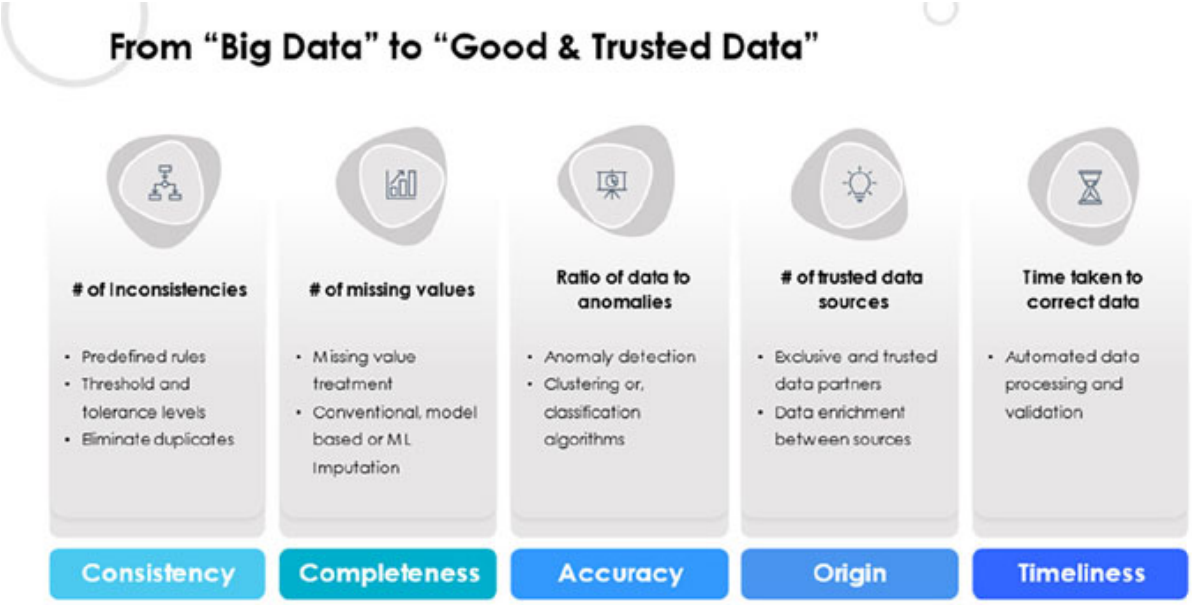
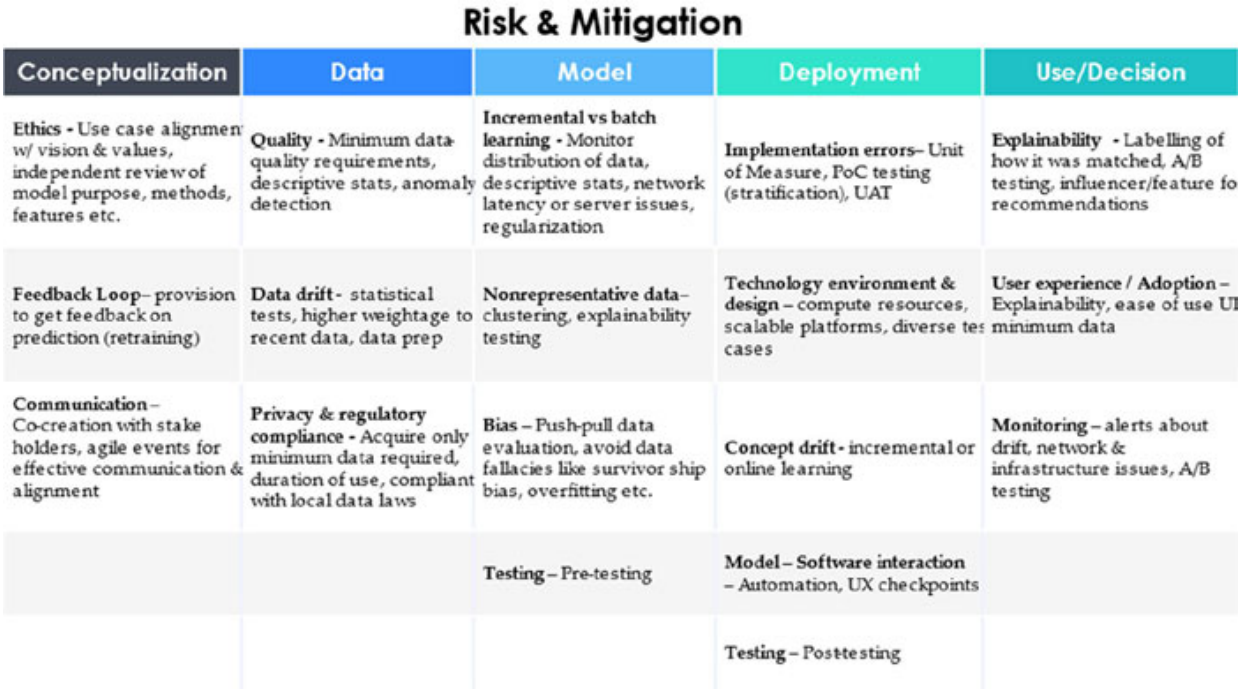


Figure 10.3: *Data Quality Measures & Best Practices*

As with any applications, machine learning applications face risks; hence, having a control strategy would ensure that the process is intact. Rather than acting as a barrier to innovation, an organization's ability to effectively manage risks is essential to the smooth integration of AI into daily operations. After multiple trials, we found that applying sig sigma principles for mitigating risks is highly beneficial. These risk factors must be analysed at various stages in the data science life cycle. Based on the frameworks and

white papers released by McKinsey and Deloitte, we have put together a basic list of potential risks and control matrix covering the end-to-end process of machine learning models. In the following figure, you can find that this process starts with conceptualising a business problem and moves on to using a fully functional machine learning model in a production environment to make informed business decisions. You can use the table in *Figure 10.4* as a handout to remind you whenever you start a machine learning project:



Risk & Mitigation

Conceptualization	Data	Model	Deployment	Use/Decision
Ethics - Use case alignmen w/ vision & values, independent review of model purpose, methods, features etc.	**Quality -** Minimum data quality requirements, descriptive stats, anomaly detection	**Incremental vs batch learning -** Monitor distribution of data, descriptive stats, network latency or server issues, regularization	**Implementation errors–** Unit of Measure, PoC testing (stratification), UAT	**Explainability -** Labelling of how it was matched, A/B testing, influencer/feature fo recommendations
Feedback Loop– provision to get feedback on prediction (retraining)	**Data drift -** statistical tests, higher weightage to recent data, data prep	**Nonrepresentative data–** clustering, explainability testing	**Technology environment & design –** compute resources, scalable platforms, diverse tes cases	**User experience / Adoption –** Explainability, ease of use UI minimum data
Communication – Co-creation with stake holders, agile events for effective communication & alignment	**Privacy & regulatory compliance -** Acquire only minimum data required, duration of use, compliant with local data laws	**Bias –** Push-pull data evaluation, avoid data fallacies like survivor ship bias, overfitting etc.	**Concept drift -** incremental or online learning	**Monitoring –** alerts about drift, network & infrastructure issues, A/B testing
		Testing – Pre-testing	**Model – Software interaction –** Automation, UX checkpoints	
			Testing – Post-testing	

Figure 10.4: *Basic Risk and Mitigation Handout for ML Projects*

Data Fallacies

Let us look at some of the challenges and look out areas that must be on top of your mind while working with data. It's common to assume that the value of data will only become apparent after it has been collected. Let's be clear: data is not the same thing as insights. There are several opportunities for missing the narrative that data is trying to convey, starting with the choice of time period used and continuing through the choice of reliable data sources. However, it is necessary to evaluate the available data and give it meaning to arrive at reliable findings. Data fallacies are pitfalls that might be encountered while attempting to draw conclusions from numerical data. The

term “*data fallacy*” refers to the practice of using flawed statistical methods to support an argument. Geckoboard, a software company that helps organisations adopt data culture, has put together a detailed list of data fallacies. You can refer to the complete list in their blog. However, we will look at some of the most important ones, which should be at the top in your mind:

- **Sampling bias** is perhaps the most prevalent kind of statistical error. It entails drawing conclusions about a population based on an insufficiently representative sample. Let us take the popular example of opinion polls, usually conducted during the times of a general election. InstaVaani’s polls for the Delhi assembly elections in 2015 looked terribly inaccurate since they were based on samples with significant bias. The issue was not with the sample size. Sample size of 800 used in InstaVaani’s polls should be sufficient, with a confidence interval of +/- 5%, to accurately predict how voters would split their ballots. The issue arose from sloppy sample selection. To avoid this, be certain that your data sample is a true reflection of the population as a whole so that any inferences you draw from it may be extrapolated to the larger population.
- **Data dredging** is when researchers ignore the possibility that a correlation between two variables occurred by coincidence. A number of tests are conducted, and findings are discussed only for the tests whose results are fascinating. Simply put, it’s the practice of drawing conclusions from a dataset when none exist. There can be no statistical significance tests without a well-defined hypothesis. It is necessary to first hypothesize before doing any kind of experiment. You shouldn’t build and test your hypothesis using the same facts.
- **Survivorship bias** is when using a subset of data to draw broad conclusions because it has “survived” a series of predetermined filters, data pre-processing techniques, etc. Let us assume that you want to start your own business venture. The first thing most of us read or research is about people or companies who have succeeded in their start-up journey. The failures of many start-ups are forgotten, while the achievements of the few are celebrated for decades. Companies who fail to make it to the finish line are often ignored, because of which the

odds of success may be overestimated. Much of what we think we know about business was similarly shaped by the victors in the past. Those who experience failure and humiliation have no meaningful voice. Those who fail to recognize the part that luck plays in the rise of the prosperous are likely to be held responsible for the misfortunes that befell them. Be careful not to extrapolate from data that has survived a screening procedure.

- **Cherry picking** is the most destructive kind of cheating in machine learning and the easiest to avoid. It's picking and discarding results that match your claim. If you have a theory that a certain action is good for your business, you might consider the data that demonstrates that the theory is true and end up ignoring all the data that contradicts it. You shouldn't do this though, because it would mean that you are not thinking critically. In the end, you might realize that your theory was wrong. While collecting data for a use case, trace whether the data was modified in any way, what pre-processing steps were used during the data transformation stage in your data warehouse or data lake, etc.
- Francis Anscombe, a statistician, created **Anscombe's Quartet** to demonstrate why it's not enough to depend just on statistical metrics when examining data. He accomplished this by constructing four data sets that would provide essentially similar statistical metrics. When plotted, though, it was evident that the datasets are completely different from one another. Graphs should be used in conjunction with statistics or summary metrics wherever possible. Visuals give a whole new meaning to the underlying data.

EDA Tips

Even before the actual data modelling begins, exploratory data analysis is clearly one of the most crucial steps in the data to insights process. Since it is a crucial step, having the right talent and skills in performing this step decides the fate of the overall project or application. Having thorough knowledge and skills of exploratory literature review, pattern recognition, and visualization will be a key differentiator. The best practice for effective EDA is as follows:

- Data observation
- Missing value and outlier treatment
- Correlation analysis

Data observation

Here, check the data for shape, size, duplicates, type, and distribution. This will help formulating and choosing the right data modelling techniques. The shape and size of the data tells you what you are dealing with. Then, check the various types of data types available as it will help determine the subsequent steps of data processing strategy. Identifying the number of duplicate entries in the data will help understand true shape and size of the data. Also, it will help you narrow down to the relevant data. The size of the dataset can inform your selection of potential models. When working with a limited dataset, for example, a neural network may produce an overfit model whose performance is unpredictable in real-time scenarios.Finally, visualize the data distribution because summary metrics will not always tell the true picture, as described earlier in Anscombe's quartet.

Missing value and outlier treatment

Good data artists know why some data isn't available, figure out how much data is missing, fill in the gaps with the right strategy, and avoid making common mistakes. We don't like it when there are inconsistencies in the data sets because that means we can't use the raw data to train our machine learning algorithms or derive actionable insights to make business decisions. There might be various reasons for missing data, like failure of the data gathering devices or tools, business rules and architecture behind data aggregation platforms, anonymizing data due to security reasons, typos or data intentionally not filled in case of surveys, forms, or data not recorded intentionally or by mistake. The three types of missing data are **missing completely at random** (**MCAR**), **missing at random** (**MAR**) and **missing not at random** (**MNAR**). If the chances of missing data is same throughout, then it is MCAR, but if it is the same only in certain groups or a cluster of data, then it is MAR. Missing data that does not fall under any of these categories will fall under MNAR. Broadly, missing value treatment can be classified into three groups: conventional, model-based and imputation. If

the missing values are less in percentage and absolute value, then dropping those data points would be fruitful. Based on the presence of outliers in the data, the mean or median might be used to replace the missing values. Mean can be used if data distribution is normal and median if there is presence of outliers. Categorical data can be imputed using mode value, i.e., the most frequently occurring data. Machine learning algorithms can also be used for treating missing values as an advanced form of treatment method. The following figure will give you an idea of the various methods for treating missing values. You can Google to get the detailed explanation of each method mentioned in *Figure 10.5*:

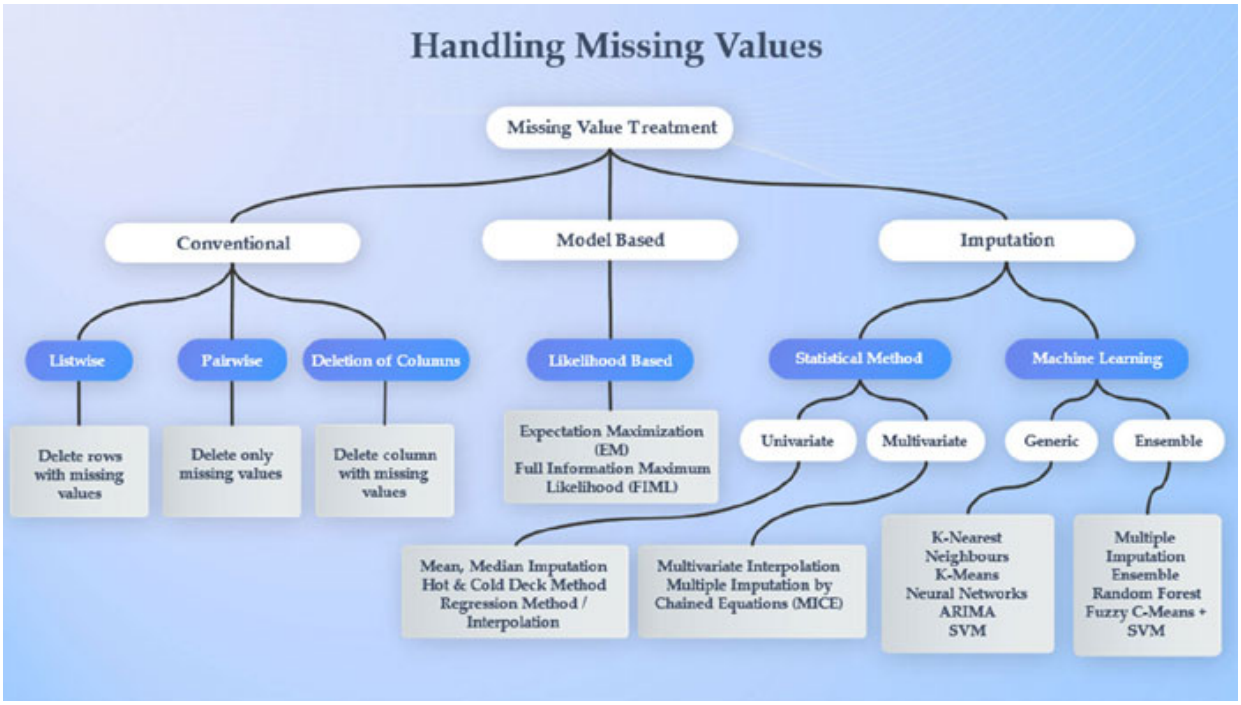


Figure 10.5: *Missing Value Treatment Techniques*

Data that significantly deviates from the normal distribution is generally referred to as an outlier. Box charts are a great tool to detect outliers. The outliers are sometimes genuine and represent a true condition. Outlier treatment is always a trade-off between the bias of our estimations if we maintain “poor” data and the loss of accuracy if we toss away “good” observations. Therefore, it is crucial to determine whether an outlying number is good or negative and evaluate its effect on the statistical metrics. The **Inter Quartile Range** (**IQR**) method can be used to identify outliers by drawing an upper (quartile 3, Q3) and lower (quartile 1, Q1) boundary. IQR

can be calculated by taking the difference of Q3 and Q1. Any value that falls outside this boundary can be considered an outlier. Typically, we would take 1.5 times the IQR, minus Q1, and add Q3, to establish this border. There are several methods for treating outliers, including deleting the outlier (if there are few observations), using data imputation methods described earlier, implementing quartile-based capping (where all values below the 95th or 90th percentile value are considered), or using machine learning models to impute outliers as a more advanced option. The following figure will help you understand IQR and outliers in a box plot:

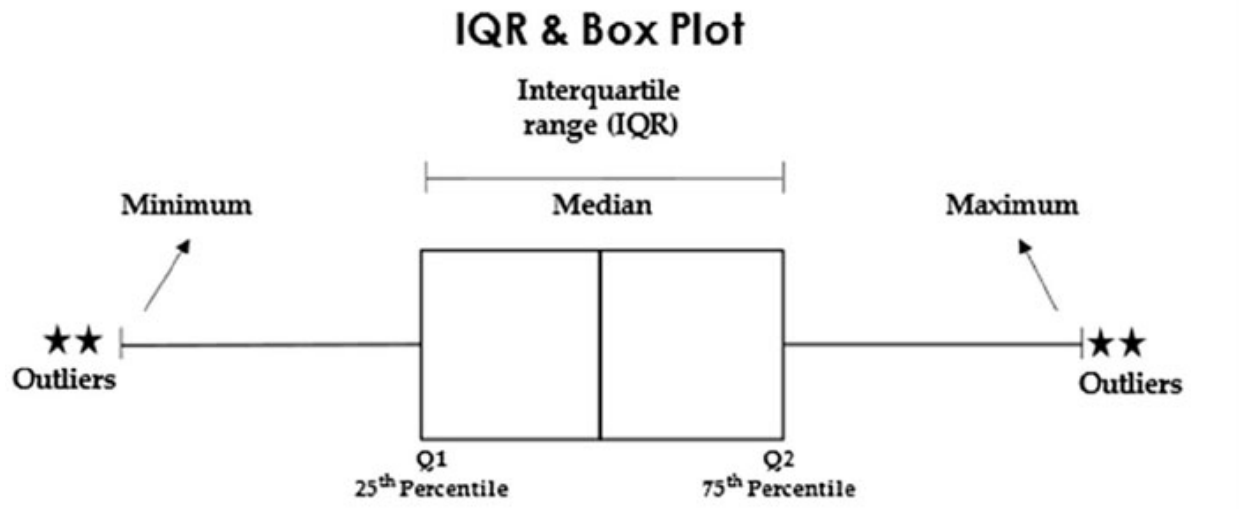


Figure 10.6: *Interquartile Range*

Correlation Analysis

The Pearson and Spearman correlation coefficients are relatively similar; however, the Kendall correlation coefficient is very different. This is due to the fact that Pearson and Spearman are virtually similar in their ability to correlate normally distributed data, but Kendall is an indicator of the strength of dependence. The outcome of all of these correlations is the same; the only difference is that Pearson and Spearman analyse the data one way, while Kendall analyses it another way. When the data is ordinal, it is easier to use spearman or Kendall (but not Pearson) since the rankings are established beforehand. The difference between the Spearman and Pearson correlation coefficients when outliers are eliminated is an important factor to consider when choosing the appropriate statistic. Due to the possibility of an outlier impact, it is recommended to use the Spearman's correlation

coefficient when working with data that does not have a normal distribution. While engineering new features or attributes from existing data using business knowledge, ensure that features are not redundant. For example, you can calculate **body mass index** (**BMI**) if you have height and weight data. However, while analysing correlation co-efficient, you might see all these to be closely related, making it redundant. In such cases, you can use BMI and drop others if needed and avoid increasing complexity of the model. The correlation coefficient is used to describe a mathematical association in which two or more variables are linked, but that does not imply causation in any way. Sometimes correlation can be spurious.

In conclusion, correlation coefficients are helpful in determining the direction and magnitude of linear correlations between variables. When comparing Pearson and Spearman correlation coefficients, the latter is more robust to outlier values.

Data presentation tips

What is the first thing that comes to mind when you hear the phrase “data analysis?” Perhaps you immediately think of “software skills” of data analysis, such as searching through data tables, Excel spreadsheets, business warehouses and statistical calculations. However, technical proficiency is meaningless without the same level of soft skills. Data analysis alone isn’t enough; you also need to master the art of data storytelling to effectively convey the insights you’ve gained. Although data visualization is an integral part of data storytelling, the two are not synonymous. It’s an engaging story that’s built on and supported by convincing data. Data storytelling involves telling the story of the past (descriptive), present (diagnostic), or future (predictive). The success is based on comprehension, retention, and appeal. The sweet spot for data storytelling lies between data insights, audience need, and compelling narrative, as shown in *Figure 10.7*:

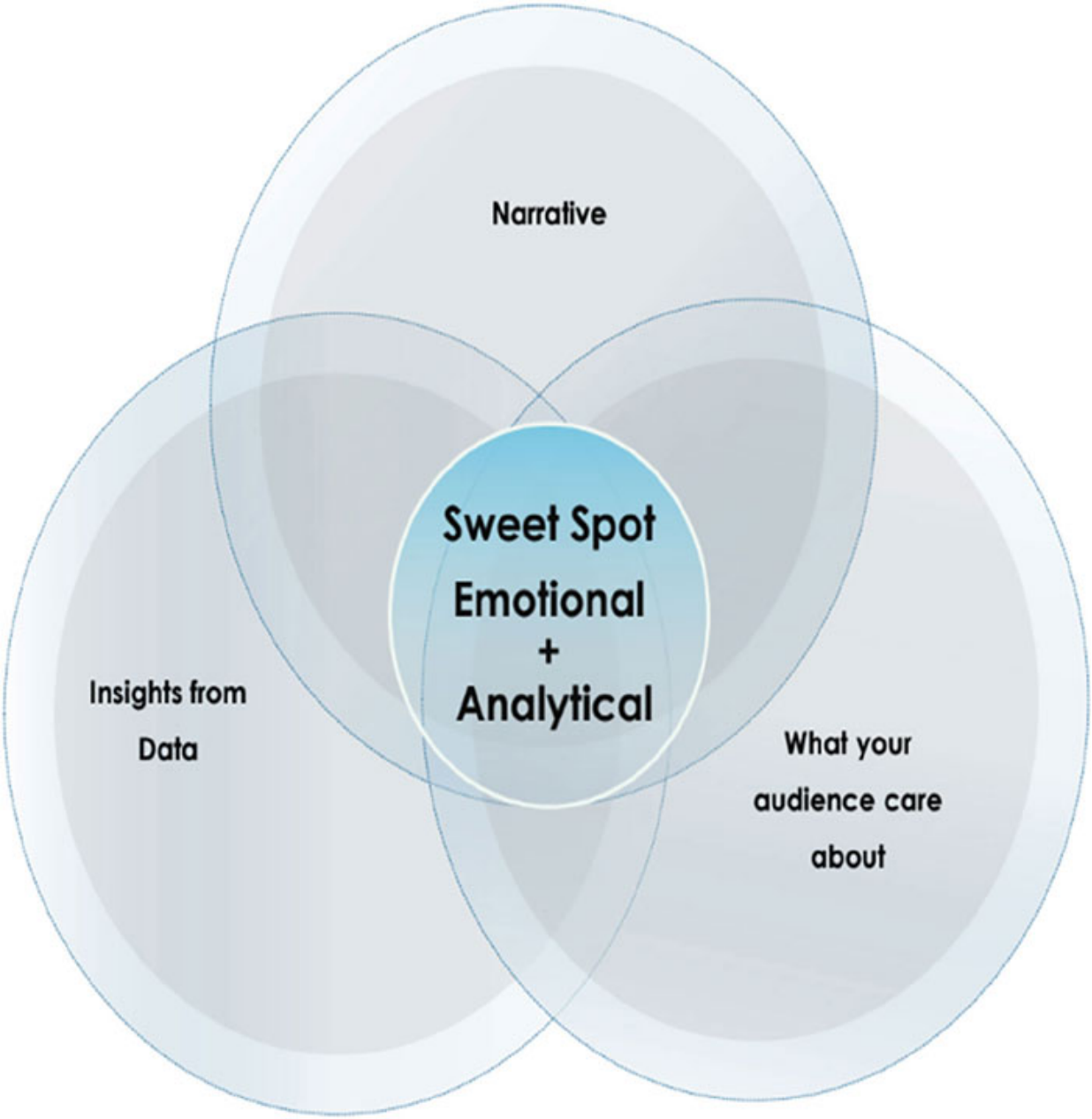


Figure 10.7: *Data Storytelling Sweet Spot*

Data is nothing but numbers and characters until it is transformed to a story. Data-driven narratives provide depth to any topic. Telling a great data-driven story drives better decision-making. Our secret formula for data storytelling can be seen in *Figure 10.8*:

Figure 10.8: *Data Storytelling Formula*

There are five key components to data storytelling framework: context, audience, visual, focus and tell a story. Organisations are increasingly prioritizing candidates who can tell a compelling story with data, including it as a requirement for analysts and creating new roles for "data storytellers" to complement existing analytics teams. Refer to *Figure 10.9*:

Figure 10.9: *Data Storytelling Framework*

Context

Define and narrow down the problem you are trying to solve. When information is delivered to readers, context is the part of the story that explains the significance of the data and why the audience should care; it will drive engagement.

Audience

Your audience is the next focal point. Context can be understood only if you can actively listen to your audience's need. Data communication begins with identifying the audience and learning their preferred mode of interaction. To guarantee the strategy's success, it's important to think about factors like who will be making the decisions and what will drive them. If you know what your audience wants to know or what they want to do with the knowledge you provide, you can confidently and succinctly answer the "so what?" question they're sure to ask. You cannot tell the same story to a subject expert and to an amateur. The level of technicalities must vary. Executive leaders will have a different need from that of a front-end employee. Hence, audience is key, and humanizing the data as per audience need is important.

Visual

Scientific studies say that people are more likely to remember information if it is accompanied with relevant visuals. Therefore, if you want to get people's attention, use an effective yet basic visual. Also, including appropriate measures in visuals, like colour, text, transition, target in case of a metric, etc., has to be carefully evaluated. Some of the best practices for compelling visuals are as follows. Always keep it simple; don't clutter it by including irrelevant information. Think about summarizing to eliminate unnecessary information and emphasize the most important points. Always follow a logical sequence by prioritising the information and how the data should be digested by your audience. Annotations can be used, but choose your words wisely. One of the worst possible outcomes would be if two individuals arrived at completely opposite conclusions after looking at the same graph. Hence, always seek feedback from a friend or colleague and follow the 30-second thumb rule: if anyone cannot understand the message

you want to communicate through the visual in the first 30 seconds without an explanation, then you might need to reconsider your visual design.

Focus

A story will have a moral, and the same applies for a data story. Draw the audience's attention and call for an action. A simple change in the visual presentation improves focus. Visual cues help process information faster; they not only help with numbers but also patterns. Patterns help identify correlation and association. The power of data storytelling and design help communicate all sorts of information. You can adjust the orientation, size, shape, colour and animate to draw attention, but always remember not to use everything at the same time. It goes with the proverb "*Too much of anything is good for nothing*".

Tell a story

To explain your data, speak as usual in normal tone but tell a story that your audience can easily relate to with emotions. Make sure there is a beginning and end to your story, along with a mid-section. Think and practise the flow of your story and ensure that it is in sync with your audience. For example, most executive or C suite leaders always like to see the summary of your story first, while subject matter experts and engineers like to go step by step to the conclusion. Conflict and tension are excellent attention-grabbers. To summarize, companies that rely heavily on data to make business decisions are upgrading their data storytelling techniques to motivate employees to act. In order to create a compelling data story, it is important to start with the context by understanding your audience and then use simple visuals to call for action; also, you must ensure that there is a logical flow to the story.

Machine learning cheat sheet

The following are three flow charts cum cheat sheet for selecting the right approach when it comes to classic machine learning techniques. The first figure acts as a simple guide to choosing the right regression algorithm. Refer to *Figure 10.10*:

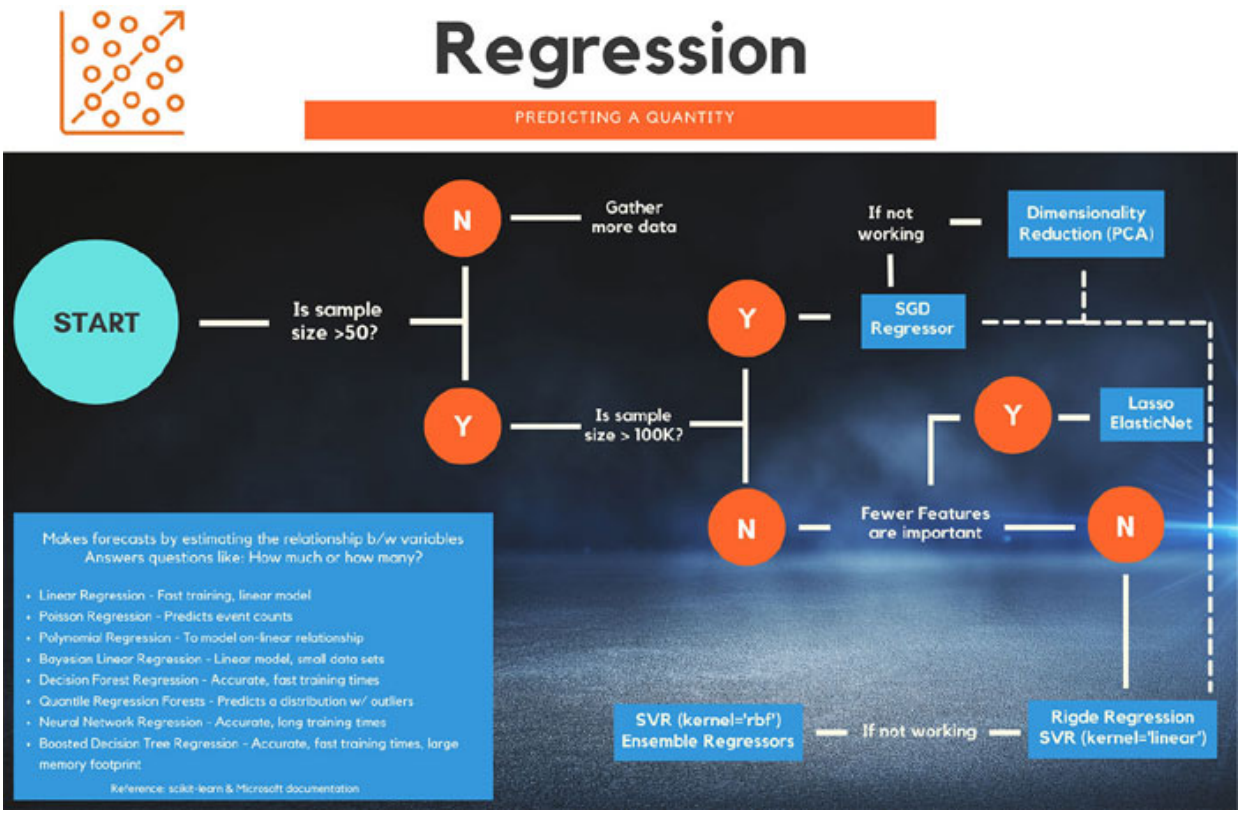


Figure 10.10: *Regression Cheat Sheet*

As discussed in *Chapter 6, Classification*, the purpose of classification is to divide things into smaller groups. The next figure will help you choose between various classification algorithms:

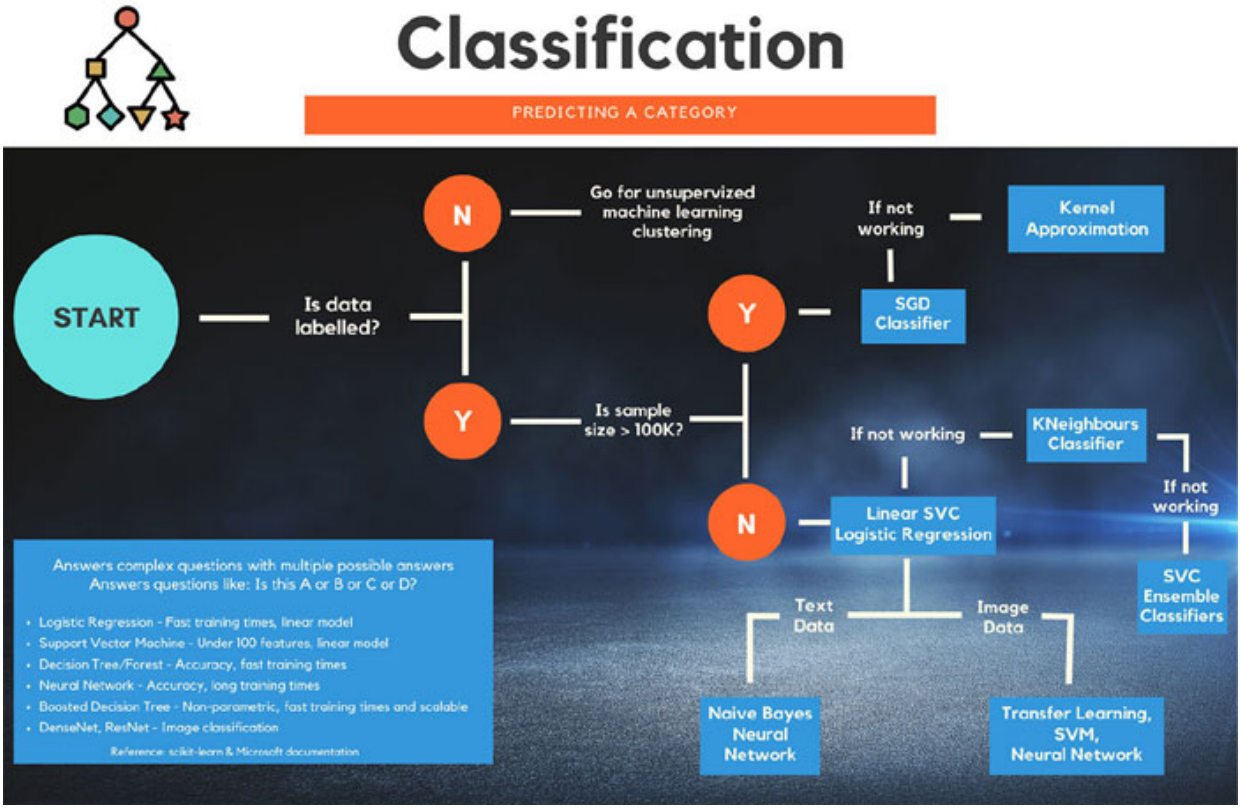


Figure 10.11: *Classification Cheat Sheet*

The next and last figure is about unsupervised learning technique of clustering:

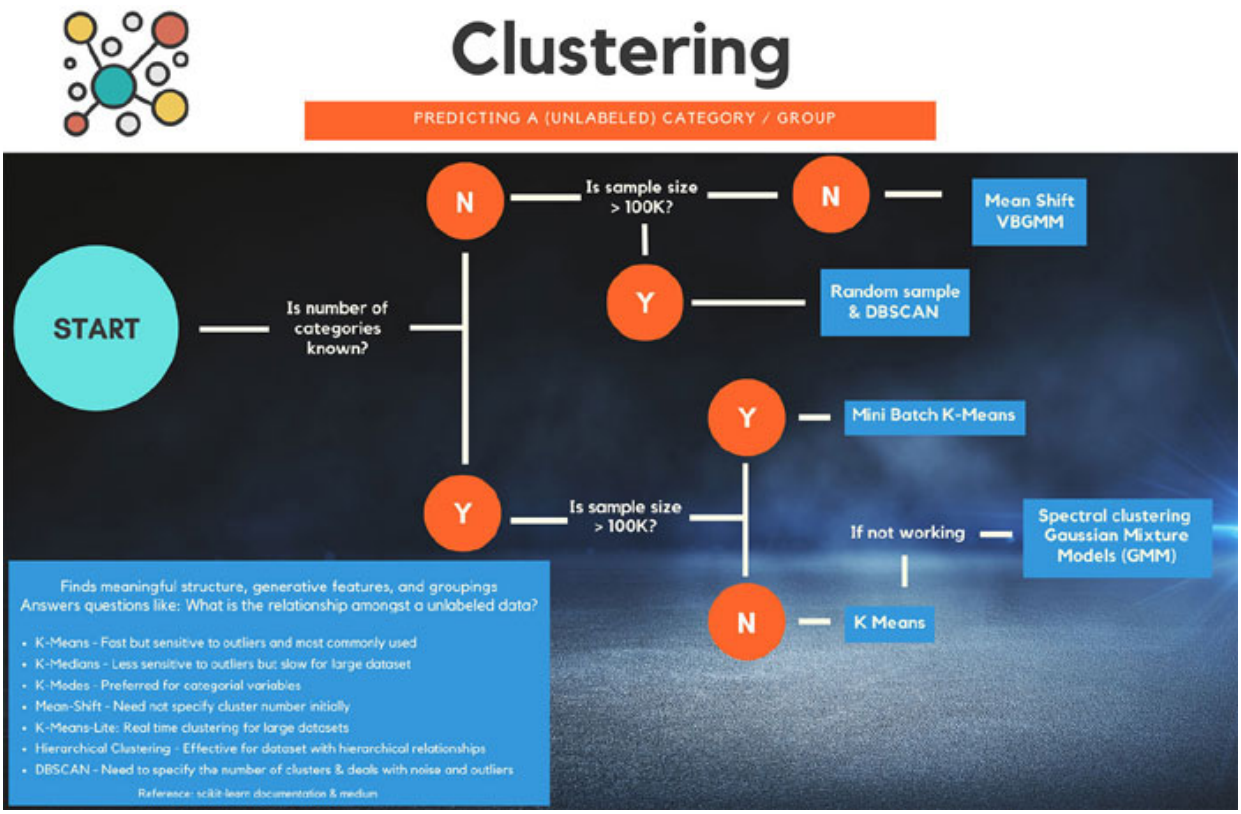


Figure 10.12: *Clustering Cheat Sheet*

Conclusion

In this chapter, we discussed some of the best practices for building a machine learning app. We zeroed in on the business and technical ends, and we learned to set and achieve objectives. We also shared some advice and recommendations on the management of code and infrastructure. Finally, we covered some of the data and model choices that may be made.

Points to remember

- Good and trusted data is better than big data.
- To start, stick with a basic model and focus on getting the infrastructure right.
- It is always a fruitful exercise to include clustering as an EDA technique as it will help uncover the hidden data clusters.

- Data story carries equal weightage with machine learning or data modelling technique.
- Always test your model before exporting it to real-time environment.

Multiple choice questions

1. What comes after model development in data science life cycle?
    a. Model deployment
    b. Business intelligence
    c. Model monitoring
    d. None of the above
2. _____ is drawing conclusions about a population based on an insufficiently representative sample.
    a. Survivorship bias
    b. Data dredging
    c. Sampling bias
    d. Cherry picking
3. Pearson correlation coefficient is robust and not subject to outliers.
    a. True
    b. False
4. Which of the following is the data storytelling formula?
    a. (Context + Focus + Visual) * Narrative
    b. (Narrative + Audience + Visual) * Context
    c. (Context + Audience + Visual) * Narrative
    d. None of the above

Answers

Question Number	Answer
1	A
2	C

3	B
4	C

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Index

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W