MINISTRY OF SCİNECE AND EDUCATION OF THE REPUBLIC OF AZERBAIJAN

AZERBAIJAN STATE UNIVERSITY OF ECONOMICS

UNEC BUSINESS SCHOOL

MASTER THESIS

on the topic:

"Application of Data Science in Decision Making"

Code of specialization and name:	**060409 Business Administration**
Specialization:	**Business and Data analytics**
Group:	**E15-20**

Master:	**Scientific Supervisor:**
Alasgarzade Kamran Yaqub	**Ph.D. in economics**
	Shafizada Elnura Rafiq
_____	_____
Program supervisor:	**Head of department**
Ph.D. in economics	**Ph.D in economics, Associate prof.**
Ghazanfarli Mirvari Khagani	**Mammadova Sevar Momin**
_____	_____

BAKU – 2023

Summary

The relevance of the dissertation: Data science provides a systematic approach to gather, analyse, and interpret large volumes of data to extract meaningful information and patterns. This information can then be used to support decision-making processes, optimize business strategies, and drive operational efficiency. By applying data science techniques, decision makers can gain a deeper understanding of their business environment, identify trends and correlations, and make more informed and evidence-based decisions.

The purpose of the dissertation. The main goal of the dissertation work is how to obtain accurate and useful information from historical data to be used in decision-making and to verify the correctness of the decisions made by Machine Learning models.

Research methods used in the dissertation: Statistical analysis and Machine Learning were used in the dissertation work. The Logistic Regression model was developed in the Python software.

The database of the dissertation: Freely accessible dataset from Amazon Customer Dataset which introduces real dataset can be downloaded using Amazon Website.

The limitation of the dissertation: Building highly accurate predictive models requires a great number of features from various datasets of news, tweets, discord, and fundamental and technical features which are not publicly available or need perfect resources to download and generate features.

Practical results of dissertation work: After dividing the Amazon data into Train and Test, applied the Logistic regression model on the Train set. Then checked how correct the decisions made on the Test set were according to the Logistic Regression. In the obtained results, the accuracy of the model is 91%. Based on the model, future decisions will be made with 91% accuracy.

Practical results of dissertation work where we may use: The practical results of dissertation work can be applied in various ways. Provide insights to improve decision-making processes, develop decision support systems, assess and manage risks, address cognitive biases, enhance organizational decision-making, and promote data-driven decision-making. Research findings have practical implications for improving decision-making practices and outcomes across different domains.

Keywords: Machine learning, deep learning, logistic regression, quantitative analysis, Naive Bayes, random forest

List of abbreviations

Abbreviation	Explanation
ML	Machine Learning
AI	Artificial Intelligence
TP	True positive
TN	True Negative
FP	False Positive
FN	False Negative
RF	Random Forest
ANN	Artificial Neural Networks

Table of contents

Introduction

The relevance of the dissertation: In today's rapidly evolving digital landscape, organizations face a multitude of challenges when it comes to decision making. The abundance of data generated by various sources, such as customer interactions, social media, sensors, and transactional records, presents both opportunities and complexities. The dissertation explores how data science can address these challenges and improve decision making across industries.

By applying data science techniques, organizations can unlock the hidden insights within large and complex datasets. The analysis of data using statistical models, machine learning algorithms, and data visualization tools empowers decision makers to uncover patterns, trends, and correlations that would be difficult to detect through traditional methods. These insights provide a solid foundation for making informed and evidence-based decisions.

The dissertation lies in the transformative potential of data science in decision making. By incorporating advanced analytics, organizations can optimize their operations, enhance efficiency, and drive innovation. For example, in finance, data science can help identify potential investment opportunities, manage risks, and improve portfolio management strategies. In healthcare, data-driven decision making can support personalized treatments, early disease detection, and resource allocation optimization.

Moreover, the dissertation highlights the importance of real-time decision making. With the speed at which data is generated, organizations need to adapt quickly to changing circumstances. Data science techniques enable organizations to process and analyse data in real-time, providing timely insights that inform immediate actions. This agility is critical in domains such as supply chain management, fraud detection, and cybersecurity, where swift and accurate decisions can have a significant impact. By extracting valuable insights from data, organizations can identify market trends, customer preferences, and competitive landscapes. These insights enable organizations to make proactive decisions,

tailor products and services to customer needs, and gain an edge over their competitors.

In the Chapter 1, Decision-making is the process of selecting the best option among alternatives to achieve goals. Factors such as problem complexity, future unpredictability, and the need for quick decisions influence the decision-making process. Decision models are used to analyse and evaluate alternatives, and the choice of model depends on the level of information and uncertainty involved in the decision. Data and big data are essential in decision making and have distinct characteristics such as being represented in binary form, stored in databases, and having high volume, diversity, and velocity. Data science encompasses various fields like analytics and machine learning, and it involves collecting, analysing, and deriving insights from data to enhance business decisions.

In the Chapter 2, the application of data science algorithms in decision-making is demonstrated using the example of Amazon data. The logistic regression model is used to determine the target audience for advertisements based on user data. Data preparation, including identifying outliers and missing values, is performed, and feature scaling is applied using the Standard Scaler method. The model's performance is evaluated using a confusion matrix, achieving an accuracy score of 91%, and cross-validation is performed to test the model's power.

Decision-making is explored as the process of selecting the best option among alternatives to achieve goals, considering factors like problem complexity and future unpredictability. Decision models are used to analyse and evaluate alternatives, adapting to the level of information and uncertainty involved. The significance of data and big data in decision-making is emphasized, with their unique characteristics and the role of data science in collecting, analysing, and deriving insights to enhance business decisions. Logistic regression is employed to determine the target audience for advertisements. The process involves data preparation, outlier identification, and feature scaling, followed by performance evaluation using a confusion matrix and achieving an accuracy score of 91%.

The purpose of the dissertation: The primary objective of dissertation work is to delve into the effective utilization of historical data to acquire accurate and valuable information for decision-making purposes. Specifically, aim to verify the correctness of decisions made by Machine Learning models. By analysing historical data patterns and relationships to intend to enhance decision-making processes and ensure the accuracy and reliability of the decisions generated by Machine Learning models.

Research methods used in the dissertation: The theoretical part of dissertation work drew upon academic journals, internet resources, and books on statistics and Machine Learning. These sources provided a solid foundation for understanding the relevant concepts and techniques. Data was collected from the Amazon website, which offers a wealth of real-world information. Using the Python programming language, built a Logistic Regression model to analyse the data. This model enables to examine the relationship between variables and make predictions or draw insights.

The database of the dissertation: The Amazon Customer Dataset, a real-world dataset introduced in the dissertation, offers a freely accessible collection of data that can be obtained through the official Amazon website. This dataset serves as a valuable resource for researchers and organizations interested in exploring and analysing customer-related information, providing an opportunity to delve into real-world data for various applications and insights.

The limitation of the dissertation: Constructing highly accurate predictive models necessitates a wide range of features sourced from diverse datasets, including news, tweets, discord, as well as fundamental and technical indicators. However, accessing these datasets and generating the required features can pose challenges as they are either not publicly available or require substantial resources for downloading and processing.

Practical results of dissertation work: After splitting the Amazon data into training and test sets, applied a Logistic Regression model to the training set.

This involved training the model on the training data, allowing it to learn the patterns and relationships within the data. Once the model was trained, evaluated its performance on the test set to assess how well it could make accurate predictions on unseen data.

The results of evaluation showed that the Logistic Regression model achieved an accuracy rate of 91%. This means that the model correctly predicted the outcome or classification for 91% of the samples in the test set. A high accuracy rate indicates that the model performed well in capturing the underlying patterns and relationships in the data, enabling it to make reasonably accurate predictions. Based on this Logistic Regression model, can now use it to make future decisions with a level of confidence. With a 91% accuracy rate, the model provides a reliable basis for decision-making. However, it's important to note that accuracy alone may not be the only metric to consider. Depending on the specific context and requirements of the decision-making process, other evaluation metrics such as precision, recall, or F1 score may also be relevant to consider the model's performance comprehensively.

It's essential to interpret the results of the Logistic Regression model in the appropriate context and consider any potential limitations or assumptions made during the modelling process. Additionally, ongoing monitoring and validation of the model's performance with new data can help ensure its continued accuracy and reliability in making future decisions.

With a precision score of 0.8966, the model correctly predicted around 89.66% of the instances it classified as positive. The recall score of 0.8125 indicates that the model accurately identified approximately 81.25% of the actual positive instances. The F1 score, which combines precision and recall, yielded a value of 0.8525. This score represents a balanced measure of the model's accuracy, taking into account both precision and recall.

These metrics collectively demonstrate the performance of the Logistic Regression model in correctly predicting positive instances, identifying actual positives, and achieving a balanced overall accuracy.

It's important to consider the specific context and requirements of your research or application when interpreting these metrics. Depending on the priorities and goals of your decision-making process, you may place more emphasis on precision, recall, or strike a balance between the two using the F1 score.

In summary, the Logistic Regression model exhibited a high accuracy rate of 91% and demonstrated a good balance between precision and recall with an F1 score of 0.8525. This indicates that the model is effective in making accurate predictions and capturing positive instances, providing a reliable basis for decision-making.

Practical results of dissertation work where we may use: Companies can leverage this dissertation work in decision-making by implementing data science techniques to improve the accuracy and effectiveness of their decision-making processes. By adopting research findings, organizations can harness the power of data analysis, predictive modelling and optimization algorithms to make data-driven decisions, optimize resource allocation, and gain a competitive advantage in their respective industries.

CHAPTER I – DECISION MAKING STRATEGIES, MODELS AND FACTORS INFLUENCING THE PROCESS

1.1 Theoretical and methodological bases of the foundation of the decision strategy

People face the decision-making process at almost every stage of their lives. In the course of his life, an individual has to choose one or more options. The concept of decision-making refers to the skills and methods learned to achieve goals and make the most appropriate choices from a variety of options under certain circumstances. In other words, decision making is the achievement of goals and objectives. In other words, decision making is to choose one of many alternative action plans to achieve a set of goals. Multiple choices, possibilities, and tactics that are applicable and feasible for all choices, actions, and goals are selected.

The following requirements must be met in order for a problem to be certified as a decision problem:

- Allow multiple action options to be used
- The effect of each action is different from each other.
- Have some goals to achieve

But if there is a way of action, the answer is complete in this situation and the question is executed without it, so it is difficult to make a decision , It is impossible to discuss it. In each of these situations, the decision maker can provide a model for solving the problem. Every day, people have to decide when to wake up, what to eat, what to wear, and when to sleep. In addition, business managers face the obstacles they face, set goals, and make consistent decisions to reach them. Managers determine not only their personal memories, but also the company they work for. Most of their time is spent on important business decisions, such as: B. Company establishment, manufacturing, promotion, financing, organization and operation. Recent competitive growth in developing markets relies on networks. The success of companies doing business in these markets depends heavily on the soundness of management's judgment.

The math model helps managers make decisions. Making the final decision together is a difficult task. The first reason is that the future outcome of the decision is unknown. The second reason is that need to make quick choices. The decision-making process depends on the breadth, importance, complexity, or simplicity of the topic. Nonetheless, the fact that they often share decision-making and functionality includes:

- All choices emphasize the choice of different methods or approaches.
- Many decisions are made consciously, and often decisions are made to carry them out.
- Actions and decisions take time. Decision-making is a series of processes that take place at different points in time.
- The decision is positive and based on future events.
- Due to the unpredictability of the future, decision makers have already achieved their goals.
- You must be responsible for recognizing some dangers and facing the possibility of failure.

Strategic decisions. Strategic decisions are actions that a company takes to improve its performance, increase revenue, and maximize value creation to become more competitive.

Why is the company trying to be more competitive?

1. Those who benefit from the performance of the company, that is. H. The owners who see the capital increase are:
2. It has a positive impact on its most important aspects: government, workforce and population.

The strategic decisions are as follows:

- Long-term management of the company by strategy.
- The scope of the plan also includes building business resources and organizational capabilities.
- Determination of company activity scope.

The characteristics of the strategic decision are as follows:

1. Their nature is complex.
2. See in a fairly unpredictable atmosphere.
3. They influence the final outcome of a company's decision.
4. They require companies to have a coordinated strategy.
5. External networks are just as important, but are set up and maintained by the enterprise.
6. Support the ambitious diversity of the enterprise.

Strategy as a Key Factor decisions. Managers primarily responsible for strategic decision making may only be able to provide the right answers to their challenges. The vast majority of human decision makers recognize and select the right solution, either alone or collectively. They only strive to find and select the most appropriate solution in extreme situations. Therefore, when an approach adds to an attractive option, it seems that the first option should be selected based on these basic criteria. At the core of a strategy is the ability to make decisions in an unpredictable environment, which involves multiple strategic choices. Developing a successful and efficient strategy includes and requires:

Predict uncertain future contours. This is difficult because uncertainties include ambiguities about the probabilities of existing routes and uncertainties about the propagation of probabilities.

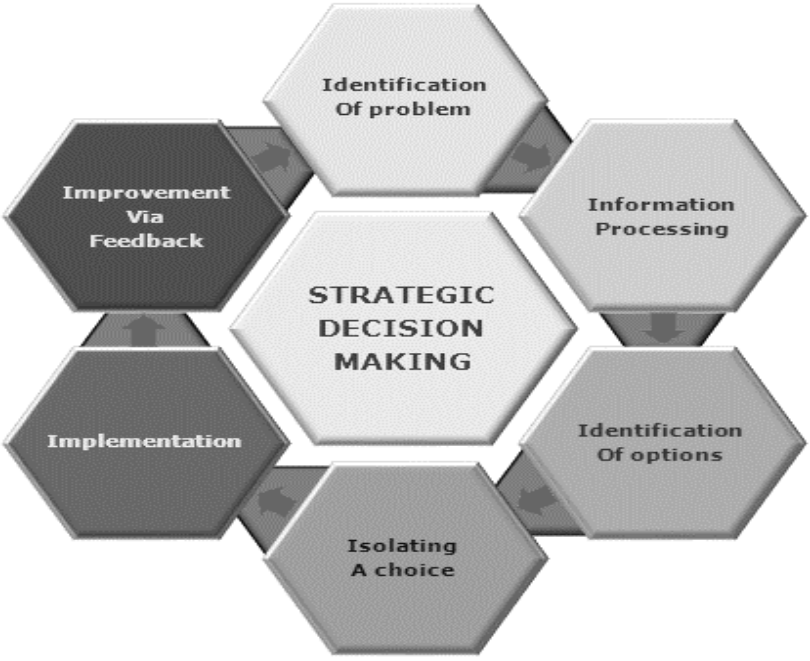
Development of new options for strategic decisions. People use their imagination and emotions when making fresh and first decisions.

Implement new options to enhance the effectiveness of your customization. Adaptation refers not only to the level of an organization's adaptation to the environment, but also to the level of the individual (“what people want and like affects what they see, what they see affects what they want and like”. To do").

It is widely recognized that competent strategic decision-making is required to grow a business and create and capture value. Decision-making issues relate to

the production and acquisition of organization and value. The challenge is to identify the optimal course of action that meets the needs of organization (Figure 1). This requires the ability to develop new strategic opportunities using imaginative means.

Figure 1. Strategic Decision Making



Source – "Thinking, Fast and Slow" by Daniel Kahneman

1. Identification of the difficulty: During this section, the difficulty that calls for a strategic preference is recognized. The end result of this section is the hassle statement.
2. Information processing is the degree whilst records series and statistics processing occur. Using method producing paradigm as a guide, that is the Strategic Assessment degree/section. We observe all outside and inner causes, have interaction in appreciative inquiry, and arrive at loads of aims.
3. The described dreams will function enter for the identity of capability opportunities. This will be the 2nd step of method producing methodology, SITP Planning Process, from an IT method viewpoint (alternatively even the 4th and fifth are associated with it). Otherwise, for you to discover any strategic preference, the dreams could be studied to decide the special manner or opportunities for accomplishing them. The emphasis must be on locating as many capability selections as feasible.

4. After spotting the severe handy opportunities, it's far important to pick the great one. There are numerous qualitative and quantitative techniques to be had for separating the choice. These strategies could be defined in next post. This might additionally offer quantifiable dreams or objectives for the approach.
5. After the choice has been identified/isolated, the implementation method need to be developed. Strategy and Pattern via way of means of Henry Mintzberg will function a catalyst for the method of the Implementation plan. Then, measures for plan execution are carried out, such as the allocation of important assets. Thus, the organization`s assets and abilities will facilitate last implementation.
6. This is the approach for development through feedback. Regular feedback, identity of gaps, and execution of corrective measures are used to evaluate whether or not or now no longer the implementation aligns with the installed quantifiable objectives. Eventually, whilst the favoured goal is reached, it's going to suggest that the strategic preference turned into capable of efficiently deal with the IT/commercial enterprise strategic dilemma.

Management Decision Making. Understanding and improving decision making is management's top priority. Psychologists recommend several options. Most rely on divide-and-conquer law. This technique also violates an important ruling entitled "Problem Fragmentation." Divide the problem into smaller components. It's not a new concept. Benjamin Franklin was one of the first to articulate the tactics of division in a letter to Joseph Priestley. In his explanation, Simon provides theoretical support for this methodology. This idea of "bounded rationality" argues that the limitations of cognitive processing force individuals to construct simpler mental representations of the world. According to Simon, "Human behaves more logically than this model ... in the real world, such behaviour is almost ideal." There are two decision-making methods for management.

- The first deals with the creation and implementation of normative decision rules based on formal logic. As a result of economics or statistics.
- The second section is dedicated to explaining how individuals behave. They really involve making decisions, choices and judgments.

Analysis of Norms. According to Neumann and Morgenstern's games and economic theory, many approaches have been developed to make the best decisions about behaviour. Risk-free options and dangerous consequences are often distinguished. Followed by two instances of each strategy.

Multi-attribute utility. Similar to the reduced MAU, this technique applies to decisions that produce more or less predictable results. As defined by Gardiner and Edwards, MAU involves determining values that will help choose each option and choosing the highest choice. The individual benefits are weighted and summed to determine the usefulness of the alternative. Multiple grades of MAU techniques have been effectively applied to administrative decisions. Personnel selection and zoning selection, like the location of a new factory.

Linear model. First, some regression analysis supported linear models. The main goal was to accurately identify and characterize the assessment. Replica (or attribute) the weight applied to the value attribute. Studies have shown that many parameters apply equilibrium or random weights in addition to the ideal weights. The durability of the linear model allows it to be used in a variety of practical settings such as graduate school admission, clinical diagnosis, and medical decision making.

Analysis of Decision Trees. "A decision tree is a graphical model that shows a set of decisions and events that generate dangerous decision scenarios." This method includes choices of choices, uncertain events, and the provision of results services. The sequence of branches is literally called a "decision tree". The expected value of each choice is determined as the average of all possible results. The solution with the largest EV is the best option. Decision trees have been used to advise on

dangerous decisions such as marketing strategies, crop growth, and public policy planning.

Bayesian theory. This strategy integrates an artificial part of Bayesian probability theory. Intelligence and graphical analysis consist of transforming decisions into analytical tools. Starting with a "fully connected" network, all potential causal relationships between the nodes in the problem area are specified. Through the process of "pruning" supported by computer algorithms, the network structure is reduced to basic node connections. As a result, the complexity of the problem is greatly reduced. Computer diagnostics use this method. Helps to estimate global bugs and problem areas.

1.2 Factors affecting the decision-making stage

In the process of making managerial choices, factors such as time, money, and manufacturing capacity are critical considerations. These factors directly affect the feasibility and practicality of different options. Additionally, elements like facilities, materials, components, technical and scientific knowledge, as well as organizational abilities, play significant roles in determining the overall viability of proposed solutions.

One characteristic of engineering solutions is that the primary attributes of components often require specific investigation or study to be accurately determined. This emphasizes the importance of conducting thorough analyses and research to gain a comprehensive understanding of the technical aspects involved. Quantifiable and precise statements are typically employed to describe the technical considerations, enabling a more focused and targeted approach to engineering analysis and design.

However, it is important to note that managerial choices are not solely based on material and technical factors. Human aspects also come into play. These aspects encompass a range of variables, including political and social convenience, ethical considerations, and moral values. The selection and implementation of alternatives are influenced by these human factors, reflecting the broader requirements and expectations of individuals and society.

Therefore, when making the best possible choice, it is crucial to take into account not only the professional ability to evaluate resources and technical considerations but also the deeply human aspects of the decision. Recognizing the significance of personal values, ethics, and morals ensures a more comprehensive and well-rounded decision-making process that incorporates both the material/technical elements and the human dimensions involved.

1.3 Data and Big Data characteristic

A calculation`s data is a collection of data that has been transformed into an efficient form for transportation or processing. In comparison to current computers and transmitters, information has been transformed into binary digital form. It is allowed to use data as either a singular or plural subject. This word is used to describe data in its most fundamental digital representation.

American mathematician and inventor of information theory Claude Shannon's work explores the origins of the idea of data in the context of computing. Using binary Boolean logic on electrical circuits, he invented binary digital notions. Numerous peripheral devices extensively used in computing today are based on binary number formats, as are CPUs, semiconductor memory, and disk drives. Information management also took the form of computer input punch cards, followed by magnetic tapes and hard drives.

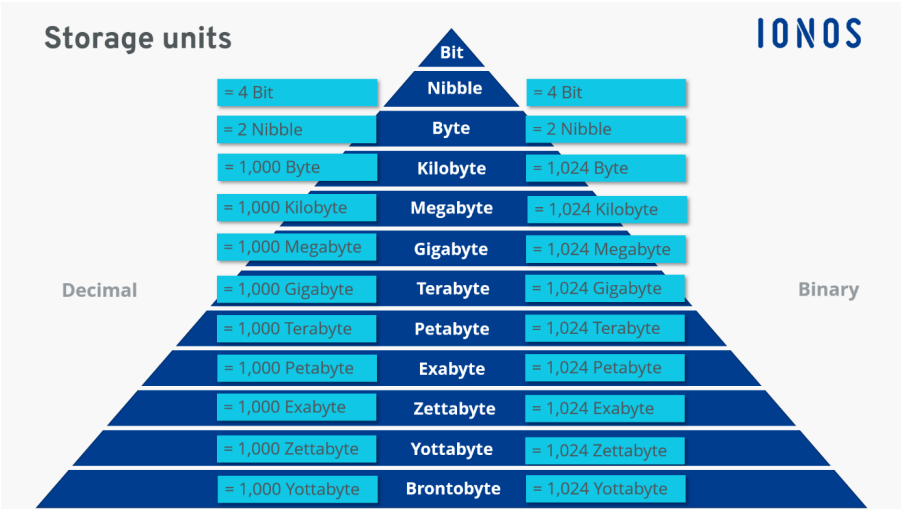
With the prominence of the words "data processing" and "electronic data processing," which formerly included all aspects of information technology, the significance of data in business calculations has been apparent for a long time. In the history of corporate data, specialization has occurred, and with the advent of corporate data processing, a new data profession has formed.

Data storage in the database. The computer uses binary values, represented by 1 and 0, to display various forms of data such as video, images, music, and text. At the fundamental level, a bit serves as the smallest unit of data, representing a single value of either 1 or 0. However, to handle larger amounts of information, data is organized into bytes, which consist of 8 digits or bits.

The storage capacity of computer systems is typically measured in higher-level units such as megabytes (MB) and gigabytes (GB). A megabyte represents approximately one million bytes, while a gigabyte represents approximately one billion bytes. This increased storage capacity allows for the accumulation and management of vast amounts of data.

With the growing expertise in the field, the development of databases has become crucial for efficient data creation and management. Databases provide structured frameworks and systems for organizing and storing data, enabling easy retrieval and manipulation. The advances in database techniques have paved the way for sophisticated data management systems, ensuring the effective handling and analysis of large volumes of information (Figure 2).

Figure 2. Data Storage



Source - "Database Systems: The Complete Book" by Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom

Through appropriate database techniques, organizations can leverage their data resources to gain valuable insights and make informed decisions. These techniques include data modeling, normalization, indexing, query optimization, and data security measures. By employing these techniques, data can be efficiently stored, retrieved, and utilized, enhancing productivity and enabling data-driven decision-making.

Overall, the progression in expertise and technological advancements has revolutionized the way data is stored, managed, and utilized. From the binary

representation of data to the development of databases and sophisticated database techniques, these advances have significantly enhanced our ability to handle and leverage data effectively in various domains and industries.

Data class. Over the past decade, advances in the Internet and smartphones have played a significant role in the creation and expansion of digital data. Nowadays, data encompasses a wide range of sources, including diary entries, online activity logs, text documents, audio recordings, and video materials. Some of these data types are unstructured, meaning they do not conform to a specific format or organization.

The concept of big data goes beyond the scale of petabytes, representing an enormous volume of information. Big data is characterized by the three V's: volume, referring to the vast amount of data generated; versatility, indicating the diversity of data types and sources; and speed, highlighting the rapid rate at which data is produced and needs to be processed.

The proliferation of web-based e-commerce platforms has given rise to data-driven business models that rely heavily on leveraging data. Companies now consider data as a valuable asset and utilize it to drive decision-making and gain insights into consumer behaviour, market trends, and operational efficiency.

With these advances, the social application of information and data confidentiality have become increasingly important. The responsible handling and protection of data have become crucial considerations to ensure privacy and security in an interconnected world.

It's worth noting that the term "data" holds meaning independent of specific data processing tools. For example, in the context of electrical components or network communications, "data" refers to the information being transmitted or exchanged, distinct from terms like "control information" or "control bits" that describe the primary content of a transmission unit. Moreover, in various fields such as science, finance, marketing, demographics, and healthcare, the term "data" is used to denote collections of information that serve specific functions and purposes.

Overall, the advances in the Internet, smartphones, and data-driven technologies have ushered in an era of unprecedented data generation and utilization. The implications of data extend beyond technical aspects and have profound societal, economic, and ethical implications that require careful consideration and responsible practices.

Big data. The term "big data" can be a bit misleading (size is one of them, but many) as it indicates that the current data is low or that its size is of primary concern. More than that). Big data often refers to data that cannot be processed or evaluated by traditional methods or technologies. Big data is the biggest obstacle companies are facing today. With access to a wealth of semi-structured or unstructured information, it is not possible to determine how much value to get from it. Still, the problem is that they can't determine if they're worth keeping. This article describes the term big data and its relationship.

- What is big data?
- Big data features
  - Data speed
  - Data volume
  - Data diversity

According to IBM research, the vast majority of business leaders do not have access to the information they need to do their jobs. These issues are faced in an environment where executives can store anything and generate unprecedented information. This presents a serious data dilemma.

Think of this as a difficult problem. Organizations have access to more potential data than ever before, but as these potential data mining accumulates, the amount of data a company can analyse is rapidly diminishing. But today, as each global scale evolves, the era of big data is in full swing.

The tool is the primary way we can feel more. If we can feel it, we will immediately try to maintain it (at least part of it). People and things are not only connected from time to time, but even more as a result of the development of

communication technologies to meet today's needs. This connection is commonly referred to as Machine to Machine (M2M) and depends on the annual growth rate of the two-digit data.

Small integrated circuits are readily available, adding intelligence to almost anything. There are hundreds of such sensors in wagons. Integrated into the wagon, these sensors monitor the current state of the wagon, the status of specific components, and GPS-based data for monitoring and transportation logistics. As a result of railroad derailments, the government is monitoring such information to prevent future disasters. To avoid such catastrophes, the government has created standards for the storage and analysis of such data. Interpret sensor data for perishable components such as bearings, identify parts that need to be repaired in advance to avoid failure, and add processors to avoid accidents. All of these affect rail cars. Trains aren't the only ones that are smart, as the actual railroad tracks contain sensors every few meters. In addition, data storage requirements apply to the entire ecosystem, including cars, trains, railroad crossing sensors, air samples that contribute to rail transport, and more.

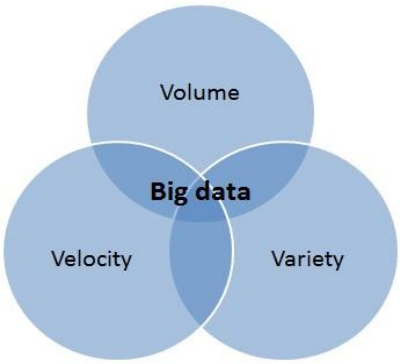
Combine this with railcar weight, arrival and departure time tracking to quickly see problems with big data. Even if all of this data is linked, it is completely unprocessed and can take many forms, making it difficult for traditional communication systems to process. Using the railcar example, we find where speed, quantity, and diversity create a big data dilemma.

What are the characteristics of Big Data? Big data is characterized by three characteristics: volume, diversity, and velocity (Figure 3). Together they define big data. The main reason we need a new skill class to change our current practice is to better understand and better manage our current knowledge and ability to act properly.

Amount of information. Currently, the amount of data stored is exploding. Statistics show that in 2000, there was 800,000,000 petabytes (PB) of data on Earth. The second point to consider is the fact that some of the information generated today

has not been evaluated at all. Over the next few years, this number is expected to exceed 35 zettabytes (ZB). Some organizations generate huge gigabytes of data hourly, daily, and yearly. Twitter alone creates 7 terabytes (TB) of data every day, while Facebook generates 10 TB per day. Individual enterprises no longer need petabyte-enabled data storage clusters. Come to think of it, the fact that it's flooded with data is pretty shocking. Save everything, including environmental data, financial data, medical data, management data, and more.

Figure 3. Data characteristics



Source - "Data Science for Business" by Foster Provost and Tom Fawcett

For example, removing the phone from the grave creates an opportunity. The opening of the entrance to the S-Bahn for boarding is an event. Registering flights, buying songs on iTunes, changing channels on TV, and choosing an electronic payment method are all actions that generate data. A single-homed PC only needs to approach gigabytes of intrusion speed as an indication that more data can be accessed than ever before. Ten years ago, we created a list of all known data stores with a size of almost terabytes. Regarding volume, everything has changed. As the phrase "big data" suggests, corporations deal with massive volumes of information. Organizations who do not know how to handle this data are thus perplexed. Nevertheless, with the appropriate technological platform, may analyse almost all of data (or at least more by recognizing important data) to get a deeper understanding of organization, consumers, and market. This results in the challenge faced by firms in all sectors today. As the quantity of information a business has access to rises,

the proportion of information it can process, comprehend, and analyse diminishes, resulting in a blind hole.

What is inside the blind spot? The exponential growth of data has given rise to a fundamental dilemma known as the "I don't know" problem. With the massive amounts of data being generated, there is both great potential for valuable insights and the challenge of extracting meaningful information from the vast sea of data. This uncertainty regarding the true value and relevance of the data creates a significant challenge for organizations and individuals alike.

As the scale of data continues to expand, the conversation has shifted from terabytes to petabytes and even zettabytes. Conventional systems and storage methods are often insufficient to handle and retain such enormous volumes of data. The sheer size of zettabytes of data necessitates the exploration of alternative approaches to data storage, management, and analysis.

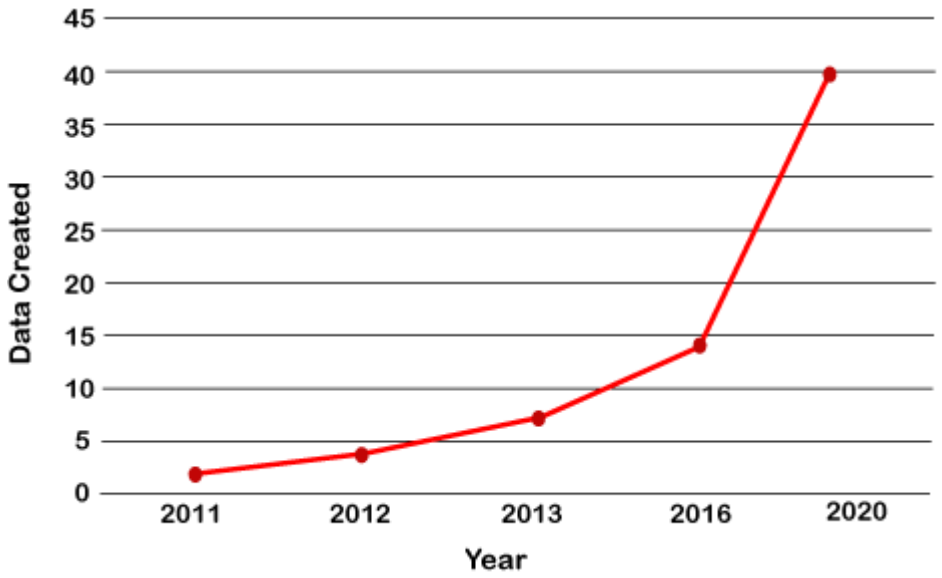
To address this challenge, organizations are turning to innovative technologies and strategies such as distributed computing, cloud storage, and big data analytics. These approaches allow for the storage and processing of vast amounts of data across multiple systems, enabling scalability and flexibility in data management.

However, it is important to note that not all data generated needs to be stored indefinitely. The value of data varies, and organizations must prioritize and determine which data is worth keeping and investing in for further analysis and insights. Data governance practices and policies are being developed to define criteria for data retention, privacy, and security, ensuring that valuable and relevant data is preserved while avoiding unnecessary storage costs and complexities.

The evolving landscape of data quantities poses both challenges and opportunities. It requires organizations to carefully consider their data strategies, exploring new storage and analysis approaches, and making informed decisions on data retention and utilization. The ability to navigate this vast sea of data effectively

will play a crucial role in unlocking its potential and deriving actionable insights for various domains, including business, research, healthcare, and beyond. (Figure 4).

Figure 4 – Data Growth



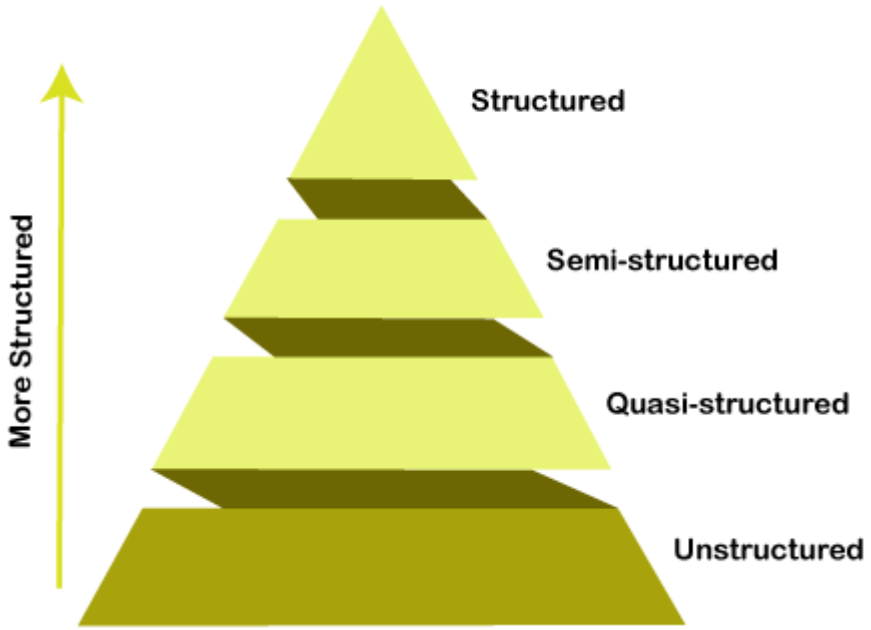
Source - "The Growth of Data and the Impact of Big Data Analytics" by McKinsey & Company

Diversity of data

The volume connected with Big Data is one of the causes of the new challenges faced by data centers attempting to cope with this phenomena. The advent of sensors and smart devices, in addition to social collaboration tools, has complicated enterprise data, since it incorporates not only conventional contact information, but also raw, semi-structured, and unstructured information through web pages and web logs. These include files, search indexes, social media forums, emails, papers, sensor data from active and passive systems, and more (Figure 5). One sample is available. In addition, conventional systems may struggle to retain and execute the necessary analytics to comprehend the content of these records, since some of the created data is incompatible with typical database technology. Some companies are moving in this direction, but think the majority are just beginning to recognize the power of big data. To put it simply, diversity includes all types of data. Examples include radical changes to the analytical requirements for including raw, semi-structured, and unstructured data along with traditional structured data as part of the decision-making and thinking process. Traditional

analytical tools cannot handle diversity. The success of an organization also depends on its ability to generate ideas from both traditional and non-traditional sources. When we concentrate on our database careers, we often discover that we spend the majority of our time on only 20 percent of the data: a sort of connection that is properly written and adheres to our rigid schemes. 80 percent of the world`s data is unstructured or at most semi-structured, and the majority of this data is responsible for establishing new speed and volume records. Examining the Twitter feed reveals its structure in JSON format, but the actual content is unstructured. Videos and photos cannot be easily or conveniently saved in a database, and some occurrences may cause data to dynamically alter (e.g., weather patterns). To take use of Big Data, businesses must be able to evaluate all forms of data, both relevant and irrelevant: text, touch data, audio, video, transactions, etc.

Figure 5. Diversity of Data

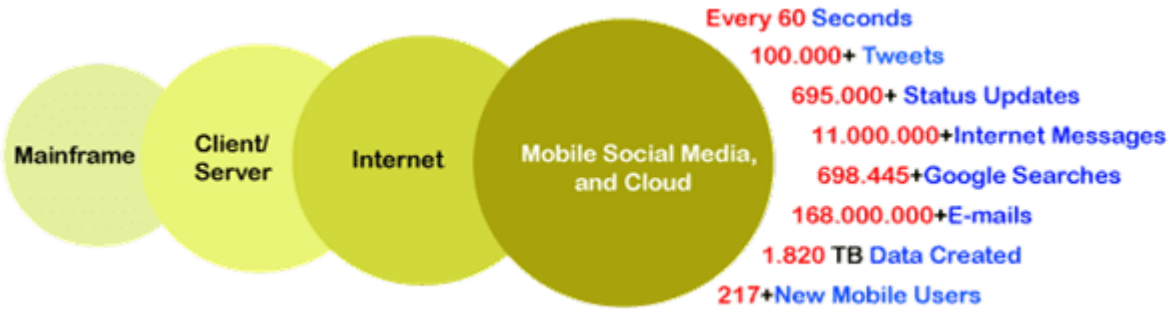


Source – "Data Feminism" by Catherine D'Ignazio and Lauren F. Klein

Data Speed. The amount of information we have and the speed at which diversity is being created must be processed. The concept of speed often refers to the speed at which data is entered and stored and the associated search speed. It is an advantage that all of this is processed quickly (Figure 6). In addition, the amount

of data investigate depends on how quickly the data arrives. New approaches to the problems that can arise with changes in speed should start with the source of the data. Rather than limiting the concept of speed to the growth rate of a data warehouse, it is better to apply this term to data movement. The speed at which information is sent. Organizations are currently processing petabytes of data instead of terabytes, and believe that the proliferation of RFID sensors and other data flows is creating data flows that traditional systems cannot manage. Finding trends, challenges, or opportunities just seconds or microseconds ahead of competitors can often lead to a competitive advantage. In contrast, most of the data produced today has a relatively short retention period. Therefore, companies must be able to evaluate this data in real time if they want to gain insights from the data. Traditional processing can be expected to execute queries against relatively static data.

Figure 6. Data Speed



Source – "Database Systems: The Complete Book" by Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom

For example, the query "Show everyone living in the ABC flood zone" is used as a weather alert list. Calculating the current allows us to perform a process similar to a continuous survey identifying people in the ABC Flood Plains, but the results are constantly updated as the spatial information in the GPS data is updated in real time. Will be done. To take advantage of big data, need to evaluate the amount and type of data not only during storage but also during transmission. From monitoring new born health to analysing financial markets, it is clear that new methods are needed to handle the volume and diversity of data.

1.4 Data Analytic, Data Science and Machine Learning concepts

What is Data Science?

Informatics has been a subject of research and exploration for more than a decade, aiming to understand and harness the power of information in various domains. One approach to conceptualize the interdisciplinary nature of informatics is through Hugh Conway's Venn diagram, created in 2010. The diagram illustrates three overlapping circles representing math and statistics, subject literacy, and hacking talent. A comprehensive understanding of information science is achieved when one possesses knowledge and expertise in all three areas.

Informatics involves the management of large volumes of data, encompassing tasks such as data cleansing, preparation, and analysis. Data scientists employ various techniques, including predictive analytics and scenario analytics, to gather data from multiple sources, analyze it using computational methods, and extract meaningful insights and patterns from the collected datasets. By examining data from a business perspective, data scientists can provide valuable forecasts and insights that aid in making informed and impactful business decisions.

The field of informatics serves as a bridge between data analysis and its application in real-world contexts. By leveraging their skills and expertise, informatics professionals play a crucial role in enhancing the understanding, interpretation, and utilization of data-driven insights. They contribute to improving business strategies, optimizing processes, and driving innovation across industries.

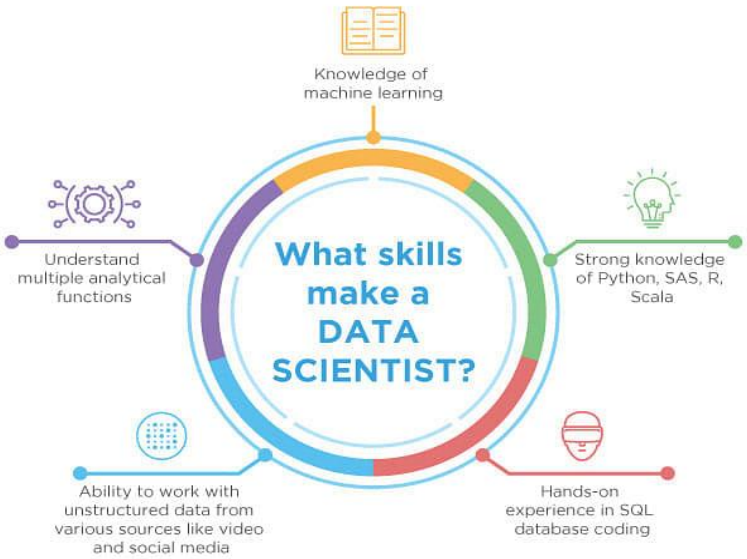
As technology continues to advance and the volume of data grows exponentially, the field of informatics will remain at the forefront of data management and analysis. The ability to effectively navigate and extract valuable insights from complex datasets will be instrumental in solving complex problems and driving evidence-based decision-making in a wide range of domains, including healthcare, finance, marketing, and beyond.

Overall, informatics provides a framework for understanding and leveraging the power of data, combining mathematical and statistical knowledge, subject expertise, and technical proficiency. Through its interdisciplinary approach,

informatics enables professionals to unlock the potential of data, driving progress and innovation in today's data-driven world.

Skills required to become a Data Scientist. Those interested in pursuing a successful career in this area need to acquire expertise in three areas: analytics, programming, and domain knowledge (Figure 7). If you want to grow in this area, the skills you need to become a computer scientist are:

Figure 7. Data Scientist Skills



Source – "Python for Data Analysis" by Wes McKinney

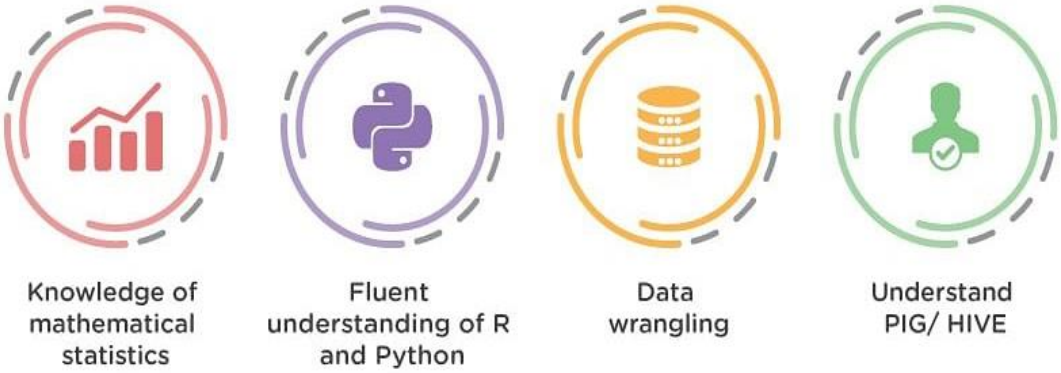
- Python, SAS, R, Scala expertise
- Ability to process unstructured material from a variety of sources, including video and social media
- Knowledge of some analytical functions.
- Machine learning skills

Data analysts can typically generate descriptive statistics, display data, and provide data points to draw conclusions. You need a basic understanding of statistics, a complete understanding of the database, the ability to generate new insights, and the ability to view data. Data analysis is also an essential aspect of informatics. Many skills required to become a data analyst.

What is data Analytics? To excel as a data analyst, individuals need to possess specific skills and abilities that enable them to effectively describe and

communicate information in response to business requests or problems. These skills are crucial for conveying insights and findings to relevant stakeholders. Here are four key abilities that are highly beneficial for aspiring data analysts (Figure 8):

Figure 8. Data Analyst Skills



Source – "Python for Data Analysis" by Wes McKinney

Knowledge of quantitative statistics: Data analysts must have a solid understanding of quantitative statistics. This includes familiarity with concepts such as probability, statistical distributions, hypothesis testing, regression analysis, and data visualization techniques. Proficiency in statistical analysis enables data analysts to extract meaningful insights from data and make data-driven recommendations.

Good understanding of R and Python: R and Python are popular programming languages extensively used in data analysis. Proficiency in these languages empowers data analysts to efficiently manipulate and analyze data, perform statistical modeling, create visualizations, and automate analytical processes. Being well-versed in R and Python opens up a wide range of analytical tools and libraries that enhance data analysis capabilities.

Information contention: Data analysts need to be skilled in managing and organizing data. They should possess the ability to gather, clean, transform, and integrate data from multiple sources. This includes expertise in data wrangling, data cleansing techniques, data preprocessing, and database querying. Effective data management ensures that the data is reliable, consistent, and ready for analysis.

Familiarity with PIG/HIVE: PIG and HIVE are data processing languages commonly used in big data environments, specifically for working with Hadoop-

based systems. Having familiarity with PIG and HIVE enables data analysts to process and analyze large-scale datasets efficiently. It involves writing queries and scripts to extract and transform data stored in distributed computing environments.

By developing and honing these abilities, aspiring data analysts can enhance their analytical capabilities, enabling them to extract valuable insights, effectively communicate findings to stakeholders, and contribute to data-driven decision-making processes within organizations.

It's important to note that the field of data analysis is constantly evolving, and there may be additional tools, technologies, and skills relevant to specific industry domains or analytical requirements. Continuous learning and staying updated with emerging trends and technologies is crucial for data analysts to remain effective and adapt to evolving data challenges.

Data Science vs Data Analytics. Information is a broad term that includes a variety of related disciplines such as scientific data analysis, data learning, and machine learning. Computer scientists are supposed to use past examples to predict the future, but data analysts collect important information from many sources. Data scientists generate questions, and data analysts find solutions to existing questions.

What is the definition of machine learning? Machine learning refers to the ability of systems to acquire data, learn from it, and utilize algorithms to identify patterns and make predictions or decisions in a specific domain. Traditional machine learning systems employ various techniques and algorithms to process and analyze data, ultimately revealing hidden insights and concepts. Here are some key aspects to consider when exploring traditional machine learning:

Data acquisition: Machine learning systems require access to relevant and high-quality data. This data can be collected from various sources such as databases, sensors, APIs, or online platforms. The data serves as the foundation for training and fine-tuning machine learning models.

Learning from data: Machine learning algorithms are designed to learn from the provided data through a process called training. During training, the algorithms

analyze the data, identify patterns, and adjust their internal parameters to optimize performance. This iterative learning process helps the system to recognize and generalize patterns, enabling it to make predictions or decisions on new, unseen data.

Statistical and predictive analytics: Machine learning systems leverage statistical and predictive analytics techniques to extract meaningful information from the data. These techniques involve analyzing the data to uncover correlations, trends, and statistical relationships, which can then be used to make predictions or derive valuable insights.

Uncovering hidden concepts: Machine learning algorithms can reveal hidden concepts and patterns that may not be easily apparent through manual analysis. By processing large volumes of data, machine learning systems can identify complex relationships and extract relevant features or factors that contribute to the observed outcomes.

Traditional machine learning encompasses various algorithms such as linear regression, decision trees, support vector machines, and neural networks. Each algorithm has its strengths and weaknesses, making them suitable for different types of problems and data characteristics.

The predictive capabilities of traditional machine learning systems have found applications in numerous domains, including finance, healthcare, marketing, and manufacturing. These systems can help businesses optimize processes, detect anomalies, personalize customer experiences, forecast trends, and make data-driven decisions.

As machine learning continues to advance, new techniques and algorithms are being developed, such as deep learning and reinforcement learning, which enable more sophisticated modeling and decision-making capabilities. This ongoing research and innovation in machine learning contribute to its evolving role in addressing complex problems and uncovering valuable insights from data.

Facebook is the best example of a machine learning application. Facebook's machine learning algorithms capture motion data for all users on social networks.

Provides articles and alerts related to the news stream based on previous actions. Similarly, Amazon and Netflix product and movie recommendations are based on past behavior and are called machine learning applications.

Skills required to become a machine learning engineer. The machine learning perspective is different from the statistical perspective (Figure 9). The following are some of the essential abilities that will assist in launching career in this fast expanding field:

- Computer experience
- Extensive programming expertise
- Knowledge of probability and statistics
- Data modeling and evaluation capabilities

Figure 9. Machine Learning Expert Skills

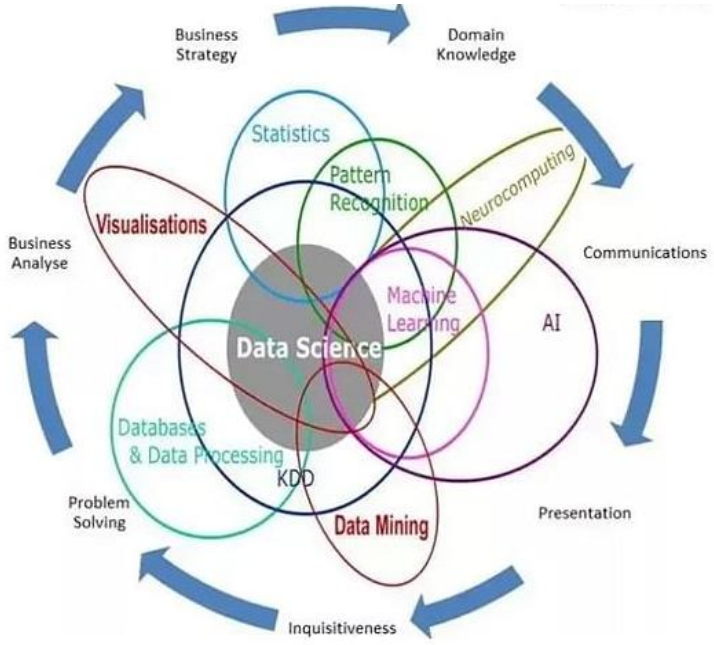
Source - "Pattern Recognition and Machine Learning" by Christopher Bishop

Data Science vs Machine Learning. As information encompasses several scientific areas, machine learning is also included (Figure 10). During machine learning, several approaches, such as regression and controlled clustering, are used. In information science, information may or may not originate from a machine or mechanical process.

Information science, as a larger phrase, focuses on all data processing approaches in addition to algorithms and statistics. Numerous ideas, such as data analytics, software engineering, information engineering, machine learning, predicting analytics, data analytics, etc., may be seen as a combination. This

encompasses the search, collection, receipt, and manipulation of vast quantities of data; together, this is referred to as big data. Information science is responsible for acquiring enormous amounts of structured data, locating relevant instances, and providing modifications to meet the business requirements of decision makers.

Figure 10. Data Science is Multidisciplinary



Source – "Data Science for Business" by Foster Provost and Tom Fawcett

Data analytics and machine learning use a number of the same tools and procedures as information science. Data Science, Data Analytics, and Machine Learning are the most in demand fields in the market right now. The combination of strong skills and real world experience in this field's most in demand fields will help develop a successful career.

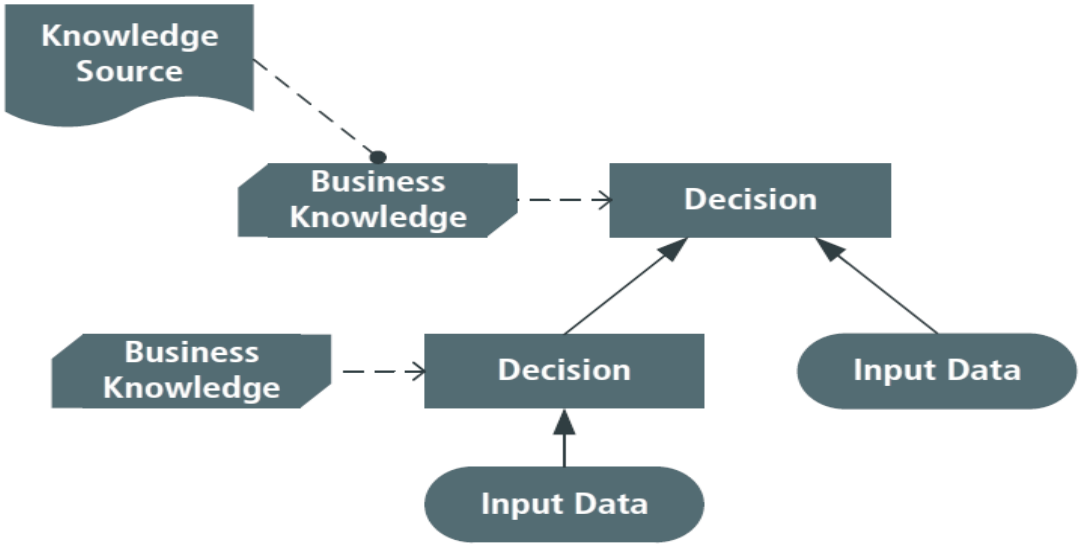
1.5 Data Science Algorithms and Models in Decision Making

The selection-making act is based at the functionality to execute the influencing elements. The availability of statistics to determine whether or not the outcome of a choice is absolutely understood and whether or not preference is advanced has a sizeable effect at the assessment of alternatives. In (Figure 11), some occurrences can be uncontrolled, whilst others can be partial and additionally unpredictable.

The traits of variables, the formation of variations, and the outcomes. Depending at the output formats, distinctive selection fashions have to be used. In a nutshell, the understanding diploma of the selection maker determines the distinction among the fashions hired in selection making. In this regard, selection-making fashions can be classified as follows:

- Decision-making with entire understanding
- Choices made at moments of uncertainty
- Make volatile judgments
- Make selections primarily based totally in element on statistics
- Make selections in an aggressive surroundings

Figure 11. How decisions are made



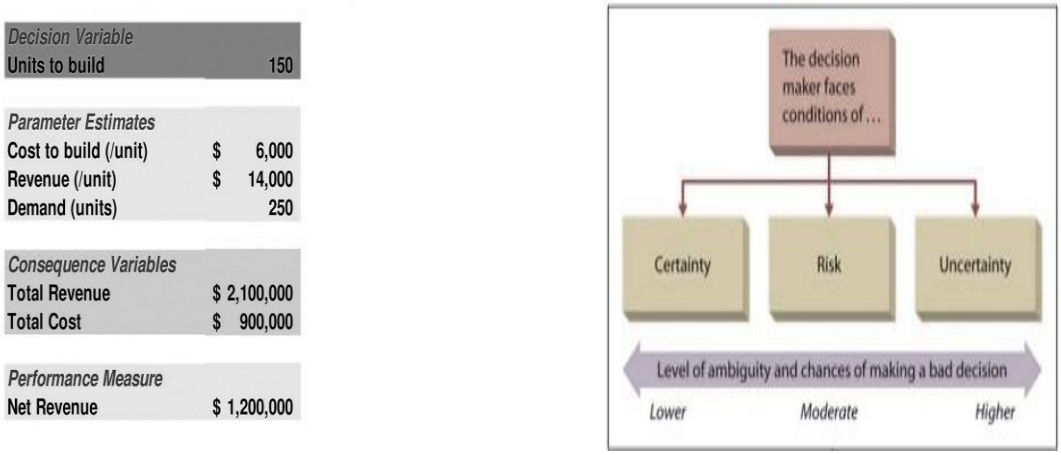
Source - "Thinking, Fast and Slow" by Daniel Kahneman

Probability-primarily based totally selection making. It is truthful to say that whilst creating a preference, it's far glaring below what situations it's far applicable. Thus, the danger of an expected prevalence is 1, making assured selection-making one of the most effective selection-making fashions. Since the selection does now no longer go away any factor of the state of affairs to danger.

Under certainty, preference problems are characterised with the aid of using complete understanding approximately every alternative, and the selection maker has get right of entry to credible records approximately the future. According to this

definition, selection-making refers to any selection-making process. These movements have predetermined repercussions.

Figure 12. Decision Making Certainty and Conditions



Source – "Smart Choices: A Practical Guide to Making Better Decisions" by John S. Hammond, Ralph L. Keeney, and Howard Raiffa

Under certainty, the selection-maker is aware the circumstance. The selection-maker is capable of effectively pick out the first-rate alternative. Consequently, the top-quality go back cost is the best cost for the objective, the achievement rate, and the choice standards is the choice with the best benefit. The use of linear programming fashions exemplifies assured selection making. Among the options, simplest those fashions are mathematically well-defined. It is suitable while it's far well matched with linear functions. In the equal manner that the perfect quantity of earnings on the way to be generated with the aid of using making an investment in authorities bonds is known, the selection to spend money on bonds is an instance of a assured preference in Figure 12.

Uncertainty and selection making. Probabilities and anticipated occurrences. Unspecified selection issues are visible as selection-making problems in conditions of uncertainty. In the occasion of ambiguity, the selection-maker is capable of offer a result.

There aren't any possibilities available. Due to a loss of historic revel in and records, there may be no chance estimate. (Figure 13) Decision-making below instances of uncertainty is the maximum hard and maximum common selection-

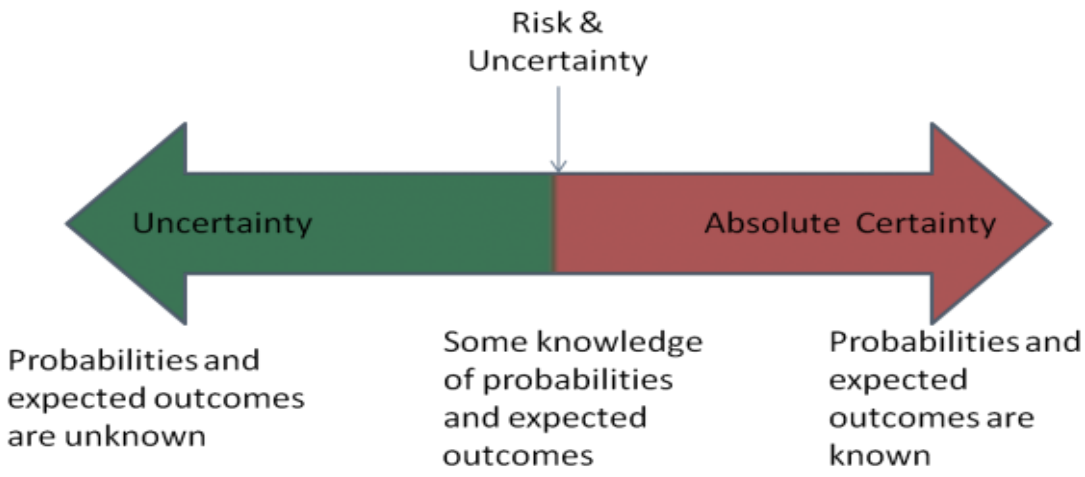
making circumstance. In the example of uncertainty-primarily based totally selection-making, there may be very little statistics at the issue. The standard context wherein enterprise managers make selections is certainly considered one among ambiguity. Due to the lack of know-how at the chance of herbal events, it's far applied on this circumstance. Depending at the selection maker`s optimism or pessimism, the strategies will vary.

In conditions of uncertainty, the subsequent elements are important for selection makers:

- The Laplace criterion
- The pessimistic standards
- The optimism (Maximax) criterion
- The Pi mana (Minimax) criterion
- The realism criterion (Hurvicz)

To make a decision at risk. High-risk choices require different numbers of decisions associated with a particular choice. All elections have pre-determined results under all circumstances. It can occur. Choose alternatives based on decision-making, or probability. Risk determination is relevant to this situation. It's not yet clear what he will do after he resigns. The decision maker in this case. It estimates the expected monetary value and selects the best choice based on the assumption that a particular situation will occur with a particular probability.

Figure 13. Risk and Uncertainty



Source – "The Black Swan: The Impact of the Highly Improbable" by Nassim Nicholas Taleb

In a natural scenario, the probability distribution of uncertain situations with risky decisions is either unknown or unknown. That is, the decision maker is objective or subjective. If you can evaluate the probability, you are operating in an ambiguous environment. If this is not possible, operating in a high-risk environment. In order to produce consistent choices, the decision maker made a few adjustments to the issue. Calculating the odds of natural occurrences is the most logical approach. Before assessing and resolving the issue, it is necessary to gather additional information. Using historical data, it is possible to estimate the likelihood of natural catastrophes, particularly if the choice to be taken has been made often in the past. In a risk scenario, probability distributions are often used to calculate the costs associated with each option alternative. Therefore, risky decisions are often based on alternative expected value criteria and expected profit or forecast cost reductions. There are three main ways to make dangerous choices. They are:

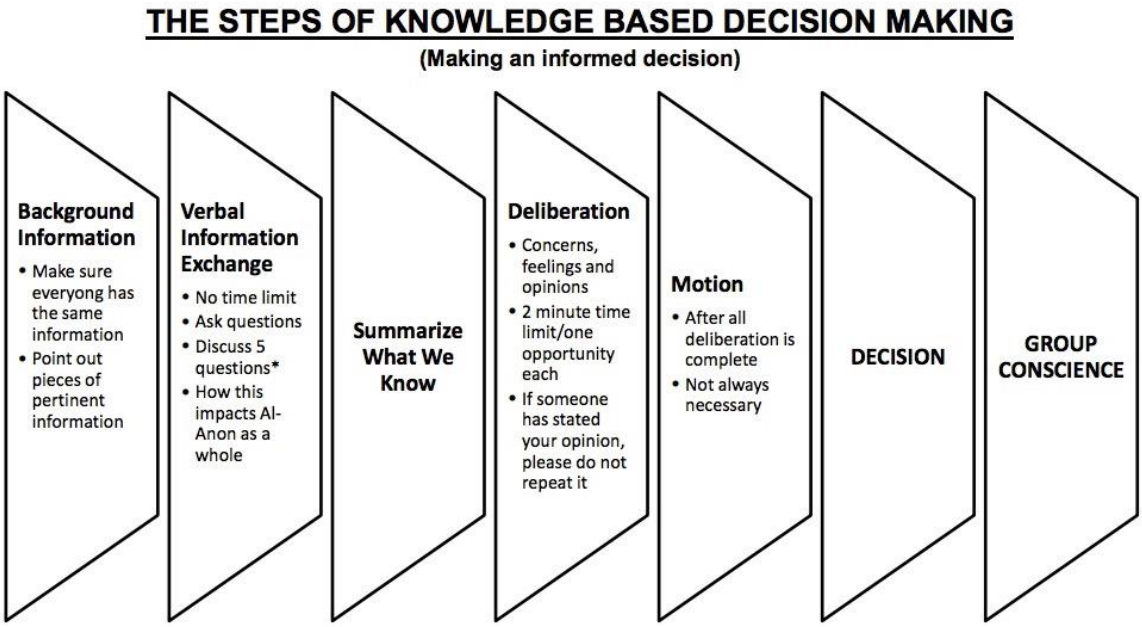
- Expected value criteria
- Expected loss (loss) criteria
- Opportunity loss criteria

Informed decisions. Future environmental conditions are known to decision makers (Figure 14). New resources can be used to improve the prediction of the potential for natural disasters. Additional resources include markets, research, research, product testing, expert commentary, samples, and simulation products. This additional data usually comes at a cost. Decision makers must first determine if it makes sense to bear these costs. In addition to data from various sources, Bayesian statistics are used to determine the probability of changes in the natural state.

Competitive decision-making. A competitive approach to decision making. Several decision makers participate in a competitive decision-making process. This includes the selection dilemma. This type of selection dilemma is called a "game." It is defined and developed through the lens of "game theory". Certain types of problems arise as a result of conflict or cooperation between two or more parties.

This includes questions about optimization decisions. Reasonable decision makers in a highly competitive business environment rely heavily on the activities of other decision makers or competitors.

Figure 14. Steps of Decision Making



Source - "Decision Making for Dummies" by Dawna Jones

In a highly competitive environment, many commercial and economic issues need to be determined. Competitor advertising and marketing strategies include new item decisions, procurement and supply policy settings, production planning in the face of unpredictable demand, capital budgets, and competitive decisions. In each of these situations he accepts the ideal choice. It will be a specific opponent for which other rivals will follow the results. It also depends on the options and the results they produce. The organization offers a variety of models and methods designed to simplify the decision-making process and specific decisions. When making decisions, we need to keep in mind that there are many elements that we can and cannot control. Moreover, all probability-based decisions we make always carry some uncertainty. In this post, you'll learn many decision-making frameworks and other tactics that can be applied to business.

Model of company decision strategy . Consider organization's decision-making methods in the light of the models considered in the next section. In

particular, reduce the cost-benefit of "bad" behavior so that an organization achieves its goals. These models help choose the best option from the options offered, taking into account the uncertainties that are always present, or the probability of error (which can be reduced as described above).

Maxim (or Wald) Model When deciding whether to use the. The Maxim Model, also known as the Wald Model, is a decision-making approach that aims to prioritize or select the option with the lowest (worst) possible outcome. It takes a "graphic" form where each possible solution is assigned a score or value based on its perceived quality or performance.

In the context of the Maxim Model, let's consider three solutions: A, B, and C. Suppose A receives a score of 1, B receives a score of 2, and C receives a score of 3, with 1 being the lowest and 3 being the highest. According to the Maxim Model, one would choose solution C because it has the best score among the three options.

However, it is important to note that the Maxim Model has its limitations. By solely focusing on the worst possible outcome and selecting the option that minimizes this outcome, there is a risk of overlooking other potentially valuable or innovative solutions. This decision-making approach can be considered "pessimistic" in nature, as it prioritizes risk avoidance and minimization over exploring alternative possibilities.

While the Maxim Model ensures a cautious and conservative approach, it may not always lead to the best or most appropriate solution for a given situation. It disregards the potential benefits or advantages that other solutions might offer, even if they come with certain risks or downsides. By solely considering the worst-case scenario, one may miss out on valuable knowledge and opportunities for improvement.

Effective decision-making requires a comprehensive assessment of various factors, including risks, benefits, trade-offs, and uncertainties. It is important to strike a balance between risk mitigation and exploring potential opportunities for growth and innovation. Different decision-making models and approaches, such as

the Maxim Model, can serve as valuable tools in specific contexts, but they should be used alongside other techniques to ensure a well-rounded and informed decision-making process.

Maximax model. The Maximax model, also known as the optimistic model, takes the opposite approach to decision-making compared to the pessimistic model discussed earlier. In this model, the focus is on selecting or working with the option that has the highest possible outcome or score, assuming the best-case scenario.

To illustrate the Maximax model, let's consider three solutions: A, B, and C. Suppose solution A receives a score of 8, solution B receives a score of 10, and solution C receives a score of 9. According to the Maximax model, the best solution would be B because it has the highest score among the options.

The Maximax model is considered an optimistic approach as it emphasizes the potential for achieving the best possible outcome. It assumes that the highest-scoring option will prevail and disregards the potential risks or downsides associated with that choice.

However, it's important to note that the Maximax model also has its limitations. By solely focusing on the best-case scenario and choosing the option with the highest score, decision-makers may overlook potential risks, uncertainties, or trade-offs associated with that choice. It may lead to overestimating potential benefits or underestimating potential downsides.

Effective decision-making involves a balanced assessment of multiple factors, including both the upside and downside of each option. While the Maximax model can be useful in highlighting the potential benefits of a particular choice, it should be used alongside other decision-making techniques and considerations.

Ultimately, the selection of a decision-making model depends on the specific context, the availability of data, the preferences and risk tolerance of the decision-makers, and the nature of the problem being addressed. Multiple decision-making models exist to cater to different perspectives and trade-offs, allowing decision-makers to make more informed and robust choices.

The application of data science algorithms and models in decision-making empowers organizations to make evidence-based, data-driven decisions. It allows for more accurate predictions, improved efficiency, optimized resource allocation, and better risk management. By leveraging the power of data science, decision-makers can gain valuable insights and enhance their decision-making processes to drive business success.

Here are some key points to consider when exploring the impact of data science in decision-making:

• Descriptive Analytics is the primary and most important aspect of analytics. This section of the analysis phase is where most of an organization`s time is spent. What Else Occurred During This Phase? The question must be answered.

• Diagnostic Analysis Diagnostic analysis is a term used to determine the reason for what is happening in the data. At this point, it's beyond observation. Why did this happen? For example, why are there slopes at the top and bottom of the chart? The ability to ask questions about information and connect those demands to business goals and requirements is the most important skill. Imagine diagnosing us, declaring us ill, and meeting a doctor leaving the room. This does not affect our health as we need treatment. Doctors need to observe us, make a diagnosis, and develop a treatment plan to make us feel better. This example is as important as analysis. Observe, create a descriptive analysis, and proceed to diagnosis.

• Predictive analytics Predictive analytics develops business strategies by making positive decisions based on the performance of the company's weakest business segments. In the first three steps, understand how organization can act. Predictive analytics uses machine learning algorithms and artificial intelligence techniques to implement the prediction phase. What is the next step? We have to find a solution to that question.

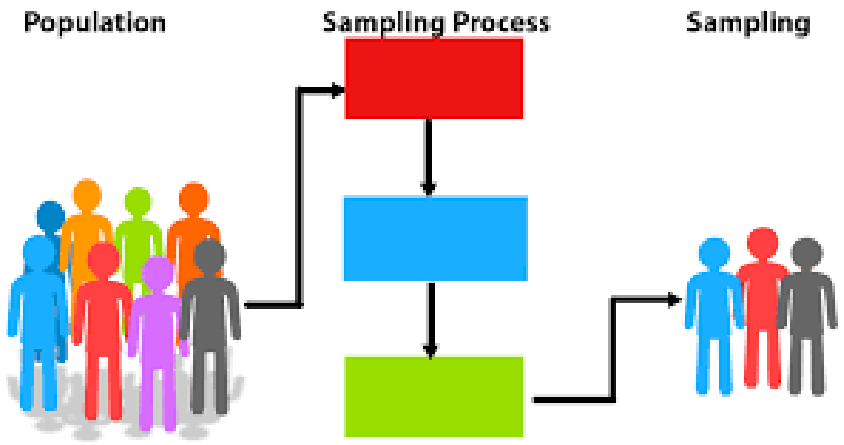
• Prescriptive analytics. This phase is very difficult. Why-why covers analysis. Processing after displaying the data and creating the forecast i. H. What are

we doing and what are we observing at this point? This is the most important stage of the company's development.

Statistics is an essential discipline in the modeling process, providing the foundation for understanding data, formulating hypotheses, estimating parameters, making predictions, and assessing model performance. By combining statistical knowledge with algorithms and other data science techniques, researchers and data scientists can build robust and reliable models to extract meaningful insights from data. Populations, samples, parameters, and statistics are basic statistical concepts. The majority of the population. This is the total data available. For example, information about everyone on the planet. The population is determined by the parameter indicated by the letter N.

X representing the sample. It is that part or subset of the population that is selected for research and identification. Case: The previous example, Population, collected data from all people around the world, but here we include data from people based on selected countries (Figure 15). The results of the population analysis are called parameters. Which group will be selected depending on the findings after examining the statistical sample, based on these results, information about large groups is provided.

Figure 15. Population and sample



Source – "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Description and Examination of Statistics - The statistical evaluation includes parts:

1. Descriptive Statistics - this path covers the collecting, summary, evaluation, and presentation of statistics.

2. Statistical Analysis - primarily based totally at the findings of a small organization, which organization is picked, and data on massive agencies is supplied primarily based totally on those results.

Variable: Indicates the kind of factors and those with this variable. For instance, the functions of the people with inside the room encompass their names, heights, genders, and weights.

Given: are a whole lot of values derived through the variable. Using the instance from the preceding paragraph as a guide, the heights of the folks with inside the room can also additionally range as follows: 1.65, 1.54, 1.74, 1.78, 1.89, 1.68, etc. Therefore, we infer that the statistics have wonderful values.

Variables are separated into qualitative and quantitative variables:

- Categorical variables are from time to time referred to as qualitative variables. Unlike quantitative variables, qualitative variables regularly do now no longer encompass numbers. Example: ranges of schooling, along with secondary school, bachelor`s degree, master's degree, and doctorate. As may be seen, schooling ranges are separated into classes and encompass express values; there aren't any numerical values. As said for qualitative variables, quantitative variables regularly encompass numbers. Quantitative variables are discrete and partitioned into wonderful portions.
- Discrete variables have constantly whole values and are appeared variables. Examples of discrete variables are the wide variety of people in a room, the amount of apples, etc.

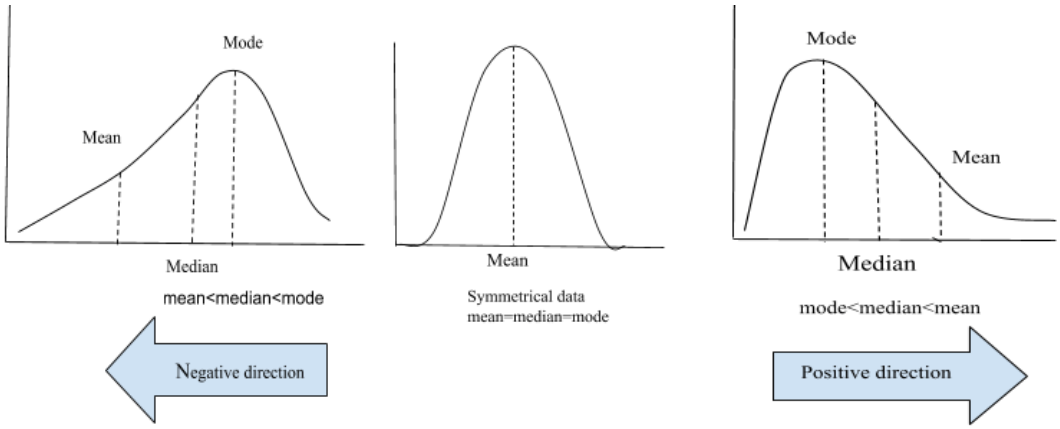
Finite variables are non-stop variables, as implied through their name. In different words, it ought to include a restricted wide variety of values and won't include numbers. For instance, we can also additionally take human weight as an

example, when consider that weight has various values, is finite, and includes fractions.

Distributions of statistics - Descriptive Statistics. The distribution of statistics follows the Kurtosis and Skewness forms. The distribution proven with inside the following graph is the Skewness distribution. Negative path - If it's far much less than the mean, median, and mode, the path of the distribution could be terrible and Skewness could be skewed to the left (Figure 16). Additionally, a terrible distribution could be very skewed if the skewness is much less than -1 or large than +1. This is a terrible distribution condition.

Normal Distribution or Symmetrical Distribution Thus, Skewness equals zero. If the skewness is among -1 and -0.five or +0.five and +1, the distribution is taken into consideration extraordinarily skewed. In this instance, the mean, median, and mode are identical.

Figure 16. Distributions of data



Source – "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Positive path - Skewed receives extra right-slanted (Figure 17). The distribution is kind of symmetrical if the skewness is among -0.five and +0.five.

Kurtosis - illustrates the shape of the statistics distribution relative to its center (Figure 18). The following are the correlation situations in diagnostic evaluation:

- There ought to exist a connection among the 2 variables.
- The correlation coefficient ought to variety among -1 and 1

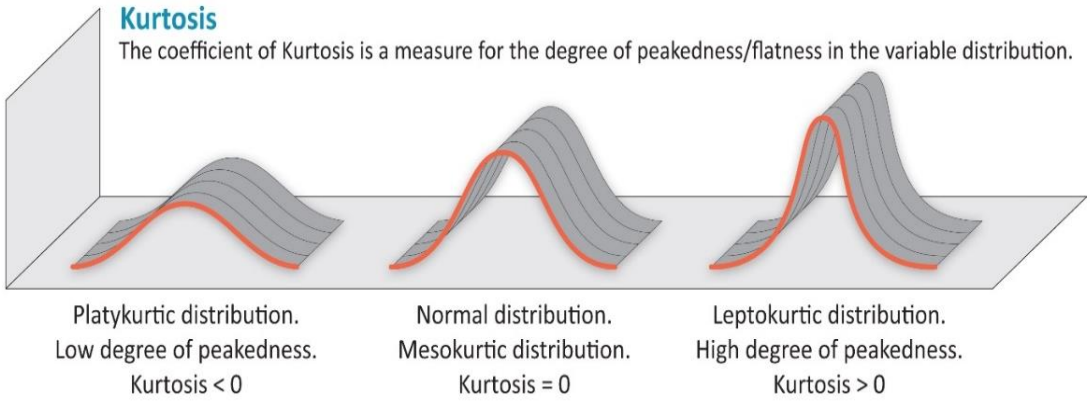
• If the price is close to 1, the affiliation among the variables is strong, i.e. the correlation is high; if the price is near 0. The hyperlink among the variables is weak.

• It shows hyperlink among values is significant.

The Pearson correlation is characterized by the following conditions (Figure 19):

- Determines the direction and strength of the connection between continuous and discrete data. The value is between[1:1].
- If the p-value is less than or equal to 0.05, the relationship between the values is important.

Figure 17. Kurtosis



Source – "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Figure 18. Statistic test

Variables	Parametric	Non-parametric
2 numeric	Pearson Correlation	Spearman Correlation
2 ordinal	N/A	Spearman Correlation
2 categorical	N/A	Chi Square
1 numerical and 1 categorical	ANOVA	Kruskal-Wallis or Mood & Friedman
Sample mean with population mean	One Sample T-test	N/A
Means of two variables, measured on the same subject	Paired Sample T-test	Wilcoxon-Sign Test
Means of two independent groups	Independent Sample T-test	Mann-Whitney Test

Source – "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

The following graph shows the strengths and weaknesses of Pearson's correlation in each context

Figure 19. Result of Pearson's correlation

Value of r	Strength of correlation
Up to 0.30	Small (weak correlation)
Between 0.30 and 0.70	Medium (Moderate correlation)
Over 0.70	Large (Strong Correlation)

R>0	Direct orpositive correlation
R<0	Inverseornegative correlation

Source – "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Normal distribution. The normal distribution is also known as the Gaussian distribution or the Bell curve (Figure 20). It is one of the most widely used distributions due to the following factors:

- Due to the large number of sample size variables, that is ss the sample size increases, the distribution approaches a normal distribution.
- It is often used in the regression analysis.

  Examples of the normal distribution:
  - Biology, most of the biological dimensions are normally distributed. For example, height, arm and leg length, nail length, blood pressure, etc.
  - IQ test
  - Stock Market Information

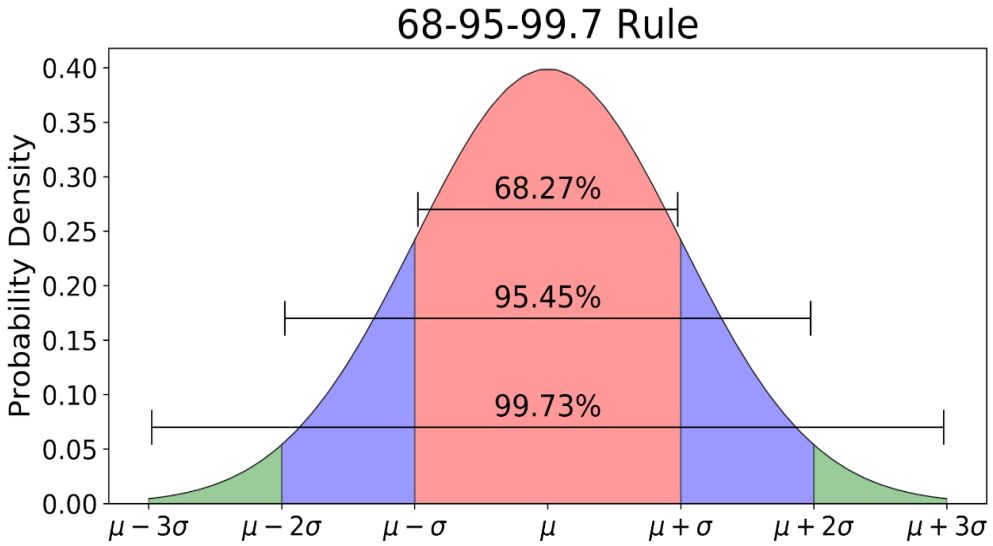
Figure 20. Normal Distribution

Source – "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Emprical Rule. "If the data distribution is bell-shaped, interval 1 constitutes 68% of the population or sample size." (Figure 21)

Standard Normal distribution. The standard normal distribution, also known as the Z distribution or the standard Gaussian distribution, is a specific form of the normal distribution. It is characterized by having a mean of zero and a standard deviation of one. By standardizing the values, it allows for comparison and calculations across different normal distributions. It covers 95% of the sample value and is between 2. It covers 99.7% of the sample value and is between.

Figure 21. Emprical Rule



Source - "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Standardization is the process of transforming a normal distribution to the standard normal distribution. This transformation involves subtracting the mean of the original distribution from each data point and dividing it by the standard deviation. The resulting standardized values, also called Z-scores, represent the number of standard deviations a particular data point is away from the mean of the distribution. (Figure 22).

$$z = \frac{x - \mu}{\sigma}$$

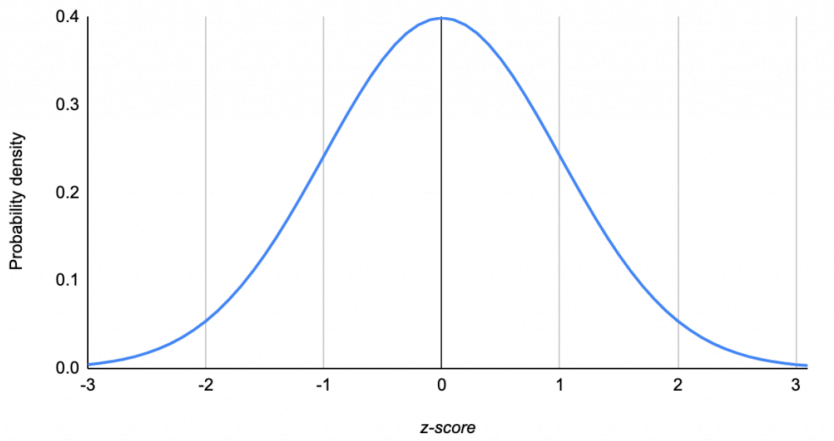
z - Z-score

x - The value being evaluated

μ - The mean

σ - The standard deviation

Figure 22. Standard Normal Distribution



Source - "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig

The standard letter Z indicates a variable with a normal distribution. May be wondering why need to standardize it (Figure 22). This is because standardization allows to do the following:

- Can compare different well-distributed datasets
- To calculate the average
- Define an external value
- Creating confidence intervals
- Probability test
- Perform regression analysis

Z score shows the deviation of the data points from the mean. Z score is in the range of [3: +3]. Identify extrema or edge values via Z score. If the value is less than 3 or greater than +3, a redundant value is displayed. The following table (Table1) corresponds to the situation where the z score is less than the negative average. In the case of the left tail.

The following table shows a scenario where the Z score exceeds the positive mean, that is, the case is positive (Table 2).

Application of machine learning algorithms in decision making. Machine learning is divided into three parts.

Supervised education

Unsupervised parenting

Reinforcement learning

Application of machine learning algorithms are using in decision making. Machine learning consists of three components:

1. Supervised learning: In supervised learning, the machine learning algorithm learns from labeled examples. It uses a dataset where each data point has input features and corresponding labels or target values. The algorithm learns a mapping function that can predict the correct label or target value for new, unseen data based on the patterns and relationships learned from the labeled training data. Examples of supervised learning algorithms include linear regression, decision trees, support vector machines (SVM), and neural networks.
2. Unsupervised learning: Unsupervised learning involves learning from unlabeled data. The algorithm analyses the data without any specific target variable or labels provided. Its goal is to discover underlying patterns, structures, or relationships in the data. Unsupervised learning techniques can be used for tasks such as clustering, dimensionality reduction, and anomaly detection. Common unsupervised learning algorithms include k-means clustering, hierarchical clustering, principal component analysis (PCA), and association rule mining.
3. Reinforcement learning: Reinforcement learning is a type of learning where an agent learns to make sequential decisions and take actions in an environment to maximize a cumulative reward signal. The agent interacts with the environment and receives feedback in the form of rewards or penalties based on its actions. Through trial and error, the agent learns to optimize his behaviour and make decisions that lead to

higher rewards over time. Reinforcement learning is commonly used in areas such as autonomous robotics, game playing, and optimization problems.

These three components encompass different learning paradigms and techniques in machine learning, allowing for a wide range of applications and problem-solving approaches.

Imagine you are a math class student and your teacher controls how the question is answered and whether the question is solved correctly. This representation is similar to a controlled learning algorithm. You can create an answer or solution for each record that appears in the record of interest. This is useful for learning the model and easily reveals the solution to the problem. Classification and regression algorithms are part of controlled learning. The classification method supports discrete value prediction. That is, it can be interpreted as information that belongs to a particular category or group. For example, suppose you have a dataset that contains images of fruits, and each image contains labels such as mango, apple, and so on. Here, the algorithm helps to classify new incoming image data.

Classification models are needed in situations where data needs to be categorized or classified into distinct classes or categories. These models are widely used in diverse fields such as pattern recognition, marketing, fraud detection, sentiment analysis, medical diagnosis, credit risk assessment, and email filtering, among others. There are an examples of a classification algorithm:

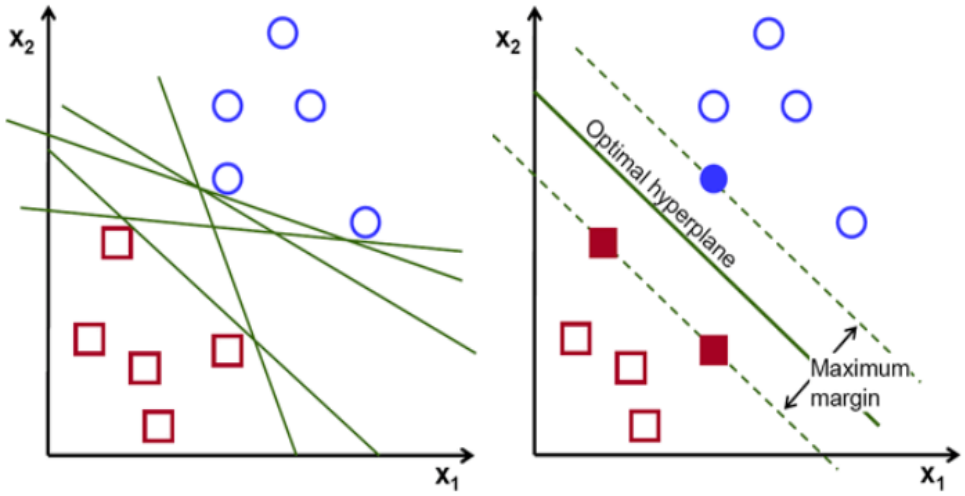
- Naive Bayes Classification
- Super Vector Machine
- Logistic regression

Naive Bayes Classification - Bayes is a collection of hypothetical facts. Bayes' Theorem is named after the 18th century British mathematician Thomas Bayes and is a mathematical formula for calculating conditional probabilities. Conditional probabilities are the probabilities that results will occur in response to past results that have occurred under similar circumstances. Bayes' theorem allows

us to rethink our current hypothesis or theory by presenting new or additional evidence. Naive Bayes assumes that the existence of a characteristic in a class is independent of the existence of other characteristics. For example, a red apple that is spherical in shape and 2 inches in size can be considered an apple. Despite their differences, each of these properties helps determine the likelihood that the fruit is an apple.

Support Vector Machines (SVMs) (Figure 23) are a versatile machine learning algorithm that can be applied to both classification and regression problems. While SVMs can be used for regression, they are predominantly used for classification tasks.

Figure 23. SVM



Source – "Support Vector Machines for Pattern Classification" by Shigeo Abe

The main objective of an SVM is to separate data points into different classes by creating hyperplanes in a high-dimensional space. In the case of binary classification, an SVM seeks to find the optimal hyperplane that maximally separates the two classes while maximizing the margin, which is the distance between the hyperplane and the closest data points from each class. SVMs achieve this by transforming the input data into a higher-dimensional feature space where it becomes easier to find a linear boundary between the classes. However, SVMs can also employ kernel functions to handle non-linearly separable data by implicitly mapping it into a higher-dimensional space, where linear separation is possible.

During the training phase, the SVM algorithm identifies a subset of the training data called support vectors, which are the data points closest to the decision boundary. These support vectors play a crucial role in defining the hyperplane and making predictions for new data points.

In the case of regression, SVMs can be used to approximate a continuous function that predicts a numeric value instead of class labels. The goal is to find a hyperplane that best fits the data points while minimizing the error between the predicted and actual values. Overall, SVMs are powerful and effective algorithms for both classification and regression tasks, offering flexibility and the ability to handle linear and non-linear relationships in the data.

Logistic regression is a statistical modelling technique used to predict the probability of an event or the occurrence of a binary outcome. It is commonly used when the target variable or dependent variable is binary or categorical, meaning it can take only two possible values or classes.

The goal of logistic regression is to model the relationship between the predictor variables (also known as independent variables or features) and the probability of the binary outcome. The output of a logistic regression model is a predicted probability, ranging from 0 to 1, which represents the likelihood of the event occurring (Figure 24).

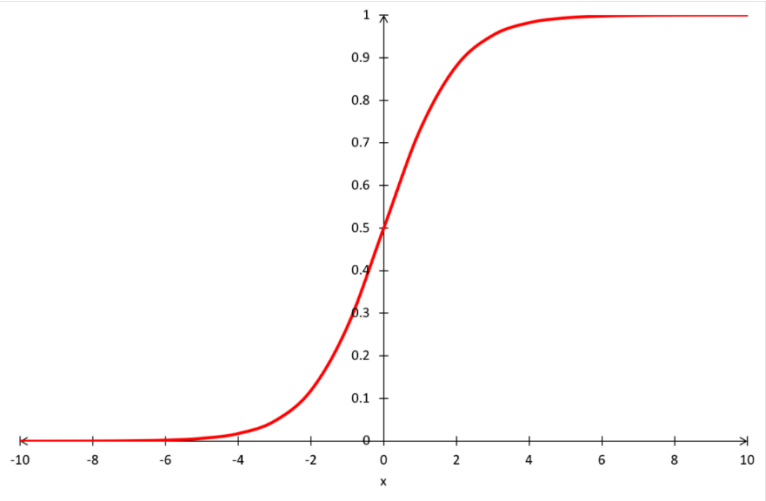
Unlike linear regression, which is used for predicting continuous numerical values, logistic regression employs a logistic function (also known as a sigmoid function) to map the linear combination of predictor variables to a probability value. This function allows the output to be confined within the range of 0 and 1.

Logistic regression is widely used in various domains, including healthcare, finance, marketing, and social sciences. It can be applied to solve classification problems, such as predicting whether a customer will churn or not, determining if an email is spam or not, or identifying whether a patient has a certain disease or not.

To fit a logistic regression model, the algorithm estimates the coefficients or weights of the predictor variables through a process called maximum likelihood

estimation. These coefficients indicate the strength and direction of the relationships between the predictors and the probability of the outcome.

Figure 24. Logistic Regression



Source – "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman

Regression algorithm - regression analysis is a statistical technique used to understand the relationship between a dependent variable and one or more independent variables. Regression analysis is a powerful tool for understanding and quantifying relationships between variables. Its applications range from prediction and forecasting to hypothesis testing and causal analysis. By using regression analysis, researchers, analysts, and decision-makers can gain valuable insights and make evidence-based decisions. The following algorithms are included in the regression.

- Linear Regression
- Nonlinear Regression
- Bayesian Linear Regression

Linear regression is the most important and most commonly used form of predictor. Regression analysis is used to explain the relationship between the dependent variable and many independent variables. (Figure 25) Linear regression can be categorized as simple or complex.

Simple regression is the simplest version of a regression equation with one dependent variable and one independent variable, defined as

$$y = kx + b$$

Where:

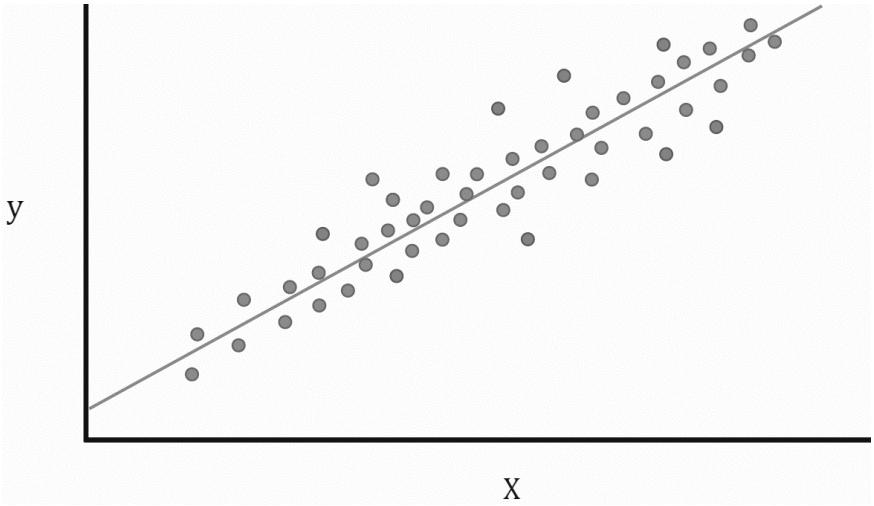
y - the estimated value,

b - is the constant value,

k - is the regression coefficient,

x - is the independent variable of the indicator.

Figure 25. Simple Linear Regression



Source – "An Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

Complex regression contains one dependent variable and two or more independent variables is an equation (Figure 26).

$$y = b + k_1 x_1 + k_2 x_2 + \cdots + k_n x_n$$

Where:

y - the estimated value,

b - is the constant value,

k - is the regression coefficient,

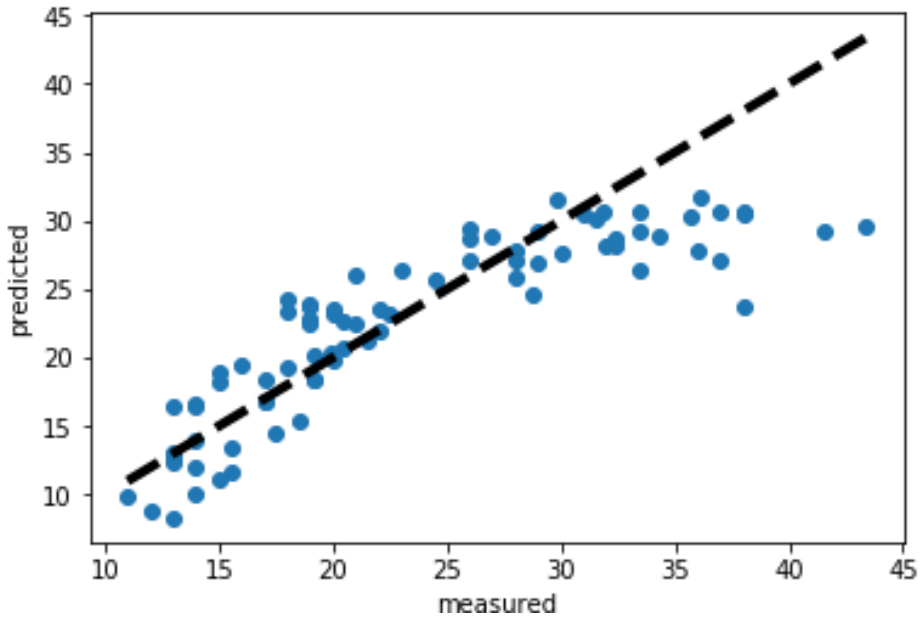
x - is the independent variable of the indicator.

Nonlinear regression is a powerful analytical method used to model the relationship between variables when the data does not follow a straight line. In contrast to simple linear regression, which assumes a linear relationship between variables, nonlinear regression allows for more complex relationships that can be

represented by curves, exponentials, logarithms, polynomials, or other nonlinear functions.

In nonlinear regression, the data is fitted to a mathematical model that describes the relationship between the variables. The goal is to estimate the parameters of the model that best fit the observed data. These parameters represent the coefficients or values that determine the shape and characteristics of the nonlinear curve.

Figure 26. Regression line



Source – "Applied Regression Analysis: A Research Tool" by John O. Rawlings, Sastry G. Pantula, and David A. Dickey

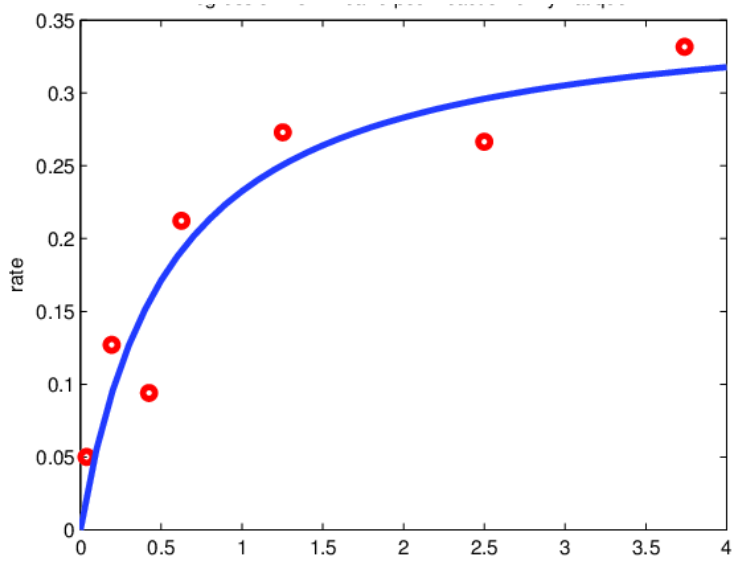
The process of fitting a nonlinear regression model typically involves iteratively adjusting the parameters to minimize the difference between the predicted values from the model and the actual observed values. This is done using optimization techniques such as the method of least squares or maximum likelihood estimation.

Nonlinear regression is particularly useful when the relationship between the variables is expected to exhibit complex patterns or when there is prior knowledge or theoretical understanding suggesting a specific nonlinear form. It allows for more

flexible modelling and can capture phenomena such as saturation, exponential growth, diminishing returns, and other nonlinear behaviours.

Nonlinear regression finds applications in various fields, including physics, biology, economics, engineering, and social sciences. It is commonly used for predicting outcomes, forecasting, trend analysis, modelling growth processes, and understanding complex systems. (Figure 27). As the name implies, unsupervised learning is the exact opposite of regulated learning. The main goal of unsupervised learning, also known as self-organizing learning, is the analysis and prediction of information.

Figure 27. Nonlinear regression



Source – "Nonlinear Regression Analysis and Its Applications" by Douglas M. Bates and Donald G. Watts

Unsupervised learning algorithms include:

1. K-means clustering: K-means is an unsupervised machine learning algorithm used for grouping similar data points into a specified number of clusters. It iteratively assigns data points to the nearest centroid and updates the centroids until convergence. K-means clustering is useful for tasks like customer segmentation, image compression, or anomaly detection.
2. Neural network: A neural network is a computational model inspired by the human brain's structure and function. It consists of

interconnected nodes (neurons) organized in layers. Neural networks can learn from labeled data to make predictions or classify new inputs. They are widely used for tasks like image recognition, natural language processing, and time series forecasting.

3. Principal Component Analysis (PCA): PCA is a dimensionality reduction technique used to identify the most important features or patterns in a dataset. It transforms the original variables into a new set of uncorrelated variables, called principal components, that capture the maximum variance in the data. PCA is helpful for reducing the dimensionality of high-dimensional data, visualizing data in lower-dimensional space, or removing redundant or noisy features.

Each of these techniques serves a specific purpose in data analysis and machine learning and can be applied depending on the nature of the problem and the characteristics of the data.

There is no difference between enhanced and uncontrolled learning. Here, the algorithm learns to react independently to the environment. In this way, the model simultaneously detects and trains new algorithms. Self-learning is mainly used in robots and video games.

CHAPTER II – APPLICATION OF DATA SCIENCE ALGORITHM IN DECISION MAKING ON THE CASE OF AMAZON

2.1 Familiarity with data and data cleaning.

We will build a logistics regression model on our data. The purpose of building our model is to determine the right target audience. Based on the sponsor's advertisement, we can conclude that the purchase of the product is successful among the elderly and the rich, but not successful among the young. The goal is to know our target audience, thereby maximizing click-through rates. Our goal is to show this ad to younger users less than we would to an older and wealthier audience, and to use that space to show ads that appeal to young users more. Our database contains information about the User ID, Gender, Age and Approximate Salary of all our users in the social network. The last column of the data set consists of a boolean column describing whether or not each person clicked on the ad (0 = False, 1 = True).

Figure 28. Raw Data

```
df1 = pd.read_csv('Social_Network_Ads.csv')
df1.head(10)
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15600575	Male	25	33000	0
9	15727311	Female	35	65000	0

Source – Developed by author

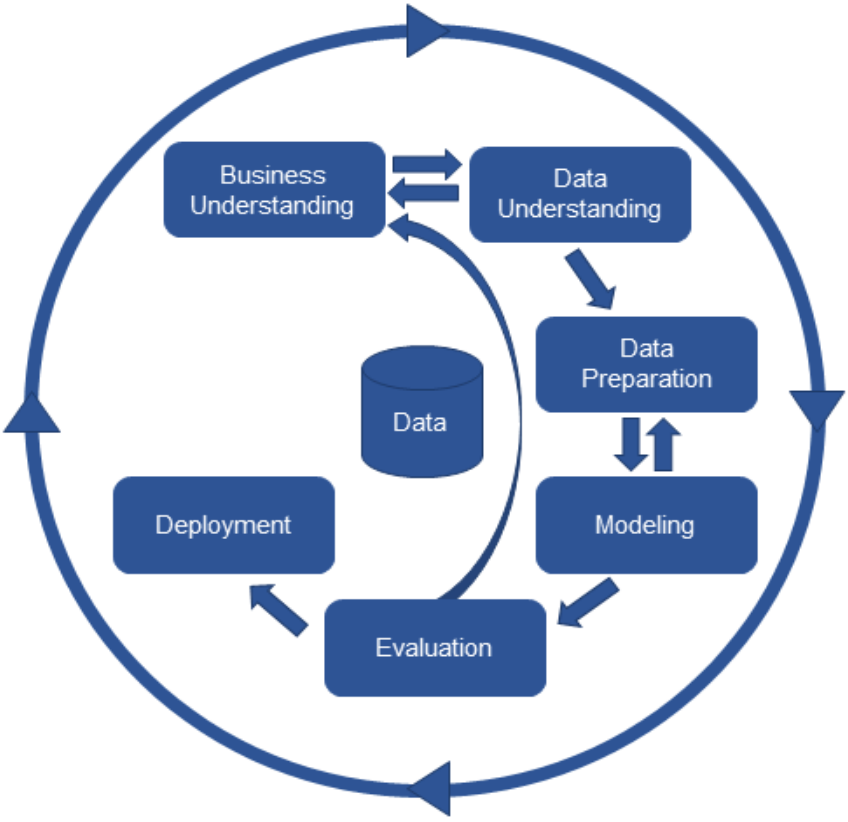
the data was entered in CSV format that it was used in the model. Because it is in CSV format, it is in the pandas library (used read_csv library). The point here

is that the data and script we enter must be in the same folder. To get acquainted with the data, when we run the code particle df1.head (), we can see the rows and columns of the data. (Figure 28) The Head function show the first 5 lines from the data by default. As it was written head (10) in the code, it will produce the first 10 lines.

If you look at the data, there is information about the customer's user ID, gender, age, income and whether they bought the product based on all this information.

According to the CRISP-DM Model (Figure 29) , data processing consists of 6 stages.

Figure 29. CRISP-DM Model



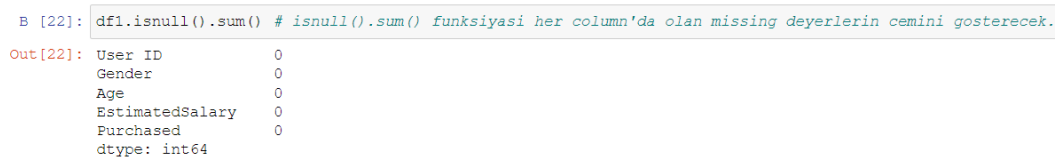
Source – "Data Mining: Concepts and Techniques" by Jiawei Han, Micheline Kamber, and Jian Pei

Understand the business: Gain a comprehensive understanding of the organization's goals, objectives, and processes. Identify key stakeholders and their requirements to align data analysis efforts with business needs.

1. Understand the data: Explore and assess the available data sources, including their quality, format, and relevance to the business problem. Identify any data gaps or limitations that may affect the analysis.
2. Data preparation: Cleanse, transform, and preprocess the data to ensure its accuracy, consistency, and suitability for analysis. This step may involve handling missing values, outliers, and data formatting issues.
3. Model building: Develop a model based on the prepared data. This may involve applying various statistical or machine learning techniques to uncover patterns, relationships, or predictions. Select appropriate algorithms and optimize their parameters to achieve the desired results.
4. Visualization: Create visual representations, such as charts, graphs, or dashboards, to effectively communicate insights from the data. Visualizations can help in understanding patterns, trends, and outliers, enabling stakeholders to make informed decisions.
5. Model deployment: Integrate the prepared model into the organization's infrastructure or application, allowing it to be utilized for decision-making or operational purposes. Monitor the model's performance and iterate as needed to ensure its continued effectiveness.

It's worth mentioning that these steps represent a high-level overview of the data analysis process and can vary depending on the specific problem, domain, and available resources.

Figure 30. Identify missing values



Source – Developed by author

We are currently in the Data Preparation phase, at which point we are identifying outliers (Figure 31), extreme values, blank values (Figure 30), and looking at how to fill and clear them. Other parameters are performed at this stage.

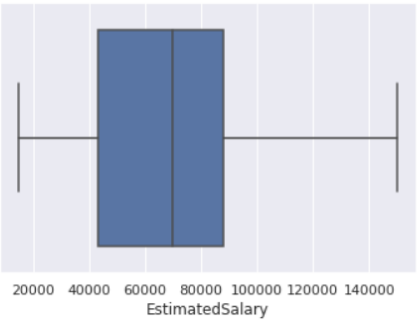
There are no extraneous values and missing values in the data. To build a model, was removed the "User ID" column from the model, thinking that it was not important. It was encoded the "Gender" column (Figure 32).

Figure 31. Identify outliers

```
B [24]: sns.boxplot(df1['EstimatedSalary']) #outlierleri mueyyenlesdirmek ucun usullardan  biride boxplot metodu ile
        # mueyyen etmekdir. EstimatedSalary column outlierlerin olmamasi asagida gosterilmisdir.

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable
word arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other argument
t an explicit keyword will result in an error or misinterpretation.
  FutureWarning

Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x7f46efa59110>
```



Source - Developed by author

Figure 32. User ID is removed from the column.

```
B [26]: df = df1.drop(['User ID'],axis=1)
        df
```

Out[26]:

	Gender	Age	EstimatedSalary	Purchased
0	Male	19	19000	0
1	Male	35	20000	0
2	Female	26	43000	0
3	Female	27	57000	0
4	Male	19	76000	0
...	...	...	...	...
395	Female	46	41000	1
396	Male	51	23000	1
397	Female	50	20000	1
398	Male	36	33000	0
399	Female	49	36000	1

400 rows × 4 columns

Source – Developed by author

Set the data to X and y values to build the model. In X, was included independent values in our data, such as the Age and Estimated Salary columns. IIt was assigned the column will predict to the value of y. The data is then divided into two parts: Train and Test. The sklearn.model_selection library is used to divide the Train and Test into sections. 75% of the data is taken as part of the Train, and 25% belongs to the Test part.

An image is a piece of code that we divide into train and test parts (Figure 33). Train is a set of information used to teach a part model. The test part is a set of data used to test an already trained model.

Figure 33. Train – test set

```
B [35]: from sklearn.model_selection import train_test_split
        X_train,X_test,y_train,y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
```

Source - Developed by author

2.2 Application of Machine Learning algorithm

The logistic regression algorithm requires a feature scaling method. It was used the sklearn.preprocessing library (Figure 34) to use this method. To make a Standard Scaler means to bring the data variables into the same price range. To do this, Python uses the transform function.

Figure 34. Standard Scaler

```
B [41]: from sklearn.preprocessing import StandardScaler
        sc = StandardScaler()
        X_train = sc.fit_transform(X_train)
        X_test = sc.transform(X_test)
```

Source – Developed by author

Figure 35. A piece of code to build a logistic regression

```
B [44]: from sklearn.linear_model import LogisticRegression
        classifier = LogisticRegression(random_state = 0)
        classifier.fit(X_train, y_train)# machine oyretmek ucun fit edilir
Out[44]: LogisticRegression(random_state=0)
```

Source – Developed by author

Figure 36. Coefficient and intercept

```
print( "coefficient -", log_reg.coef_, 'intercept -',log_reg.intercept_)
```

```
coefficient - [[2.24802196 1.12653177]] intercept - [-1.05821646]
```

Source – Developed by author

$$y = -1.058 + 2.248x_1 + 1.127x_2$$

If we put the coefficients and intersection points (Figure 36) we got into the formula, the formula we will get will be in this form. The predictions we will give will work on this formula.

$$y = \ln\left(\frac{P}{1-P}\right):$$

$$P = \frac{e^{-1.058+\ 2.248x_1+1.127x_2}}{1+e^{-1.058+\ 2.248x_1+1.127x_2}}$$

Where:

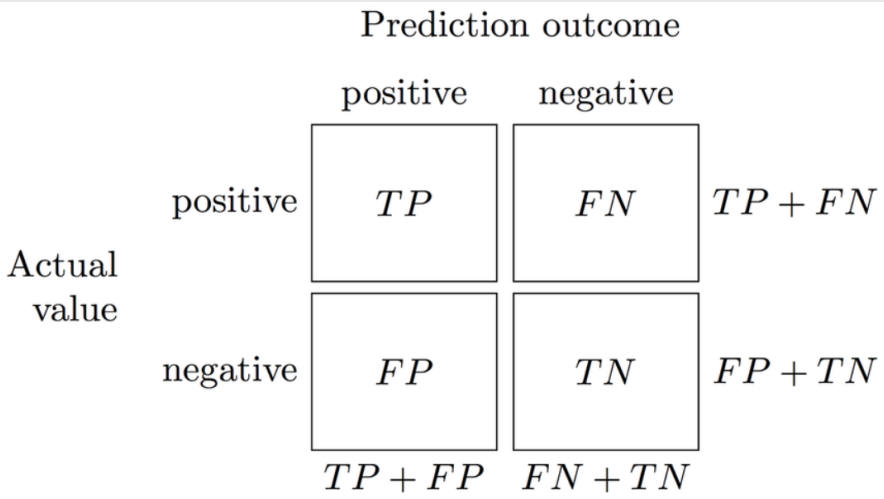
P - The probability of a 1 (the proportion of 1s, the mean of Y);

e - is the base of the natural logarithm (about 2.718).

Figure 37. Construction of Confusion Matrix

```
[48]: from PIL import Image
      img = Image.open('Confusion-matrix-for-a-two-class-problem-TP-is-the-number-of-correct-predictions-that-an.png')
      img
```

Out[48]:



Source – Developed by author

(Figure 37) Confusion Matrix - The error matrix is used as a matrix to compare the real and predictable values of the target mass to evaluate the performance of classification models. The classification forecast has 4 assessments:

1. TP - True Positive

2. TN - True Negative

3. FP - False Positive

4. FN - False Negative

Accuracy Score - achieved a result of 91% out of 100%, our result is good (Figure 38)

Figure 38. Confusion Matrix and accuracy score

```
B [51]: from sklearn.metrics import confusion_matrix, accuracy_score
        cm = confusion_matrix(y_test, y_pred)
        print(cm)
        accuracy_score(y_test, y_pred)# dustura gore cixarir 65 TN dir

        [[65  3]
         [ 6 26]]

Out[51]: 0.91
```

Source - Developed by author

To test the power of the model we built in the data, we divide the data into 3 parts (Figure 39) and look at the power of the model in each part. Our results are better than 50%.

Figure 39. Cross Validation Score

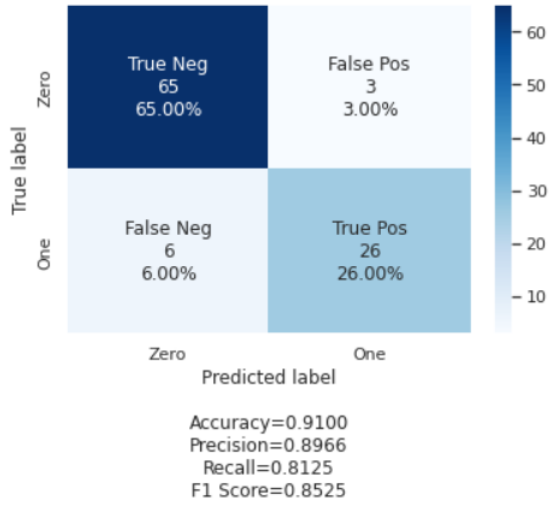
```
B [53]: from sklearn.model_selection import cross_val_score
        cross_val_score(classifier, X_train, y_train, cv=3, scoring="accuracy")

Out[53]: array([0.78, 0.76, 0.91])
```

Source – Developed by author

Figure 40. Roc curve score

```
B [55]: labels = ['True Neg','False Pos','False Neg','True Pos']
        categories = ['Zero', 'One']
        make_confusion_matrix(cm,
                              group_names=labels,
                              categories=categories,
                              )
```




```
B [59]: from sklearn.metrics import roc_auc_score
        roc_auc_score(y_test, y_pred)

Out[59]: 0.8841911764705883
```

Source - Developed by author

2.3 Results of Machine Learning algorithm

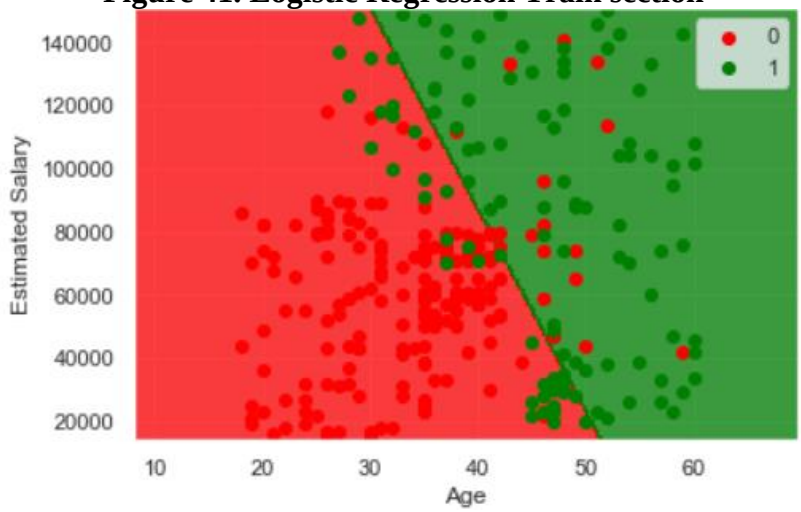
The figure (Figure 41) above helps us to the correlation between the dependent variable (probability of clicking on the name) and the independent variables (age and approximate salary) indicates that there is a positive relationship. As the age and approximate salary increase, the probability of a person clicking on the name also increases. This suggests that age and salary have a significant impact on the likelihood of clicking on the name.

Regarding the accuracy of observations, achieving approximately 80-90% accuracy is a positive outcome. It means that model is performing well in predicting the probability of clicking on the ad based on age and approximate salary.

However, was mentioned that result is close to 100%, which can be a cause for concern. A perfect 100% accuracy could indicate a potential overfitting issue, where the model has memorized the training data and may not generalize well to

new, unseen data. Overfitting can lead to overly optimistic performance and may not accurately reflect the model's true predictive ability.

Figure 41. Logistic Regression Train section



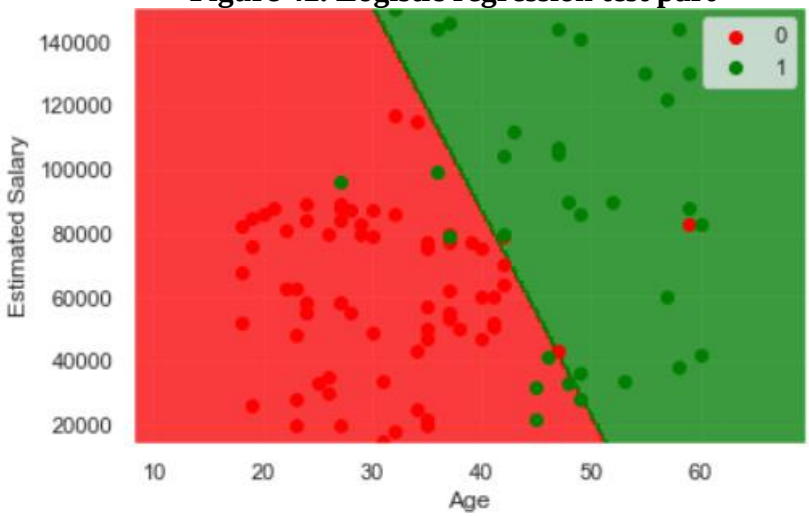
Source – Developed by author

Nevertheless, was mentioned that there is no overfitting problem in data, which is reassuring. It suggests that model has learned the underlying patterns and relationships without solely memorizing the training data. This indicates that model is performing well and is able to capture the correlation between age, salary, and the probability of clicking on the name.

Overall, achieving a result that is close to 100% accuracy is considered a very good result. It indicates that model is effective in predicting the likelihood of ad clicks based on age and salary. However, it is important to continue evaluating model's performance on new data and consider further validation techniques to ensure its generalizability and robustness.

In Figure 42, the test part of the model shows a similar pattern to the train part. This similarity suggests that the model is performing consistently and effectively on new, unseen data. As the number of observations increases, the information in the graph becomes more pronounced and aligns closely with the patterns observed in the train part.

Figure 42. Logistic regression test part



Source - Developed by author

The presence of red dots in the red and green regions of the model indicates instances where individuals who are not expected to click on the ad have actually viewed the product. However, was mentioned that these occurrences are relatively few. This suggests that the model is correctly identifying the majority of individuals who are unlikely to click on the name, as evidenced by the scarcity of misclassified instances in those regions.

The consistency between the train and test parts, along with the limited number of misclassifications, further supports the good performance of the model. It indicates that the model has learned meaningful patterns and generalizes well to new data.

It's important to continue monitoring and evaluating the model's performance over time. Assessing its accuracy, precision, recall, and other relevant metrics can help ensure that the model maintains its predictive power and effectively captures the relationship between the predictors (age and salary) and the probability of ad clicks.

That is, when users register on the website, we will know based on the information we have whether those people will click on the ad. Thus, we have defined our target audience.

CONCLUSION

The dissertation gives the answer to a scientific issue about the formulation of a marketing plan to enhance the effectiveness of a trading company`s marketing efforts on the virtual market, and draws the following findings.

1) Considering the notion of virtual market in the digital economy has enabled us to conclude that a virtual market is a special method of interaction between sellers and buyers in the virtual environment utilizing internet based information and communication technologies. It has been determined that the primary characteristics of the virtual market are: the operation of the subjects of interaction in the information space; accessibility for any buyer, regardless of their geographical location; the acceleration of exchange, administrative, and commercial operations between participants of virtual interaction; prompt response to customer inquiries; and the ability to complete online sales of goods.

2) The research of corporate marketing activities in virtual marketplaces revealed that conventional content is supplemented by new marketing instruments, such as the establishment of a website in the form of an electronic showcase or online shop, marketing on social networks, interactive buyer conversation, offer customisation; targeted, contextual, and media advertising. This allows for additional competitive advantages through instant data exchange, the attraction of more consumers, the reduction of transaction costs, the use of new methods to collect information about consumer preferences, the simplification of business processes, and the modelling of risk situations. Cyberattacks, risks connected with disruptions in the functioning of information services and providers, virtual fraud, data integrity breaches, the danger of exposure of personal customer information, low quality content, and SEO integrity are mentioned as hazards that arise in the course of marketing efforts in the Internet environment.

3) A marketing company's marketing strategy in the virtual market is a set of means to promote products and establish customer connections in the Internet environment. It allows me to have a more personalized relationship with target

audience-focused influence and consumer. Focuses on consumer communication using the latest internet marketing technology. And it ensures a positive image of the trading company and consumer confidence. Developing a company's marketing plan in the virtual market takes into account the company's activities in a virtual environment where the information space, communication space, and operation space are separated.

4) Pandemic situations have changed the way individuals shop and behave, and they are shopping more online.

5) The pandemic scenario re-emphasizes the importance and need for enterprises to monitor customer behaviour.

References

1. Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

2.– "Data Mining: Concepts and Techniques" by Jiawei Han, Micheline Kamber, and Jian Pei

3. "Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

4. "Pattern Recognition and Machine Learning" by Christopher Bishop

5. Herfeld C. The diversity of rational choice theory: A review note. TOPOI. An International Review of Philosophy, 39, 2020, p.329-347.

6. Karasek J., Musiał A., Gaponiuk K. New reality: the consumer in the age of COVID-19 How have the shopping habits of Poles changed during the coronavirus. 2020, p.1-28.

7. Lodziana G.J. Marketing determinants of consumer market behavior in Poland, CeDeWu, Warsaw 2015, p.28-38.

8. Malecka, M. Values in economics: A recent revival with a twist. Journal of Economic Methodology, 8(1), 2021, p.88-97.

9. Manosuthi N., Lee J. & Han H. Predicting the revisit intention of volunteer tourists using the merged model between the theory of planned behaviour and norm activation model. Journal of Travel & Tourism Marketing, 37(4), 2020, p.510-532.

10. Ostrowska I., Gracz L. Young buyers on e-shopping, Wydaw. PLACET, Warsaw. 2014: p.10-39.

11. Pallant J., Sands S., Karpen I. Product customization: A profile of consumer demand. Journal of Retailing and Consumer Services. 2020, T. 54. p.102030-102039

12. Wojcik P.Consumer psychography, Wydaw. Words and Thoughts, Lublin. 2017: p. 27-28.

13.https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

14. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
15. https://www.investopedia.com/terms/b/bayes-theorem.asp
16. https://www.pluralsight.com/blog/data-professional/data-informed-decisions
17. http://dx.doi.org/10.2139/ssrn.1819486
18. https://www.sciencedirect.com/science/article/abs/pii/S0167923621001173
19. https://betasciencepress-publishing.com/journals/data-science-and qaror-making/
20. https://www.dataversity.net/data-science-vs-decision-science-a-new-era dawns/
21. https://towardsdatascience.com/data-science-for-decision-makers 7248beddf948
22. https://www.analyticsvidhya.com/blog/2020/10/data-science-the-go-to tool-for-solving-daily-problems-taking-better-decisions/
23. https://ideas.repec.org/a/spr/jknowl/v12y2021i1d10.1007_s13132-016-0396-2.html
24. https://www.liebertpub.com/doi/10.1089/big.2013.1508
25. https://www.simplilearn.com/decision-analysis-principles-for-dataSciences-article
26. https://www.projectpro.io/article/data-scientist-vs-decision-scientist/171
27. https://thordrc.com/data-science-vs-machine-learning/
28. https://www.sqlservercentral.com/blogs/big-data-and-hadoop-part-1
29. https://philpapers.org/rec/HERTDO-19
30. https://www.investopedia.com/terms/b/bayes-theorem.asp
31. https://www.slideshare.net/statalex/statistical-analysis-training-course
32. https://en.wikipedia.org/wiki/P-value

Appendix

Table 1. Negative Z Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Source – Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

Table 2. Positive Z Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Source – Statistics for Business and Economics" by Paul Newbold, William L. Carlson, and Betty Thorne

List of Figures