**Machine learning (ML)** is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to learn from and make predictions or decisions based on input data, without being explicitly programmed to do so.

In other words, **Machine Learning** refers to the ability of the machine to learn from data, without being explicitly programmed to do so.

**Key Points**

1. Algorithms
2. Data

**Data**

Data is a collection of facts, figures, or other information that can be analysed to gain insights or make decisions.

Data can be categorised into:-

1. **Structured data**, refers to data that is organised in a specific format, such as a spreadsheet or database. Some common structured data formats include csv (Comma Separated Values), json (JavaScript Object Notation), xml (Extensible Markup Language), sql (Structured Query Language), excel, etc.
2. **Semi-structured data**, refers to data that contains some structure, but is not fully organised. Some common examples of semi-structured data formats includes log files, email, html, xml (flexibility to define tags), json (flexibility in defining attributes and values), etc
3. **Unstructured data**, refers to data that does not have a specific structure. Some common examples of unstructured data formats include text files, audio files, image files, social media, sensor data etc.

**Dataset** is a collection of data that is organised in a specific way for a particular purpose. A dataset can include various types of data, such as text, images, numbers, and other types of information.

Datasets can be used for a variety of purposes, such as research, analysis, and modelling. For example, a dataset could include data on sales transactions, customer demographics, or weather patterns.

To be useful, a dataset should be well-organised and contain a sufficient amount of data to support meaningful ***analysis*** to gain insights or to help in making decisions.

**Analytics**

Analytics refers to the process of using data and statistical techniques to gain insights and make decisions. It involves the collection, processing, and analysis of data in order to identify trends, patterns, and relationships.

Analytics can be used in a wide range of applications, including business, finance, healthcare, science, and more. The insights gained from analytics can help organisations make informed decisions and improve their operations.

**Types of Analytics**

1. **Descriptive analytics**: This type of analytics is used to describe what has happened in the past. It involves analysing historical data to gain insights into trends and patterns.
2. **Predictive analytics**: This type of analytics is used to predict what might happen in the future. It involves analysing historical data to identify patterns and make predictions about future outcomes.
3. **Prescriptive analytics**: This type of analytics is used to prescribe what actions should be taken to achieve a desired outcome. It involves analysing data to identify the best course of action to take.

**Applications of Data Analytics**

1. **In business**, data analytics is used to analyse customer behaviour, improve operations efficiency, identify customer needs and inform decision-making processes.
2. **In finance**, data analytics is used to identify trends and patterns in financial data, detect fraud, and manage risk.
3. **In healthcare**, data analytics is used to analyse patient data, develop treatment plans, and improve patient outcomes.
4. **In science**, data analytics is used to analyse experimental data, develop models, and make predictions.

**Data Analytics Tools and Statistical Techniques**

1. **Excel**: Excel is a widely used spreadsheet software that can be used for basic data analysis, including sorting, filtering, and creating charts and graphs.
2. **SQL**: SQL is a programming language used for managing and manipulating structured data in databases. It allows users to perform operations such as querying, inserting, updating, and deleting data.
3. **Tableau**: Tableau is a data visualisation tool that allows users to create interactive dashboards and visualisations from a variety of data sources.
4. **Python**: Python is a popular programming language used for data analysis and machine learning. It has a variety of libraries and frameworks, such as NumPy, Pandas, and Scikit-learn, that make it easier to analyse data and build models.
5. **R**: R is a programming language used for statistical analysis and data visualisation. It has a variety of packages, such as ggplot2 and dplyr, that make it easier to analyse and visualise data.
6. **Regression Analysis**: Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables.
7. **Cluster Analysis**: Cluster analysis is a technique used to group similar objects or observations together based on their characteristics.
8. **Time Series Analysis**: Time series analysis is a technique used to analyse time-dependent data, such as stock prices or weather patterns, to identify trends and patterns.
9. **Machine Learning**: Machine learning is a branch of artificial intelligence that focuses on building models that can learn from data and make predictions or decisions. It includes techniques such as decision trees, random forests, and neural networks.

**Job Titles in Data Analytics**

1. Data Analyst
2. Data Scientist
3. Business Analyst
4. Data Engineer

**Big Data**

Big Data refers to very large and complex sets of data that are difficult to manage and analyse using traditional tools and techniques. Despite the size, the data can come in many different forms and generated at a very fast pace.

**Characteristics of Big Data**

1. **Volume**: Big data size can range from terabytes to petabytes and beyond, and can include both structured and unstructured data.
2. **Velocity**: Big data is generated at an incredibly fast pace, and new data is constantly being added. This means that the speed at which data is generated and processed is a critical characteristic of big data. For example, social media platforms generate large amounts of data in real-time.
3. **Variety**: Big data can come in many different forms, including text, images, audio, video, and sensor data. This data can be structured, semi-structured, or unstructured, and may come from a variety of sources, such as social media, IoT devices and transactional databases.
4. **Veracity**: Big data may be unreliable, incomplete, or inconsistent, which can make it difficult to analyse and draw meaningful insights from. Ensuring data quality is a critical characteristic of big data, as inaccurate data can lead to incorrect conclusions and decisions.

**Tools for Big Data**

1. **Hadoop**: Hadoop is an open-source framework that allows for distributed processing and storage of large data sets across clusters of computers. It is designed to handle both structured and unstructured data and can be used for a variety of data processing tasks, such as data preparation, data transformation, and data analysis.
2. **Spark**: Spark is a fast and general-purpose distributed computing system that is designed for big data processing. It can process data in memory, making it much faster than traditional disk-based systems like Hadoop.
3. **NoSQL Databases**: NoSQL databases are non-relational databases that are designed to handle unstructured and semi-structured data. They can be used to store and retrieve large amounts of data quickly and efficiently.
4. **Data Visualisation Tools**: Data visualisation tools allow for the creation of visual representations of large data sets, making it easier to understand and analyse the data. Popular data visualisation tools include Tableau, QlikView, and Power BI.

**Data Analytics In Action**

Data Analytics involves the process of analysing data to gain insights to make informed decisions.

Here are the steps involved in data analytics:

1. **Define the problem**: Identify the problem you want to solve or the question you want to answer with data analytics. This will help you focus your efforts and guide the rest of the process. This involves stating the problem, scope, hypothesis, data sources, analysis techniques that will be used, and the stating the conclusion and what might be the recommendations
2. **Collect and clean the data**: Collect the data that you need to answer your question, and then clean it by removing any errors, inconsistencies, or missing values. This will ensure that your analysis is accurate and reliable.
3. **Explore the data**: Use exploratory data analysis techniques to get a better understanding of your data. This can include visualisations, summary statistics, and clustering.
4. **Analyse the data**: Use statistical analysis techniques to identify patterns, relationships, and trends in the data. This can include hypothesis testing, correlation analysis, and regression analysis.
5. **Visualise the results**: Use data visualisation techniques to communicate your findings in a clear and concise manner. This can include charts, graphs, and dashboards.
6. **Interpret the results**: Once you have analysed the data and visualise the results, interpret the findings to understand the implications for your problem or question. This can include making recommendations or identifying areas for further investigation.
7. **Communicate the results**: Finally, communicate your findings to stakeholders in a clear and concise manner. This can include visualisations, dashboards, or reports.

**Data Collection Techniques**

Refers to the methods used to gather data from various sources. There are several techniques used for data collection depending on the problem, type of data, data quality, sample size, resources, etc.

Here are some common data collection techniques:

1. **Surveys**: Surveys are a common method of collecting data from individuals or groups. They can be conducted in-person, over the phone, or online.
2. **Observations**: Observations involve gathering data by observing people, objects, or events in their natural environment. Observations can be recorded manually or through automated sensors.
3. **Experiments**: Experiments involve manipulating one or more variables to observe the effect on a particular outcome. Data can be collected through controlled experiments in a laboratory setting or in the field.
4. **Interviews**: Interviews involve gathering data through one-on-one conversations with individuals or groups. They can be conducted in-person or remotely.
5. **Web Scraping**: Web scraping involves using automated tools to extract data from websites. This data can be used for market research, competitor analysis, and other applications.
6. **Social Media Monitoring**: Social media monitoring involves collecting data from various social media platforms to gain insights into customer behaviour, brand sentiment, and industry trends.
7. **Sensor Data Collection**: Sensors can be used to collect data on a variety of physical parameters such as temperature, humidity, pressure, and motion. This data can be used for environmental monitoring, building automation, and other applications.

**Data Cleaning Techniques**

Refers to the process of identifying and correcting errors, inconsistencies, and inaccuracies in data. It is important to carefully clean and prepare the data before conducting any analysis to ensure that the results are reliable and accurate. The data cleaning techniques used depend on the nature of the data and the problem.

Here are some common data cleaning techniques:

1. **Removing Duplicates**: This involves identifying and removing any duplicate records or observations in the data.
2. **Remove Irrelevant Data**. Removing any unnecessary data that adds no meaning to the data.
3. **Handling Missing Data**: Identifying any missing values in the data and deciding how to handle them. Options include imputing missing values using statistical methods or removing observations with missing data.
4. **Correcting Errors**: Identifying and correcting errors in the data, such as typos or incorrect values.
5. **Formatting Data:** Formatting data in a way that is acceptable by the analysis techniques that you have chosen without changing the meaning.
6. **Standardising Data**: Ensuring that data is consistently formatted and labelled throughout the dataset.
7. **Handling Outliers**: Identifying and handling any outliers in the data. This can include removing outliers, transforming the data, or using robust statistical methods that are less sensitive to outliers.
8. **Data Validation**: Verifying that the data is accurate and complete. This can involve cross-checking the data against other sources or conducting additional data collection to fill in any gaps.

Examples on some of these techniques can be found on this link [Google Colab](https://colab.research.google.com/drive/1jfS5vp2bcJukHl2UufparC0N4A4bAkFL?usp=sharing)

An Example of Data Analysis [Zoom Tanzania Used Cars](https://github.com/shebogholo/data_analysis_with_python)

**Algorithm**

An algorithm is a set of instructions or rules that a computer follows to solve a specific problem or complete a task. Algorithms are used in a wide range of applications, from simple calculations to complex artificial intelligence systems.

An algorithm typically consists of a series of steps that are executed in a specific order to achieve a desired outcome. The steps of an algorithm may involve performing calculations, making decisions based on specific criteria, or executing specific operations.

Algorithms can be classified into many different categories, depending on the specific problem they are designed to solve. For example, a sorting algorithm is designed to organise a list of items in a specific order, while a search algorithm is designed to find a specific item in a list.

In recent years, Machine Learning Algorithms have become increasingly popular. These algorithms are used to analyse large datasets and identify patterns, making them useful for applications such as image recognition, speech recognition, and natural language processing.

**Machine Learning Algorithms**

1. **Supervised Learning**: This is where the algorithm is trained on labelled data, meaning that the correct output is already known. The algorithm tries to learn a mapping between the input features and the correct output.
2. **Unsupervised Learning**: This is where the algorithm is trained on unlabeled data, meaning that there is no correct output. The algorithm tries to find patterns or structure in the data on its own.
3. **Reinforcement Learning**: This is where the algorithm learns by trial and error, receiving feedback in the form of rewards or penalties based on its actions.

**Supervised Learning Algorithms**

The goal of supervised learning is to learn a mapping between the input features and the correct output.

Here are some common types of supervised learning algorithms:

1. **Linear Regression**: This is a regression algorithm that tries to fit a line to the data by minimising the sum of squared errors between the predicted output and the actual output.
2. **Logistic Regression**: This is a classification algorithm that tries to fit a logistic function to the data by maximising the likelihood of the correct class labels given the input features.
3. **Decision Trees**: This is a classification algorithm that creates a tree-like model of decisions and their possible consequences based on the input features.
4. **Random Forest**: This is an ensemble learning algorithm that combines multiple decision trees to make a final prediction.
5. **K-nearest Neighbours**: This is a classification algorithm that classifies new data points based on the class labels of the k-nearest neighbours in the training data.
6. **Support Vector Machines**: This is a classification algorithm that finds the hyperplane that maximally separates the classes in the data.
7. **Neural Networks**: This is a type of machine learning that uses artificial neural networks with multiple layers to learn complex patterns in data.
8. **Naive Bayes**: This is a probabilistic algorithm that assumes that the features in the data are independent, and calculates the probability of each class given the input features.
9. **Gradient Boosting**: This is an ensemble learning algorithm that combines multiple weak models to make a final prediction, by iteratively fitting new models to the residual errors of the previous models.

**Unsupervised Learning Algorithms**

The goal of unsupervised learning is to find patterns or structure in the data on its own.

Here are some common types of unsupervised learning algorithms:

1. **Clustering**: This is a family of algorithms that group similar data points together based on some similarity measure, such as distance or density. Examples include k-means clustering, hierarchical clustering, and DBSCAN.
2. **Principal Component Analysis (PCA)**: This is a dimensionality reduction algorithm that projects the data onto a lower-dimensional space while retaining the maximum amount of variance in the data.
3. **Singular Value Decomposition (SVD)**: This is a matrix factorization technique that decomposes a matrix into its constituent singular values and vectors, which can be used for a variety of tasks, including matrix approximation, data compression, and feature extraction.
4. **Independent Component Analysis (ICA)**: This is a technique for separating a multivariate signal into its independent, non-Gaussian components.
5. **Autoencoders**: This is a type of neural network that learns to encode the input data into a lower-dimensional representation and then decode it back to the original input, while minimising the reconstruction error.
6. **Generative Adversarial Networks (GANs)**: This is a type of deep learning model that consists of a generator and a discriminator that play a minimax game to learn to generate realistic samples from a given distribution.
7. **Anomaly Detection**: This is a family of algorithms that try to identify rare or unusual data points that deviate significantly from the rest of the data.
8. **Association Rule Mining**: This is a technique for finding interesting relationships between items in a dataset, such as frequent itemsets for association rules.
9. **Density Estimation**: This is a family of algorithms that try to estimate the underlying probability distribution of the data, which can be used for tasks such as anomaly detection, clustering, and generation of new data. Examples include Gaussian mixture models, kernel density estimation, and variational autoencoders.

**Reinforcement Learning Algorithms**

The goal of unsupervised learning is to learn to take actions in an environment to maximise a reward signal. Here are some popular reinforcement learning algorithms:

1. **Q-Learning**: Q-learning is a model-free, off-policy algorithm that seeks to find the optimal action-selection policy for any given environment. It uses a table of values (called Q-values) that are updated as the agent interacts with the environment to learn the best action to take in each state.
2. **Deep Q-Networks (DQN)**: DQN is an extension of Q-learning that uses deep neural networks to approximate the Q-values instead of a lookup table. It has been shown to perform well on a range of complex tasks.
3. **Policy Gradient Methods**: Policy gradient methods learn a policy directly, rather than estimating a value function. They are typically model-free and can learn both discrete and continuous actions. One popular policy gradient method is REINFORCE.
4. **Actor-Critic Methods**: Actor-critic methods combine elements of policy gradient methods with value-based methods. They use an actor to select actions and a critic to estimate the value of state-action pairs. One popular actor-critic method is Advantage Actor-Critic (A2C).
5. **Deep Deterministic Policy Gradient (DDPG)**: DDPG is an actor-critic method that can learn policies for continuous action spaces. It uses a neural network to represent the actor and the critic and has been shown to perform well on a range of continuous control tasks.
6. **Monte Carlo Tree Search (MCTS)**: MCTS is a tree search algorithm that builds a search tree by simulating the environment and using a selection strategy to choose the next action. It has been successful in a range of games, including Go and Chess.

**Steps on Machine Learning**

The general steps in a typical machine learning pipeline are as follows:-

1. **Problem Definition**: The first step in any machine learning project is to clearly define the problem you want to solve. This involves identifying the business problem or opportunity, defining the scope of the project, and establishing measurable success criteria.
2. **Data Collection:** The second step is to collect relevant data for the machine learning model. This includes identifying the sources of data, obtaining the data, and cleaning the data to ensure it is accurate, complete, and consistent.
3. **Data Preparation:** The third step is to prepare the data for machine learning. This involves transforming the data into a format that can be used by the machine learning algorithms. This may include data preprocessing, feature selection, feature engineering, and data splitting into training, validating and testing datasets.
4. **Model Selection:** The fourth step is to select an appropriate machine learning model for the problem at hand. This involves selecting the appropriate type of model, such as supervised or unsupervised learning, and selecting the appropriate algorithm for the problem.
5. **Model Training:** The fifth step is to train the machine learning model using the training dataset. This involves selecting the appropriate hyperparameters for the algorithm, tuning the model to improve performance, and validating the model using the validating dataset and testing using the testing dataset.
6. **Model Evaluation:** The sixth step is to evaluate the performance of the machine learning model. This involves measuring the model's accuracy, precision, recall, F1 score, and other performance metrics to determine whether it meets the success criteria.
7. **Model Deployment:** The final step is to deploy the machine learning model in a production environment. This involves integrating the model into the existing systems, monitoring the model's performance, and making adjustments as necessary to ensure it continues to meet the business needs.

Example of Machine Learning in Action [https://github.com/shebogholo/Machine-Learning-Algorithms](https://github.com/shebogholo/Machine-Learning-Algorithms/blob/master/Machine%20Learning%20Workflow.ipynb)

**Deep Learning**

Deep Learning is a subset of machine learning that focuses on creating neural networks capable of learning from and making decisions based on large amounts of data.

**Types of Neural Network**

1. Feedforward Neural Networks
2. Convolutional Neural Networks (CNNs)
3. Recurrent Neural Networks (RNNs)
4. Generative Adversarial Networks (GANs)

**Categories of Deep Learning**

1. Computer Vision
2. Natural Language Processing

**Applications of Computer Vision**

1. **Object Recognition and Detection:** Computer vision can be used to identify objects within images or video streams. This can be used in a variety of applications, such as facial recognition, licence plate recognition, and identifying products on a store shelf.
2. **Image and Video Analysis:** Computer vision can be used to analyse images and videos to identify patterns, trends, and anomalies. This can be useful for monitoring traffic, detecting medical conditions, or identifying defects in manufacturing.
3. **Autonomous Vehicles:** Computer vision is a critical component of self-driving cars and other autonomous vehicles. Cameras and sensors are used to identify objects in the vehicle's environment and make decisions based on that information.
4. **Augmented Reality:** Computer vision can be used to overlay digital information on the real world, creating augmented reality experiences. This can be used in applications like gaming, advertising, and industrial design.
5. **Robotics:** Computer vision is used in robotics to help robots navigate and interact with their environment. Robots can use cameras and sensors to identify objects and make decisions based on that information.
6. **Medical Imaging:** Computer vision is used in medical imaging to help diagnose and treat medical conditions. It can be used to analyse X-rays, MRIs, and other medical images to identify anomalies and predict outcomes.

**Applications of Natural Language Processing**

1. **Sentiment Analysis:** Sentiment analysis is the process of analysing text to determine the sentiment or emotion behind it. NLP algorithms can be used to analyse social media posts, customer reviews, and other forms of text to determine whether they are positive, negative, or neutral.
2. **Chatbots:** Chatbots are computer programs that are designed to simulate human conversation. NLP algorithms are used to understand user inputs and generate appropriate responses. Chatbots can be used for customer service, sales, and other applications.
3. **Language Translation:** NLP algorithms can be used to translate text from one language to another. This can be useful for businesses that operate in multiple countries or for individuals who need to communicate in different languages.
4. **Information Extraction:** NLP algorithms can be used to extract useful information from unstructured data sources, such as news articles, research papers, and social media posts. This can be used for a variety of applications, such as market research, competitive intelligence, and trend analysis.
5. **Speech Recognition:** NLP algorithms can be used to convert spoken language into text. This can be useful for applications such as voice assistants, dictation software, and automated call centres.
6. **Text Summarization:** NLP algorithms can be used to summarise large volumes of text into shorter, more digestible summaries. This can be useful for news articles, research papers, and other types of long-form content.