**Sheza Nauman**

**assignment**

**Artificial intelligence**

**1)Machine learning Vs Data science:-**

The study of data and how to get meaning from it is known as **data science**. This discipline employs a variety of techniques, algorithms, systems, and tools for extracting information from both structured and unstructured data. This knowledge is then utilized by organizations in business, government, and other sectors to increase revenue, develop new goods and services, improve infrastructure and public processes, among other things. whereas **Machine learning** is a branch of artificial intelligence that uses algorithms to extract data and then predict future trends. Engineers can run statistical analysis on data using software that has been programmed with models that help them identify trends in the data. Information on users is gathered by social media sites including Facebook, Twitter, Instagram, YouTube, and TikTok. It predicts your interests and desires based on past behaviour and suggests goods, services, or articles that are pertinent to you. Machine learning is used in data science but also emerges in domains outside of it as a set of concepts and technologies. When it makes sense, data scientists frequently use machine learning in their work to accelerate the collection of information or to aid with trend analysis.

|  |  |
| --- | --- |
| **Characteristics of ML** | **Characteristics of DS** |
| Algorithm-Centric: The main focus of machine learning (ML) is developing and improving algorithms that can learn from data. These algorithms are in charge of classifying or making predictions depending on the incoming data.  ML includes both supervised learning (where models are trained on labelled data) and unsupervised learning (where models find patterns in unlabeled data).  Narrow Focus: ML models are designed with a specific objective in mind, such as recommendation systems, picture recognition, or natural language processing.  Model Evaluation: Depending on the nature of the issue, metrics including as accuracy, precision, recall, and F1-score are used to gauge how well ML models perform.  Iterative Training: ML models are trained on data repeatedly, and when more data and optimisation are applied, the models' performance gets better. | Data-Centric: The main goals of data science are to gather, purify, and explore data in order to produce useful insights. We call this procedure "data wrangling."  The entire data lifecycle—from data collection and preprocessing to visualisation and storytelling—is involved in data science.  Domain Knowledge: In order to contextualise their findings and offer sound advice, data scientists frequently need domain knowledge.  Impact on corporate: The ultimate objective of data science is to deliver actionable insights that can support corporate decisions, streamline procedures, or resolve challenging issues.  Tool Diversity: Depending on the issue at hand, data scientists employ a variety of tools and techniques, such as statistics, data mining, data visualization, and machine learning. |

**Comparison:-**

A specialized area of data science called machine learning that focuses on developing prediction models is its scope. In addition to data collecting, cleansing, and analysis on a larger scale, data science also includes machine learning.

**Objective:** While data science seeks to gather insights, address issues, and influence decision-making, machine learning's main objective is prediction or categorization.

**Process:** Model training and evaluation are part of ML, which is more algorithm-centric. Data acquisition, exploration, and dissemination are all components of the comprehensive approach used by data science.

**expertise:** A combination of technical and domain-specific expertise is frequently required of data scientists. ML engineers and researchers, in contrast, concentrate more on the creation and improvement of algorithms.

**Outputs:** While data science creates reports, visualizations, and useful insights, machine learning often creates prediction models.

**2) Data science VS Artificial intelligence**:-

1. Data science involves pre-processing, analysis, prediction, and visualization, which is a big difference. AI is the application of a model that predicts future events.

1. A broad term for statistical approaches, design procedures, and development methodologies is data science. Algorithm design, development, effectiveness, conversions, and deployment are all aspects of artificial intelligence.
2. In contrast to TensorFlow, Kaffee, and scikit-learn, which are utilised in AI, Python and R are the tools used in data science. Data analysis and data analytics, which use past and present data to anticipate future data, are the main topics of data science. Machine learning is a topic in artificial intelligence.
3. The goal of data science is to uncover underlying patterns and trends in data. The discipline's goal is to gather usable data, process it, interpret it, and then apply it to arrive at significant conclusions. On the other hand, artificial intelligence is employed to manage data on its own, freeing up the human from any further involvement in the process.
4. Complex models can be created utilizing data science to extract numerous information, statistical methods, and insights. On the other side, artificial intelligence is designed to create models that, to a certain extent, mimic human intellect and comprehension. The goal is to achieve self-sufficiency, which would mean that the machine would no longer require any human input, via simulating cognition.

**Problem statement for DS**:-

* To increase sales, use future demand forecasting.
* To effectively manage campaigns through consumer segmentation.
* To detect and prevent fraud.

**Problem statement for AI**:-

* Devices that recognize speech, such as Apple's Siri and Amazon's Alexa
* engines for recommendations: Netflix watch suggestions and Amazon product catalogue suggestions

**Reasons of using DS:-**

There are several compelling reasons for using data science in various fields and industries. Here are some of the key reasons:

1. Informed Decision-Making
2. Predictive Analytics
3. Efficiency and Optimization
4. Customer Understanding: Competitive Advantage
5. Product and Service Improvement
6. Risk Management
7. Healthcare and Research
8. Environmental and Sustainability Impact
9. Personal and Professional Development
10. Education and E-Learning
11. Humanitarian and Social Good:

Overall, data science is a powerful tool that can generate insights, drive innovation, and improve decision-making in a wide range of domains. Its versatility and ability to extract valuable information from data make it a critical asset for organizations and individuals in today's data-driven world.

**Reasons of using AI:-**Artificial Intelligence (AI) is used for a variety of reasons across different industries and domains due to its versatility and potential. Here are some key reasons for using AI:

1. Automation
2. Data Analysis
3. Predictive Analytics
4. Personalization.
5. Natural Language Processing (NLP.)
6. Image and Video Analysis
7. Healthcare
8. Fraud Detection:
9. Recommendation Systems
10. Customer Support
11. Autonomous Systems
12. Scientific Research
13. Education
14. Energy Efficiency
15. Space Exploration
16. National Security Entertainment and Gaming Humanitarian Efforts

In summary, AI is adopted across various sectors to improve efficiency, decision-making, customer experiences, and innovation. Its ability to process data, learn from it, and adapt makes it a transformative technology with widespread applications.

|  |  |
| --- | --- |
| **Advantages of DS** | **Disadvantages of DS** |
| Data science aids organizations in making data-driven decisions, which can result in better strategies and better results.  Data scientists can find important insights and trends in data that would be overlooked without their knowledge, giving organizations a competitive edge.  Predictive Analytics: Data science enables organizations to predict upcoming patterns and outcomes, assisting in proactive planning and risk reduction.  Personalization: Businesses can utilize data science to customize goods, services, and content to suit different customer preferences, which will increase client pleasure.  Efficiency Gains: Automating and optimizing based on data can result in cost savings and streamlined procedures.  Enhanced Customer Experience: By examining consumer input and behaviour to address concerns and preferences, data science may enhance customer service. | Data Quality: Data science heavily relies on data quality, and if data is inaccurate, incomplete, or biased, it can lead to misleading results.  Knowledge Gap: Because of the strong demand for skilled data scientists, it might be difficult to attract and keep talent.  Privacy Concerns: Collecting and analyzing vast amounts of data can raise privacy concerns, and businesses must navigate regulatory challenges.  Costly Infrastructure: Building and maintaining data infrastructure can be expensive, especially for small businesses.  Skill |

|  |  |
| --- | --- |
| **Advantages of AI** | **Disadvantages of AI** |
| Automation: AI can perform routine jobs more effectively and with less human error.  Scalability: AI systems have the capacity to process and analyze large datasets at a rate that is faster than what is possible for humans.  Personalization: Users may receive highly customized content and product recommendations from AI-powered recommendation systems.  24/7 Availability: AI applications can run continuously, offering support and service.  AI can address complicated problems that are challenging for people to solve, such as picture recognition and natural language processing.  Safety: AI can be used to enhance safety in risky contexts like driverless vehicles or hazardous industrial settings. | AI-powered automation may result in employment displacement in some industries, posing economic and social difficulties.  Fairness and Bias: AI systems may produce unfair or discriminating results if their training data contains biases.  Lack of imagination: AI's ability to solve problems is constrained since it lacks human imagination and the capacity to think beyond its training data.  Costly to Develop and Maintain: Creating and maintaining AI systems may be pricey, especially for intricate tasks.  AI presents ethical concerns regarding decision-making openness, privacy, and the possible abuse of cutting-edge technology.  Technical Difficulties: Building and optimising AI models can be technically difficult, requiring a lot of knowledge and computer power. |

**Conclusion:-**

In conclusion, data science and AI have many advantages, but they also have a unique set of difficulties and drawbacks. When incorporating data science and AI into their goals and operations, organizations must carefully take into account these considerations.

**3) Machine learning types:-**

The ability for a machine to automatically learn from data, enhance performance based on prior experiences, and make predictions is known as machine learning. A collection of algorithms used in machine learning operate on vast amounts of data. These algorithms are fed data to train them, and after training, they develop a model and carry out a certain task. These ML techniques support the resolution of numerous business issues, including clustering, associations, forecasting, classification, regression, and others.Machine learning is primarily split into four types based on the techniques and modes of learning, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Machine learning

**Supervised Machine learning:-**

Supervised machine learning is based on supervision, as its name suggests. In the supervised learning technique, this means that we train the machines using the "labeled" dataset, and then the machine predicts the output based on the training. Here, the marked data indicates which inputs have already been mapped to which output. More precisely, we may state that after training the machine with input and related output, we ask it to predict the outcome using test dataset.

Supervised machine learning can be classified into two types of problems, which are given below:

* Classification
* Regression

**a) Classification**

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are Spam Detection, Email filtering, etc.

Some popular classification algorithms are given below:

* Random Forest Algorithm
* Decision Tree Algorithm
* Logistic Regression Algorithm
* Support Vector Machine Algorithm

**b)** **Regression**

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc. Some popular Regression algorithms are given below:

* Simple Linear Regression Algorithm
* Multivariate Regression Algorithm
* Decision Tree Algorithm
* Lasso Regression

**Unsupervised Machine Learning:-**

Unsupervised learning is distinct from the supervised learning method because, as its name implies, supervision is not required. In unsupervised machine learning, this means that the system is trained on an unlabeled dataset and makes output predictions without any human supervision. In unsupervised learning, the models are trained on data that has neither been categorized nor labeled, and they are then allowed to behave autonomously on that data. The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

Unsupervised Learning can be further classified into two types, which are given below:

* Clustering
* Association

**Clustering:-**

When looking for the innate groups in the data, we employ the clustering technique. It is a method of clustering the items so that those that share the most similarities stay in one group and share little to none in common with those in other groups. Grouping clients based on their purchase habits is an illustration of the clustering method. Some of the popular clustering algorithms are given below:

* K-Means Clustering algorithm
* Mean-shift algorithm
* DBSCAN Algorithm
* Principal Component Analysis
* Independent Component Analysis

**Association:-**

An unsupervised learning method called association rule learning identifies intriguing relationships between variables in a sizable dataset. This learning algorithm's primary goal is to identify the dependencies between data items and then map the variables in a way that maximizes profit. This algorithm is mostly used in continuous production, web usage mining, market basket analysis, etc. Apriori , Eclat, and FP-growth algorithms are a few of the well-known algorithms for learning association rules.

**Semi supervised machine learning:-**

Between supervised and unsupervised machine learning, there is a form of method known as semi-supervised learning. It uses a combination of labeled and unlabeled datasets during the training phase and stands in the middle of supervised learning (with labeled training data) and unsupervised learning (without labeled training data) techniques. While semi-supervised learning acts on data that contains a few labels and is a middle ground between supervised and unsupervised learning, the majority of the data it uses is unlabeled. Labels are expensive, however for corporate needs, there might not be many labels. Because supervised and unsupervised learning are dependent on the presence or lack of labels, it is entirely distinct from those methods.

The idea of semi-supervised learning is presented in order to address the shortcomings of supervised learning and unsupervised learning methods. Instead of using solely labeled data as in supervised learning, the primary goal of semi-supervised learning is to make optimal use of all accessible data. An unsupervised learning technique is first used to cluster comparable data, and it also aids in labeling the unlabeled data into labeled data. The reason behind this is that labeled data is more expensive to acquire than unlabeled data.

**Reinforcement Machine learning:-**

With reinforcement learning, an AI agent (a software component) automatically explores its surroundings by striking and trailing, acting, learning from experiences, and increasing performance. Reinforcement learning operates on a feedback-based method. The objective of a reinforcement learning agent is to maximize the rewards since the agent is rewarded for every good activity and penalized for every negative one. In contrast to supervised learning, reinforcement learning relies only on the experiences of the agents.

Due to its way of working, reinforcement learning is employed in different fields such as Game theory, Operation Research, Information theory, multi-agent systems.A reinforcement learning problem can be formalized using Markov Decision Process(MDP). In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

**Categories of Reinforcement Learning:-**

Reinforcement learning is categorized mainly into two types of methods/algorithms:

**Positive Reinforcement Learning**: Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.

**Negative Reinforcement Learning**: Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

**4) Python language:-**

High-level, adaptable, and interpreted Python is a programming language renowned for its clarity and readability. It was made by Guido van Rossum and initially made available in 1991. Python is widely used in many different fields, including web development, data analysis, scientific computing, artificial intelligence, machine learning, automation, and more. It has experienced tremendous growth in popularity over the years.

Here are some of the Python programming language's salient qualities and attributes:

1. Readability and Learning Ease: Python is a great choice for novices because its syntax is made to be simple to understand and write. Indentation (whitespace) is used to denote code blocks, encouraging clear and dependable coding practices.
2. Python is an interpreted language, therefore you don't have to compile it before running any code written in it. Because of this, the development process is quicker and more engaging.
3. Cross-Platform: Linux, macOS, and Windows are just a few of the platforms that support Python. Code can be created on one platform and run with little to no changes on another.
4. Large Standard Library: The standard library that comes with Python is extensive and contains modules and packages for a variety of activities, from working with files and data to network programming and web development.
5. Python is dynamically typed, thus you don't need to explicitly specify the kinds of your variables. Code is more flexible since the type is decided by the interpreter at runtime.
6. Strong Typing: While Python is dynamically typed, it enforces strong typing, meaning that operations between incompatible types will result in a TypeError. For example, the popular data science, machine learning, and web development libraries NumPy, pandas, TensorFlow, and Django are included.
7. Extensible: You can easily extend Python by writing C or C++ code and integrating it with your Python programs, allowing you to leverage existing libraries and tools.
8. Community and Ecosystem: Python has a vibrant and active community of developers, which has led to the creation of a vast ecosystem of third-party libraries and frameworks. This includes libraries like NumPy, pandas, TensorFlow, and Django, which are widely used in data science, machine learning, and web development.
9. Python is adaptable for a variety of application types since it supports both procedural and object-oriented programming paradigms.
10. Python is free to use, distribute, and change because it is open source. This has helped it become widely used and keep becoming better.
11. Python is a great option for both beginning and expert developers because of its readability and simplicity. Building simple scripts to complex web apps and machine learning models, it is appropriate for a wide range of applications thanks to its adaptability and vast ecosystem of libraries.

**5) Statistics Vs Mathematics in Machine learning:-**

AI and machine learning are fundamentally based on math and statistics. In essence, mathematics is an approach to expressing concepts using numbers. Statistical communication of ideas is more abstract. Machine learning from experience is the most straightforward definition of artificial intelligence (AI). Both statistics and mathematics are essential to machine learning (ML), but they are utilised at different stages of the ML pipeline and for different reasons. Their functions and how they connect to ML are broken down as follows:

**Statistics in Machine Learning:**

**Data Analysis**: Statistics is fundamental in understanding and analyzing data before building machine learning models. Descriptive statistics (mean, median, standard deviation, etc.) help in summarizing and visualizing data, identifying outliers, and assessing data quality.

**Probability Theory**: Probability theory is essential for understanding uncertainty in data and model predictions. Concepts like probability distributions, Bayesian inference, and conditional probability are used in various ML algorithms, especially in probabilistic models.

**Hypothesis Testing:** Statistical hypothesis testing helps in making decisions about the significance of observed effects or differences in data. It is used in A/B testing and model evaluation to determine if differences are statistically significant.

**Regression Analysis:** Statistical regression techniques are used in linear regression and other regression models in ML. These techniques aim to model the relationship between variables and make predictions based on that relationship.

**Statistical Learning Theory**: Concepts from statistical learning theory, such as bias-variance trade-off, over fitting, and model selection, are crucial for understanding the generalization properties of ML models.

**Validation and Cross-Validation:** Statistical techniques like k-fold cross-validation are used to assess the performance of ML models, helping to avoid over fitting and selecting the best models.

**Mathematics in Machine Learning:**

**Linear Algebra**: Linear algebra is essential for understanding the mathematical underpinnings of many ML algorithms, especially those involving matrix operations. It is used in techniques like principal component analysis (PCA), singular value decomposition (SVD), and deep learning.

**Calculus:** Calculus plays a role in optimization algorithms used to train ML models. Gradient descent, a fundamental optimization technique, relies on calculus concepts like derivatives and gradients.

**Probability and Statistics**: While statistics focuses on data analysis and hypothesis testing, probability theory is used in probabilistic models, such as Bayesian networks and Gaussian processes, which are used in various ML applications.

**Information Theory**: Information theory concepts like entropy and mutual information are used in feature selection and dimensionality reduction techniques, helping to identify relevant features and reduce computational complexity.

**Optimization:** Mathematical optimization techniques are central to training ML models. Gradient-based optimization algorithms, such as stochastic gradient descent (SGD), are used to find model parameters that minimize the loss function.

**Conclusion**:-both statistics and mathematics are integral to machine learning. Statistics helps in understanding data, making inferences, and evaluating models, while mathematics provides the mathematical foundation for building and optimizing machine learning algorithms. A solid grasp of both areas is essential for a comprehensive understanding of machine learning concepts and their practical application.

**6) Measure of central tendency in ML:-**

Measures of central tendency are frequently employed in machine learning to describe and comprehend the central or typical value within a dataset. For a variety of tasks, including as data preprocessing, visualization, and model evaluation, these measurements offer insights into the "centre" of the data distribution. The three most popular central tendency measurements in machine learning are:

**The mean**, sometimes known as the average, is the most widely used indicator of central tendency. By adding up all the values in a dataset and dividing by the overall number of values, it is calculated. Extreme values have an impact on the mean, and it is sensitive to outliers. Sometimes outliers can considerably affect the mean, making it difficult to appropriately depict the typical value:

Mean is calculated as (Sum of all Values) / (Number of Values).

**The middle** value in a sorted dataset is called the median. When there are an even number of values, the data are first arranged in ascending order to get the median, which is the average of the two middle values. The median is a reliable indicator of central tendency because, in contrast to the mean, it is less susceptible to outliers.

Median = Middle Value (for an odd number of values)

Median = (Value at position N/2 + Value at position N/2 + 1) / 2 (for an even number of values).

**Mode:-**The value that appears the most frequently in a dataset is the mode. Depending on the dataset, there may be one mode (unimodal), multiple modes (multimodal), or no modes at all. When working with discrete or categorical data, the mode is helpful. It is less frequently used for continuous data.

Formula: Mode = Value(s) with the highest frequency

These measures of central tendency serve different purposes in machine learning:

* Mean is often used for continuous numerical data when the data distribution is approximately symmetric and not heavily affected by outliers.
* Median is preferred when dealing with skewed or non-normally distributed data, as it is less influenced by extreme values.
* Mode is useful for categorical data or discrete data where you want to identify the most common category.

It's crucial to pick the right central tendency measure based on the characteristics of your data and the precise objectives of your research or ML model. To fully comprehend the central tendency of the data, it may be helpful in some circumstances to take into account all three measurements.

The end