# Assignment 1:

Part 1:

Q1. Probability can be defined as the likelihood of an occurrence of an event or circumstance. For example the likelihood of the dice to land on the number six or for a coin to land with the heads side facing up.Theoretically probability can be calculated by dividing the number of desired outcomes by the total number of possible outcomes. In the example of a dice the probability of it landing on the number six is ⅙ since it is the number of desired outcomes(1) by the total number of outcomes(6).

● 0 means the event is impossible.

● 1 means the event is certain.

● 0.5 means the event is equally likely to occur or not occur.

Q2. Referring back to the first question we can use the formula of considering the number of favourable outcomes(1 since we are looking for the number 3) divided by the total number of outcomes(6 since that is the number of faces in a regular dice). In this case the likelihood of this event would be ⅙.

The number of favourable outcomes is 1(3).

The total number of outcomes is 6(1,2,3,4,5,6).

Q3. The central tendencies are:

- Mean: sum of all value divided by number of values
- Median: The middle value when the data is sorted(Ascending orders), if the number of values are even then it is an average of the two middle numbers.
- Mode: The value that has the highest frequency in the array or the value that appears the most

Q4. Descriptive statistics is used to summarize the data that has been collected or to point out its main features. There are different branches of descriptive statistics:

1. Measures of Central Tendency:

- Mean
- Median
- Mode

2. Measures of Dispersion/Variability:

- Range
- Variance
- Standard Deviation

Q5. Range is the difference between the highest and lowest value of the array. It gives the spread of the data and helps us understand the range of values which the data is varying between. For example in the array provided(60, 70, 80, 90, 100) we can say that the highest value is 100 and the lowest value is 60. Therefore the range will be 100(highest value) - 60(lowest value) which will be 40.

Q6. Variance and standard deviation both measure the spread of data around the mean. Variance is the average of the squared differences and is expressed in squared units, making it less intuitive. Standard deviation, the square root of variance, has the same units as the original data, making it easier to interpret and compare.

7. Probability helps us in dealing with the uncertain by helping us to predict the outcome of a particular event. With machine learning models that rely heavily on data, using this data the machine or the user can use probability to create predictions or predict the outcome of the event. ML models usually output probabilities and the ability to read these probabilities is crucial. It also helps us to confidence in our research and predictions based on the data collected.

Some examples of this are:

Rolling a Die: What's the probability of rolling a 4?

● Possible outcomes: 1, 2, 3, 4, 5, 6 (6 total)

● Favorable outcome (rolling a 4): 1

● Probability = 1/6 ≈ 0.167 or 16.7%

Coin Toss: What's the probability of getting Heads?

● Possible outcomes: Heads, Tails (2 total)

● Favorable outcome (Heads): 1

● Probability = 1/2 = 0.5 or 50%

Q8. Median is better when used for skewed data or data with outliers.

Some scenarios include:

1.) house prices in a city

2.) incomes

Both of these scenarios include sets of data which aren't arranged properly or are unclean. Using the median to clearly sort the data.

Q9. Data explorations can be defined as understanding new patterns, distributions, and relationships. Statistics plays a major part in data exploration helping it to identify newer patterns and distributions using statistics based of previous data collected on the matter.

Q10. A case study on Friedreich's Ataxia (FRDA) highlights the importance of both data and methods like statistics and machine learning (ML) by showing how:

- Data (e.g., genetic sequences, clinical symptoms, progression rates) is crucial for understanding the disease's causes, patterns, and patient variability.

- Statistical and ML methods help analyze complex datasets to identify key risk factors, predict disease progression, and support diagnosis or treatment decisions.

Together, they demonstrate how meaningful insights and advancements in medical research depend on high-quality data and the appropriate use of analytical techniques.

Q11. A large standard deviation in a dataset of house prices suggests there is wide variation—some houses are much cheaper, while others are extremely expensive (e.g., small apartments vs. luxury villas).

If you only looked at the mean (average) price, it could be misleading because:

- High-priced outliers can inflate the mean, making it seem like most houses are more expensive than they actually are.
- Similarly, low priced outliers will diminish the mean making it seem to be cheaper to buy a house
- It may not reflect what a typical buyer would actually end up paying.

In such cases, using the median (middle value) gives a more accurate sense of the “typical" house price.

Q12. The volcano plot shown in the slide is a visual tool commonly used in gene expression studies to identify potential biomarkers—genes that show significant changes in activity between two conditions (e.g., healthy vs. diseased).

How to interpret the plot:

- X-axis: $\log_2$(fold change)
    - Measures how much gene expression changes.
    - Positive values (right side): Genes that are up-regulated (more active in the condition being studied).
    - Negative values (left side): Genes that are down-regulated (less active).
- Y-axis: $-\log_{10}$(adjusted p-value)
    - Measures statistical significance of the change.

    - Higher points mean more statistically significant results.
    - Adjusted p-values account for multiple testing to reduce false positives.
- Colored dots:
    - Blue (left): Significantly down-regulated genes.
    - Red (right): Significantly up-regulated genes.
    - Gray (center): Genes with no significant change.

What "up-regulated" and "down-regulated" mean:

- Up-regulated = The gene is more active (expressed more) in the disease or test condition.
- Down-regulated = The gene is less active (expressed less) in that condition.

Q13. "A field of study that gives computers the ability to learn without being explicitly programmed." - Arthur Samuel, 1959

Q14. The big three types of machine learning are:

1.) Supervised learning

2.) Unsupervised learning

3.) Reinforcement learning

Q15. In supervised learning, the main difference between classification and regression lies in the type of output they predict:

- Classification:
    - The goal is to predict a category or class label.
    - Output is discrete (e.g., "spam" or "not spam", "disease" or "no disease").
    - Example: Predicting whether an email is spam.

Eg: Is this tumor Malignant or

Benign?

    - 
- Regression:
    - The goal is to predict a numerical value.
    - Output is continuous (e.g., house price, temperature).
    - Example: Predicting the price of a house based on its features.

Eg: What will be the price of this house?

Q16.The main goal of Unsupervised learning is in Finding hidden patterns. It is used to find hidden structures, patterns r relationships in data.

Q17. Principal Component Analysis (PCA)

Primary Purpose in Unsupervised Learning:

PCA is used for dimensionality reduction—it helps simplify large, complex datasets by:

- Transforming the original features into a smaller set of new variables called principal components.
- These components capture the most important patterns (i.e., the directions with the most variance) in the data.

Why it’s useful:

- Reduces noise and redundancy.
- Makes data easier to visualize (e.g., in 2D or 3D plots).
- Speeds up other unsupervised tasks like clustering.

Q18. The difference between traditional programming and machine learning lies in how they use inputs and outputs:

Traditional Programming:

- Inputs: Rules (logic/code) + Data
- Output: Answers/results
- You write explicit rules to tell the computer what to do with the data.

Example:
If you code: if score > 90: grade = 'A', the program applies this rule to student scores to generate grades.

---

Machine Learning:

- Inputs: Data + Answers (labeled outcomes)
- Output: Model (learned rules)
- The computer learns patterns from examples instead of being told the rules.

Example:
Feed a model past student scores and their grades. It learns how scores relate to grades and can then predict grades for new scores.

Q19. The core idea of "Learning from Examples" in Machine Learning is that instead of explicitly programming rules, a model learns patterns from labeled data.

Reffering to the cat recognition analogy it refers to the machine finding patterns in assessing whether the object in front of it is a cat or not based on previous data provided to it.

Q20. An agent learns to take actions in an environment to maximise rewards. It learns through feedback loop where it is continuously fed bad or good feedback which helps it in learning.

Q21. Supervised learning:

1.)Classification

2.)Regression

Unsupervised learning: 1.) Clustering

Q22.)

Why are “Data Preprocessing“ and “Feature Engineering“ marked as “IMPORTANT!“ in the Machine Learning Workflow? What problems might occur if they are not done properly?

Answer:

Data Preprocessing and Feature Engineering are marked as “IMPORTANT!” in the machine learning workflow because they directly affect the quality of input data, which in turn determines how well a machine learning model can learn patterns and make accurate predictions.

If these steps are not done properly, the following problems can occur:

- Missing or noisy data may lead to incorrect model training, resulting in poor performance.
- Incorrect or inconsistent formatting can break the training pipeline or skew the results.
- Irrelevant or redundant features may confuse the model, reduce accuracy, and increase computation time.
- Unscaled features can negatively impact algorithms that rely on distance metrics (e.g., k-means, logistic regression).
- Unencoded categorical variables may not be interpretable by the model, leading to errors or meaningless outputs.

These steps are critical because they ensure the dataset is clean, consistent, and meaningful, enabling the model to learn effectively.

Q23. In the context of spam email detection, if a spam filter incorrectly marks an important email from your school as spam, what type of error is this? Why is it particularly problematic?

Answer:

This is a False Positive error.

In classification tasks:

- A False Positive (FP) occurs when the model incorrectly predicts a positive class (e.g., "Spam") for an instance that is actually negative (e.g., "Not Spam").
- A False Negative (FN) would be when a spam email is marked as not spam.

Marking a legitimate email (like one from the school) as spam is particularly problematic because:

- Critical or time-sensitive information may be missed, such as deadlines, class announcements, or exam schedules.
- Trust in the system decreases, as users may no longer rely on the spam filter.
- It may lead to unintended consequences such as missing opportunities or failing to act on required information.

False Positives in spam filters are thus more disruptive to user experience compared to False Negatives.

24. What is the broad definition of Artificial Intelligence (AI) provided in the slides?

AI is defined as the field of computer science focused on creating systems that can perform tasks that typically require human intelligence. These tasks include reasoning, problem-solving, learning, perception, language understanding, and decision-making.

25. According to the concentric circles diagram, what is the relationship between AI, Machine Learning (ML), and Deep Learning (DL)?

The relationship is hierarchical:

- AI is the broadest field encompassing all intelligent systems.
- ML is a subset of AI, focused on systems that learn from data.
- DL is a further subset of ML, dealing specifically with neural networks with many layers, inspired by the human brain.

26. List the three types of AI based on capability discussed in the slides. Which type do we have today?

1. Artificial Narrow Intelligence (ANI) / Weak AI
2. Artificial General Intelligence (AGI) / Strong AI
3. Artificial Superintelligence (ASI)

Today, we only have ANI — systems that are specialized in one area, like Siri, Alexa, or self-driving cars.

27. Name two key areas that are considered “Foundations of AI.”

Two key foundational areas of AI mentioned in the slides are:

- Natural Language Processing (NLP)
- Computer Vision

Others include robotics, knowledge representation, planning, and problem-solving.

28. Briefly explain the difference between AI “Thinking Humanly” and “Acting Rationally” as goals of AI, according to Russell & Norvig’s categories.

- Thinking Humanly refers to designing AI systems that mimic human cognitive processes, such as reasoning and memory, like simulating how a human solves problems.
- Acting Rationally focuses on building systems that make optimal decisions to achieve goals, regardless of whether the approach is human-like — prioritizing outcomes over imitation.

29. What is Natural Language Processing (NLP)? Give one example application mentioned.

NLP is the area of AI that enables computers to understand and generate human language.
An example mentioned in the slides is Machine Translation, such as Google Translate.

30. What is Generative AI, and how does it differ from AI models that only analyze existing data? Give an example.

Generative AI refers to models that can create new content (e.g., images, text, music), not just analyze data.
It differs from traditional AI models, which only detect patterns or classify existing data.
An example is DALL·E, which can generate images from text descriptions.

31. Explain how an AI model might learn biases from data and give a hypothetical example of an unfair outcome that could result.

AI models learn from historical data. If that data contains biases (e.g., underrepresentation of a group), the model may replicate those patterns.
Example: A hiring algorithm trained on past hiring data may discriminate against women if previous hiring decisions favored men, leading to unfair rejection of qualified female candidates.

32. Why is it important to understand *how* an AI model makes its decisions, especially in critical applications like healthcare?

Understanding how an AI model makes decisions ensures:

- Trust in the system
- Accountability when outcomes affect human lives
- Detection of errors or bias in decision-making

In healthcare, this is essential because wrong decisions (e.g., misdiagnosing a condition) can lead to serious harm or legal consequences. Explainability allows professionals to validate and justify the AI's recommendations.