Part 1: Probability and Statistics

1. Probability is the measure of how likely an event is to occur. A probability of 0 means the event will not happen, 0.5 means it has an equal chance of happening or not, and 1 means it is certain to happen.

2. A standard die has 6 sides. There's only 1 side with a '3'.
   Favorable outcomes = 1
   Total outcomes = 6
   Probability = 1/6 ≈ 0.1667

3. The three main measures of central tendency are: Mean, Median, and Mode.

4. The primary purpose of descriptive statistics is to summarize and describe the main features of a dataset in a clear and understandable way.

5. Range = Highest score - Lowest score
   Given: 60, 70, 80, 90, 100
   Range = 100 − 60 = 40

6. Variance measures the average squared deviation from the mean, and its units are the square of the original data's units. Standard Deviation is the square root of variance, so it is in the same units as the data, making it easier to interpret.

7. Understanding probability helps interpret ML outputs, like prediction confidence. For example, in medical diagnostics, a model may predict a disease with 90% probability. This guides decisions. Without grasping probability, the prediction might be misunderstood or misused.

8. Use the Median when data has outliers. Example: If house prices are ₹50L, ₹60L, ₹65L, and ₹5Cr, the Mean will be skewed by ₹5Cr, but the Median (₹62.5L) gives a more typical value.

9. Data exploration involves analyzing datasets to find patterns, trends, or anomalies. It helps data scientists decide which models or preprocessing steps are appropriate.

10. The FRDA case study showed that both high-quality data and methods like statistical modeling are crucial. It demonstrated how statistical techniques helped identify biomarkers, which could have been missed without proper analysis.

11. A large standard deviation suggests house prices vary a lot (e.g., ₹30L to ₹10Cr). This makes the mean misleading, as it may suggest a typical price that few people actually pay. It highlights the need for other metrics like median.

12. A volcano plot uses:

- x-axis: log2(fold change) – how much expression changed
- y-axis: −log10(p-value) – how statistically significant the change is
  Colored dots show genes that are significantly “up-regulated” (increased expression) or “down-regulated” (decreased). It helps identify important genetic markers.

Part 2: Machine Learning Fundamentals

13. Arthur Samuel (1959) defined ML as the field of study that gives computers the ability to learn without being explicitly programmed.

14. The “Big Three” types of ML are:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

15. In classification, the output is a category (e.g., spam or not spam). In regression, it’s a continuous value (e.g., predicting house price).

16. The main goal of Unsupervised Learning is to find hidden patterns or structures in unlabeled data.

17. PCA stands for Principal Component Analysis. It reduces data dimensions while retaining the most important variance in the dataset.

18. In traditional programming, we provide rules (logic) and data to get an output. In ML, we provide data and outputs to the algorithm, and it learns the rules on its own.

19. ML learns from examples by identifying patterns. For instance, in cat recognition, a model sees many labeled cat images and learns features (ears, whiskers) that distinguish cats from other animals.

20. In Reinforcement Learning, an agent interacts with an environment, takes actions, and learns through rewards or penalties to maximize long-term outcomes.

21. Supervised Learning: Decision Trees, Support Vector Machines (SVM)
    Unsupervised Learning: K-Means Clustering

22. Data preprocessing and feature engineering are marked “IMPORTANT!” because poor-quality data leads to bad models. If not done properly, issues like missing values, irrelevant features, or inconsistent formats can drastically reduce accuracy and reliability.

23. This is a False Positive—the model wrongly labels a non-spam email as spam. It’s problematic because you might miss important messages, causing delays or misunderstandings, especially in critical contexts like education or work.

Part 3: Artificial Intelligence Concepts

24. AI is broadly defined as the science of making machines that can perform tasks that require human intelligence, like learning, reasoning, and problem-solving.

25. According to the concentric diagram:

- AI is the broadest category.

- Machine Learning is a subset of AI.

- Deep Learning is a subset of ML.

26. Three types of AI by capability:

- Narrow AI

- General AI

- Super AI
  Today, we only have Narrow AI.

27. Two key foundations of AI:

- Mathematics
- Cognitive Science

28. “Thinking Humanly” aims to model how humans think, while “Acting Rationally” aims for the most logical, optimal decisions, even if not human-like. The first focuses on replicating thought; the second focuses on outcomes.

29. NLP (Natural Language Processing) is the field that allows machines to understand and generate human language. An example is machine translation, like Google Translate.

30. Generative AI creates new content (text, images, music). Unlike analytical AI that only evaluates, generative AI can produce original outputs, like ChatGPT writing poems or Midjourney creating art.

31. AI learns from data, so if the training data reflects societal biases (e.g., hiring decisions that favor men), the AI may replicate and reinforce those biases. For instance, a resume screener might unfairly reject women applicants.

32. Explainability is important because users need to trust and understand AI decisions, especially in areas like healthcare. If an AI recommends a treatment, doctors must understand the reasoning to ensure it’s safe and appropriate.