ASSIGNMENT #2

Applied Data Science with ML & AI - Horizons 25

1. **Define probability in your own words. What do probabilities of 0, 0.5, and 1 signify for an event? -** Probability is the measure of how likely an event is to occur. The higher the probability, the more likely the event is to happen.

   0 - means an impossible event - it can never happen
   0.5 - means an event that is equally likely to happen or not happen (a 50/50 chance)
   1 - means a sure event - it will definitely happen

2. **What is the probability of rolling a '3' on a standard six-sided die? Show the favorable outcomes and total possible outcomes.**

   Favourable Outcomes: '3'
   Total Possible Outcomes: '1', '2', '3', '4', '5', '6'
   Probability: 1/6

3. **List the three main measures of central tendency discussed.**

   1. Mean - The sum of all values divided by the number of values
   2. Median - The middle value of sorted data
   3. Mode - The value that appears most frequently in the data

4. **What is the primary purpose of descriptive statistics?**

   Descriptive statistics summarises and describes the main features of a dataset, showing us where the centre of the data is and how spread out it is.

5. **Define "Range" as a measure of dispersion. How is it calculated using the example test scores: 60, 70, 80, 90, 100?**

   Range is the difference between the highest value and the lowest value of a dataset. It gives us a quick sense of how wide the data varies.

   Highest = 100, Lowest = 60
   RANGE = 100-60 = **40**

6. **What is the key difference between “Variance” and “Standard Deviation” in terms of their units and interpretability?**

   VARIANCE - expressed in squared units, harder to relate to the original data
   STANDARD DEVIATION - expressed in the same unit as the original data, easier to interpret and understand as the typical distance from the mean.

7. **Explain why understanding probability is crucial when working with Machine Learning models. Give one example from the slides.**

   - Real-world data is complex. Thus, ML models often provide outputs in the form of probabilities, which makes understanding it crucial.
   - Probability helps us measure the model’s confidence in its predictions and is the foundation for many statistical tests and ML algorithms.

   Eg. "80% chance this email is spam"

8. **When would you prefer to use the Median over the Mean to describe the central tendency of a dataset? Provide an example scenario**.

   Using the median is preferred when the data is skewed or has outliers.
   Eg. House prices in Mumbai

9. **The slides mention “Data Exploration” as a reason why statistics is important in Data Science & ML. Explain what this means in a sentence or two.**

   Data exploration means finding patterns, understanding how values are spread out, and seeing how different variables relate to each other in a dataset. Statistics helps with this.

10. **Briefly describe how a Case Study, like the one presented on Friedreich’s Ataxia (FRDA), highlights the importance of both data and methods.**

    Case studies — like the FRDA case study — show how both **data** (like gene expression levels) and **methods** (like statistics and ML) work together to discover important **patterns** or outliers that may not be obvious otherwise.
    Eg. A **volcano plot** was used to spot genes that were extremely different between groups, helping identify possible **biomarkers** for the disease.

11. **Imagine a dataset of house prices in a city. Why might the standard deviation be very large? How could this affect your interpretation of the "average" house price if you only looked at the mean?**

    A large standard deviation in house prices means there's a wide gap between the cheapest and most expensive homes—like small apartments vs luxury villas. If you only look at the mean, it could be misleading because it gets warped by extreme values, and might not actually show you what **most people** pay.

12. **The slides show a "Volcano Plot" in the context of discovering biomarkers. Without needing to understand all the biology, what do you think the plot is trying to show based on its axes ("log2(fold change)" and "−log10(adjusted p-value)") and the colored dots? What might "up-regulated" and "down-regulated" mean in simple terms?**

    I think the volcano plot shows which genes are changing the most (x-axis) and how confident we are in those changes (y-axis). The coloured dots show significant changes.

    Up-regulated - gene activity has increased
    Down-regulated - gene activity has decreased

13. **What is Arthur Samuel's 1959 definition of Machine Learning?**

    Arthur Samuel defined Machine Learning as the field that gives computers the ability to learn from data without being explicitly programmed for every task.

14. **List the "Big Three" types of Machine Learning.**

    1. Supervised Learning
    2. Unsupervised Learning
    3. Reinforcement Learning

15. **In supervised learning, what is the difference between "Classification" and "Regression" tasks? Give one example of each from the slides.**

    Classification - Predicting a category or class. The output is discrete.
    Eg. Is this email spam or not spam?

Regression - Predicting a continuous value. The output is a number.
Eg. What will be the price of this house?

16. What is the main goal of Unsupervised Learning, according to the slides?

The main goal of Unsupervised Learning is to discover hidden structures, patterns, or relationships in the data.

17. What does PCA stand for and what is its primary purpose in unsupervised learning?

PCA stands for Principal Component Analysis. Its main purpose is to reduce the number of features in a dataset while keeping the important information—making models faster and easier to understand.

18. Explain the difference between traditional programming and machine learning in terms of their inputs and outputs.

In traditional programming, you give the computer **rules + data** → it gives **results**. In machine learning, you give the computer **data + results** → it learns the **rules** (model). ML is better for problems where writing rules manually is too hard.

19. Briefly describe the core idea of "Learning from Examples" in Machine Learning, using the cat recognition analogy.

"Learning from examples" means the model is trained using many labeled examples. Like showing a child lots of pictures of cats, so they learn what a cat looks like. The ML model does the same—it finds patterns and uses them to recognize new examples.

20. What is an "agent" in the context of Reinforcement Learning, and how does it learn?

An **agent** is like a virtual decision-maker in Reinforcement Learning. It learns by interacting with an environment, trying actions, and getting rewards or penalties—like training a dog with treats.

21. List two common ML algorithms for Supervised Learning and one for Unsupervised Learning mentioned in the slides.

Supervised Learning - Logistic Regression, Decision Trees
Unsupervised Learning - K-Means (clustering)

22. **The "Machine Learning Workflow" includes "Data Preprocessing" and "Feature Engineering." Why do you think these steps are marked as "IMPORTANT!" and what kind of problems might occur if they are not done properly?**

If your data is messy or your features are poor, your model will perform badly—even if you use a great algorithm. Thus, these steps are marked as "IMPORTANT!"
Eg. Missing values, irrelevant features, or wrongly scaled data can lead to incorrect predictions.

23. Consider the spam email detection example. If a spam filter incorrectly marks an important email from your school as spam, what type of error is this in the context of classification (e.g., False Positive, False Negative)? Why might this type of error be particularly problematic?

If a spam filter marks an important school email as spam, that's a False Positive (predicting "spam" when it's not). This is a big problem because you might miss something important without even knowing it was there.

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

AI is a branch of computer science focused on building systems that can do tasks requiring human-like intelligence, such as learning, reasoning, and decision-making.

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

AI is the broadest field, ML is within AI and Deep Learning is a subfield of ML.
(So all DL is ML and all ML is AI)

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

- Artificial Narrow Intelligence

- Artificial General Intelligence
- Artificial Super Intelligence

We currently have ANI - AI that's good at one specific task (like Siri or Google Translate). The other two types are still hypothetical.

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

NLP and Computer Vision

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

Thinking Humanly - AI that mimics human thought processes (like problem-solving the way a human would).
Acting Rationally - AI that chooses the best action to achieve a goal, based on logic—even if it doesn't think like a human.

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

NLP is about teaching machines to understand and generate human language.
Eg. Google Translate or chatbots like Siri.

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

Generative AI **creates new content** (like images, text, or music). This is different from normal AI models that just analyze existing data.
Eg. DALL·E creates images from text prompts.

31. The slides discuss "Ethical Considerations in AI," including "Bias." Explain how an AI model might learn biases from data and give a hypothetical example of an unfair outcome that could result.

AI models can pick up biases if they're trained on data that's already unfair. For example, if past hiring data mostly includes men, the AI might learn to prefer male candidates—even if a woman is just as qualified. This can lead to unfair decisions and repeated discrimination.

32. The concept of "Explainability" or "Transparency" in AI is becoming increasingly important. Why do you think it's important to understand how an AI model makes its decisions, especially in critical applications like healthcare?

Understanding *how* an AI model makes its decisions is really important—especially in areas like healthcare, where the stakes are high. If a model suggests the wrong diagnosis, doctors need to know why it made that choice. Without explainability, it's hard to trust or fix the system when something goes wrong.