1. Random forest and XGBoost which one is more prone to overfitting and why?
2. In RF if we change n\_estimators from 150 to 100 will it help to reduce overfitting? or what will happen?
3. If we’ve too many features in our data, and for unsupervised clustering which matrix would you choose our ot two: Euclidian and Manhattan and why?
4. In imbalanced data, for classification, which metric would you use?

## **GenAI, LLM:**

1. How can LLM be effectively utilized for domain-specific tasks?
2. In LangChain, could you please list-out the available text splitters and how they function?
3. How can we ensure that tokens are not split when using the RecursiveCharacterTextSplitter in LangChain for character-level chunking?
4. If we store all PDF documents in chunk format within VectorDB and retrieve the top 10 chunks, how can we manage the scenario where chunks are extracted from different PDFs?
5. How do you evaluate your LLM model response?
6. Explain ROUGE and BLEU Scores?
7. Could you please explain what kind of mathematical calculation is used in Re-ranking models?
8. Can we implement the LLM model to classify whether a given retrieved chunk is relevant or not? How effective would it be?

## **Deployment**

1. Explain CICD pipeline?
2. How Docker is different from .bash or shell files?

## **Natural Language Processing:**

* What is tokenization and stemming?
* What is bag of words?
* What is word2vec?
* What is standardization, normalization, and scaling?
* What are transformers?
* What is BERT?
* Why are transformer models considered superior to other models?
* What is LSTM?
* What is the main purpose of BERT?
* What is Tf-IDF?
* What are BERT embeddings?
* When should BERT embeddings be used?

## **GenAI, LLM:**

1. What are the criteria for defining a model as an LLM model?
2. Can you explain Prompt Fine-tuning?
3. Have you had experience working with Lora and Alora?
4. What are the characteristics of normal distribution?  
   2. What is central limit theorem?  
   3. How much percentage data it covers when we have 2sigma value in normal distribution around mean?  
   4. What is P-value?  
   5. What is type-1 error?  
   6. Why the mean and variance of poison distribution is same?  
   7. If we have low variance data/variable and high variance data/variable which one do you think is good for machine learning model?
5. Explain how NER typically works  
   2. Difference between traditional NER and spacy

Part 4: Probability questions — 5 mins  
1. Let’s assume we have initially just one ball in a bag(and we don’t know the color of a ball) and then we added another red ball in a bag and shuffled it. What is the Probability of getting a red ball from a bag?  
2. If we have numbers from 0 to infinity, what is the probability of number is divisible by 120?