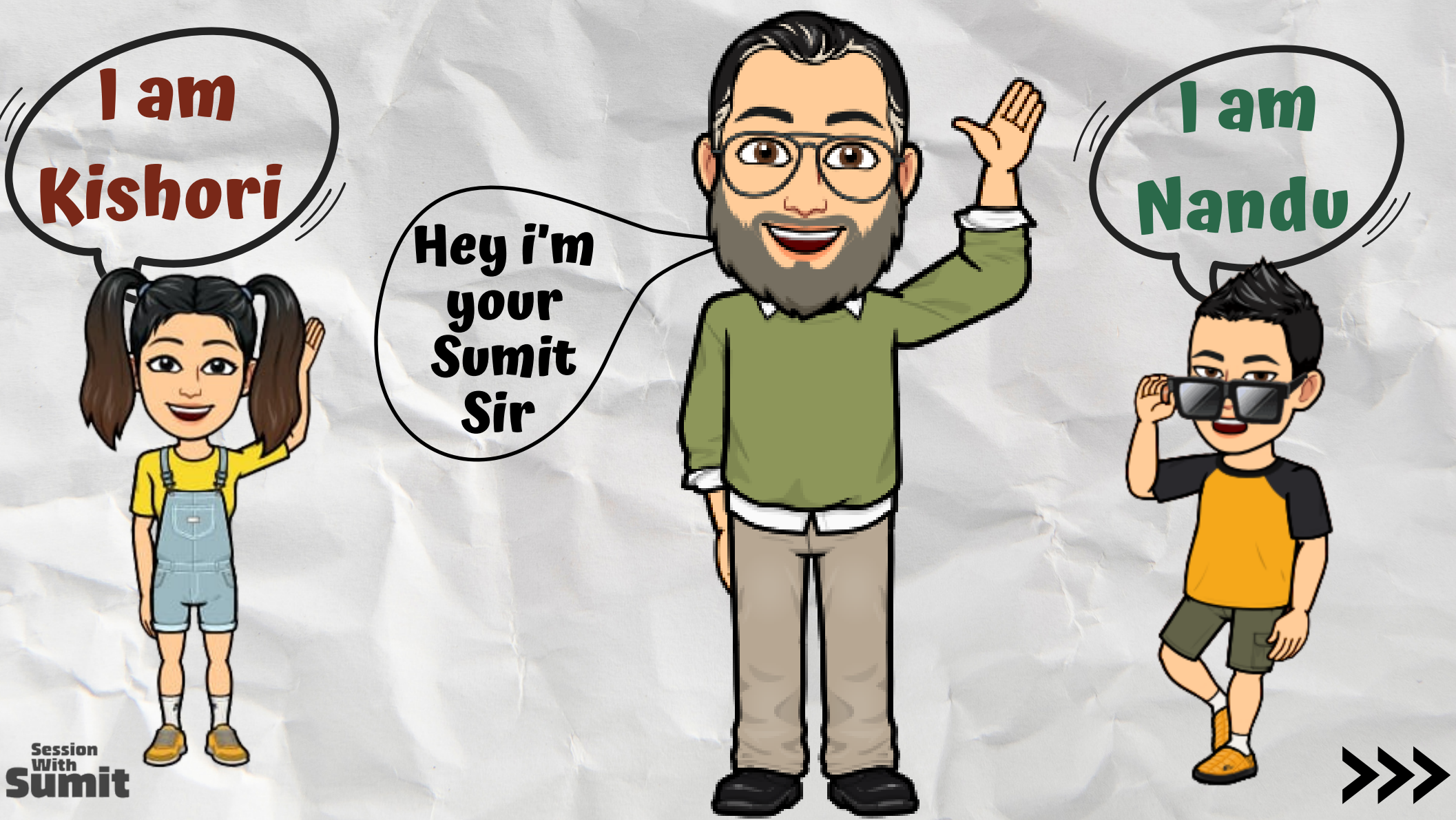
Interview Question

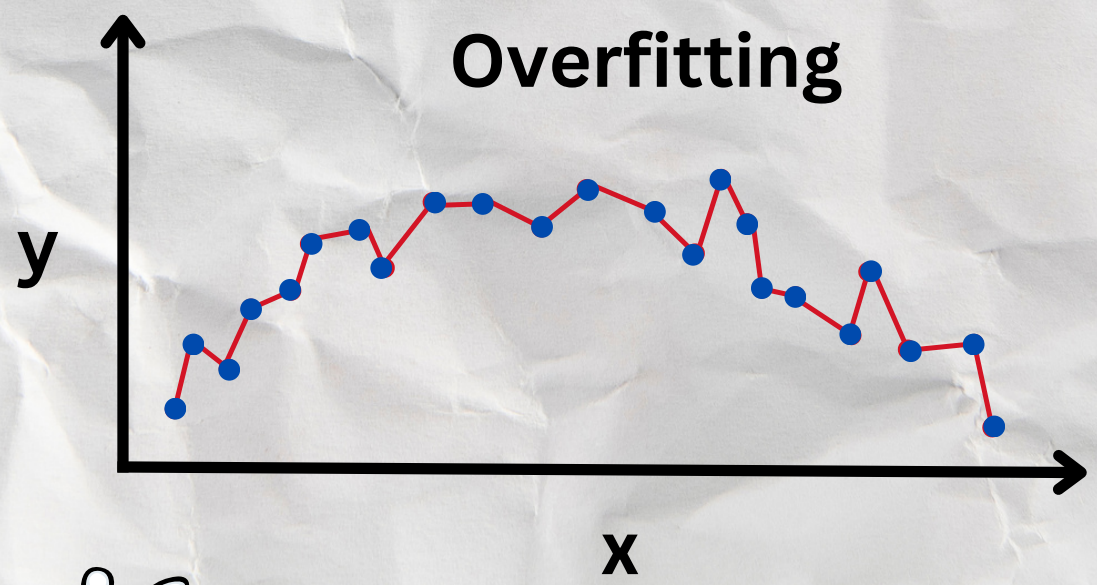
Explain, Various Ways To Avoid Overfitting?

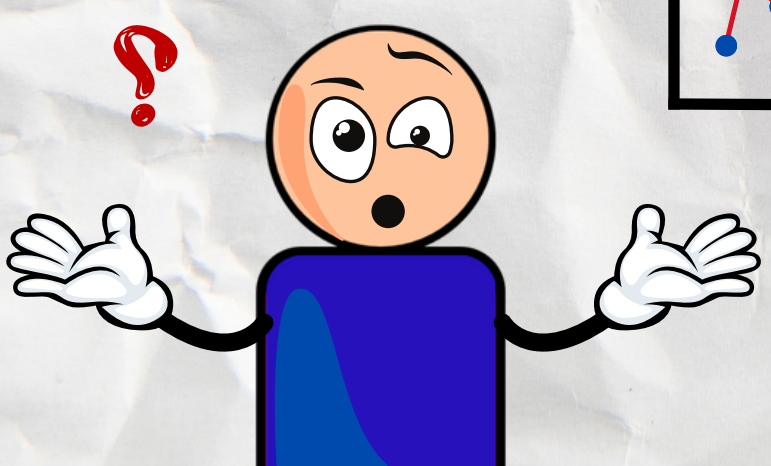


Session With Sumit

Let's first understand, what overfitting is.

Overfitting is a statistical modeling error where a function is too closely aligned to a limited set of data points. It is useful for an initial data set and not for any other data set.





Session With Sumit

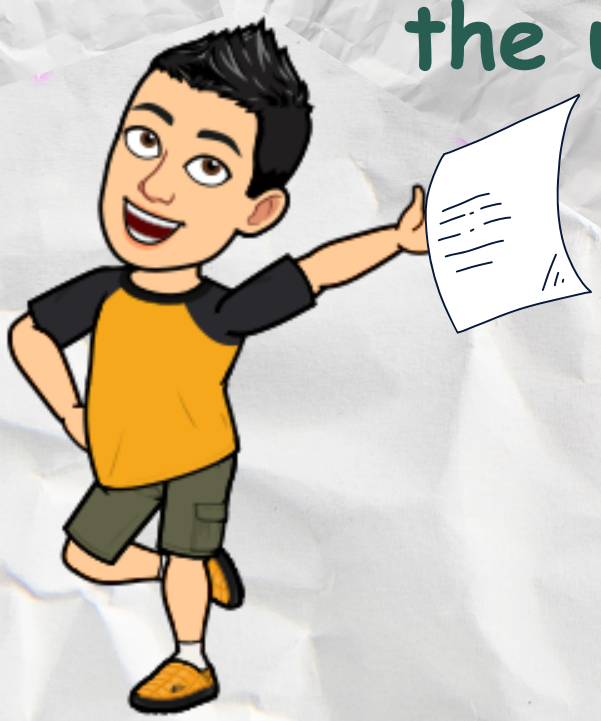
I have not consumed Bournvita today. Will you please explain in simple terms?

Session With Sumit

Let's take one example.

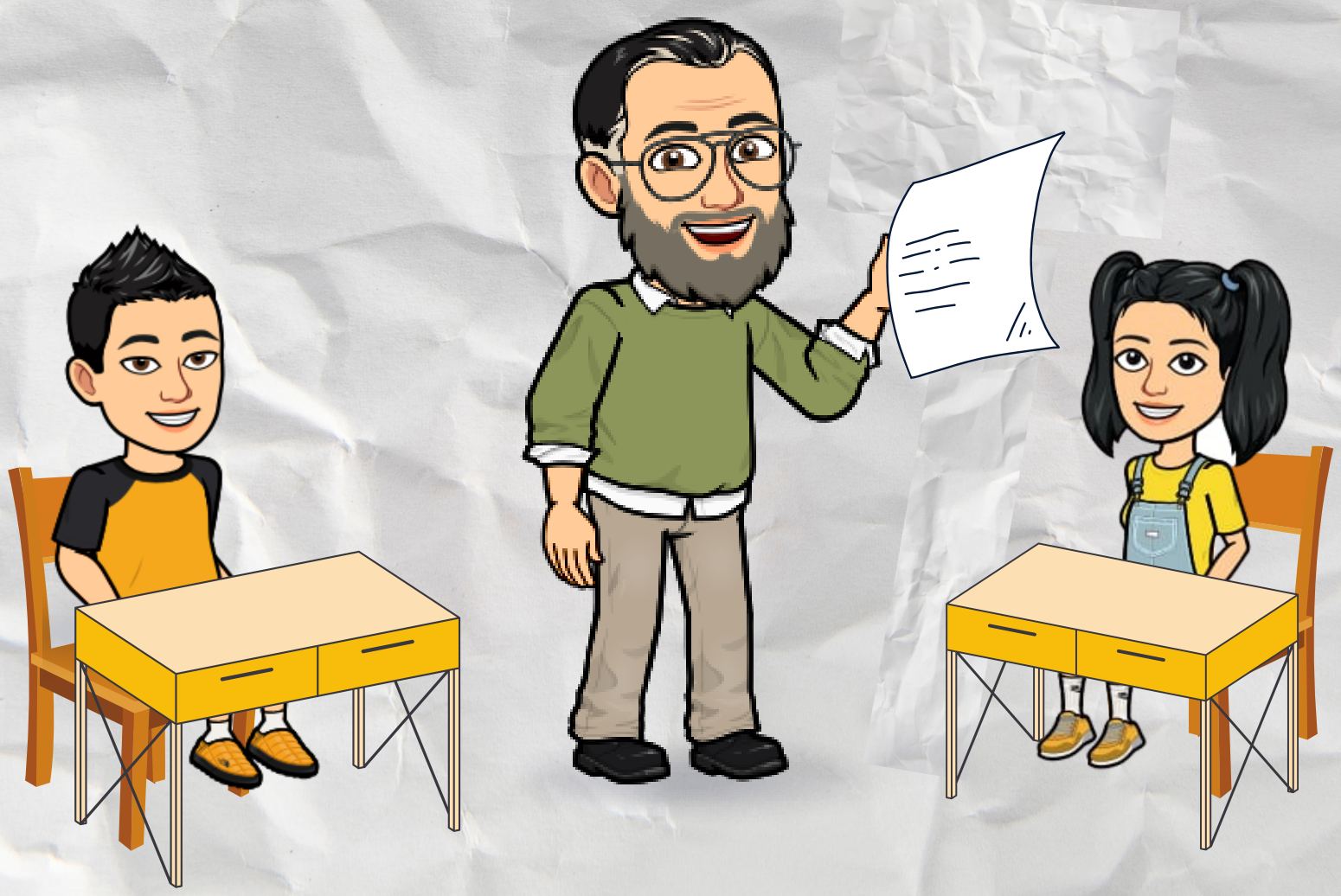
Nandu and Kishori are preparing for the science exam that they have to attempt tomorrow. Nandu being lazy and unserious towards exams chose to prepare for the examination using sample paper (paper 1)

Kishori is a knowledge seeker and doesn't study just to pass the exam. So, she chose to go through the chapters and understand the underlying concepts.



Session With Sumit

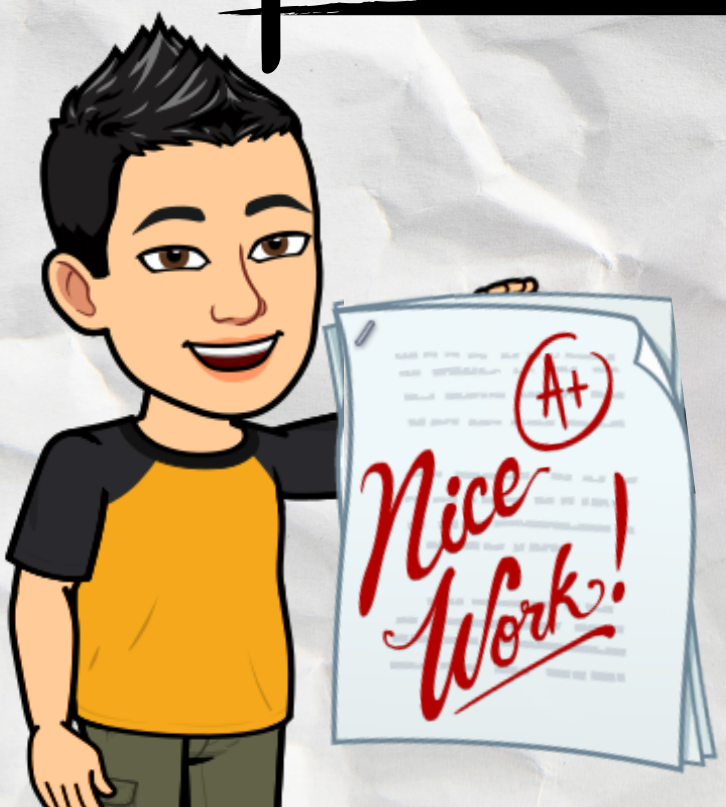
To test both of them, Sumit Kaka (Their home tutor) asked them a few questions to solve. Nandu being a cunning kid, handed over the same sample papers he was preparing from. Sumit Kaka asked 10 questions to both of them.

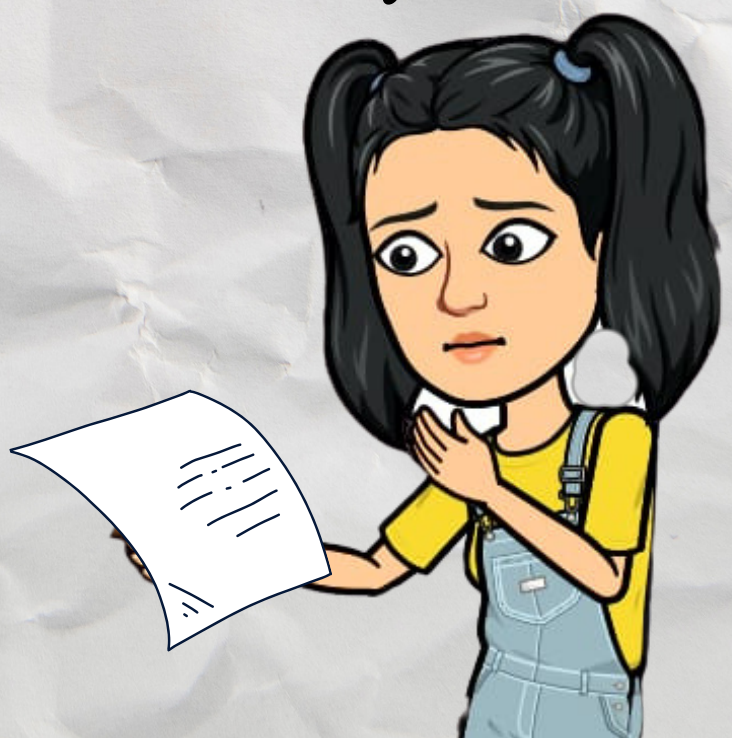


Session With Sumit

When Sumit Kaka, evaluated the answers, Nandu got 9/10 while Kishori scored 7.5/10.

So was nandu well prepared for the examination?





Session With Sumit

THE NEXT DAY...

Session With Sumit

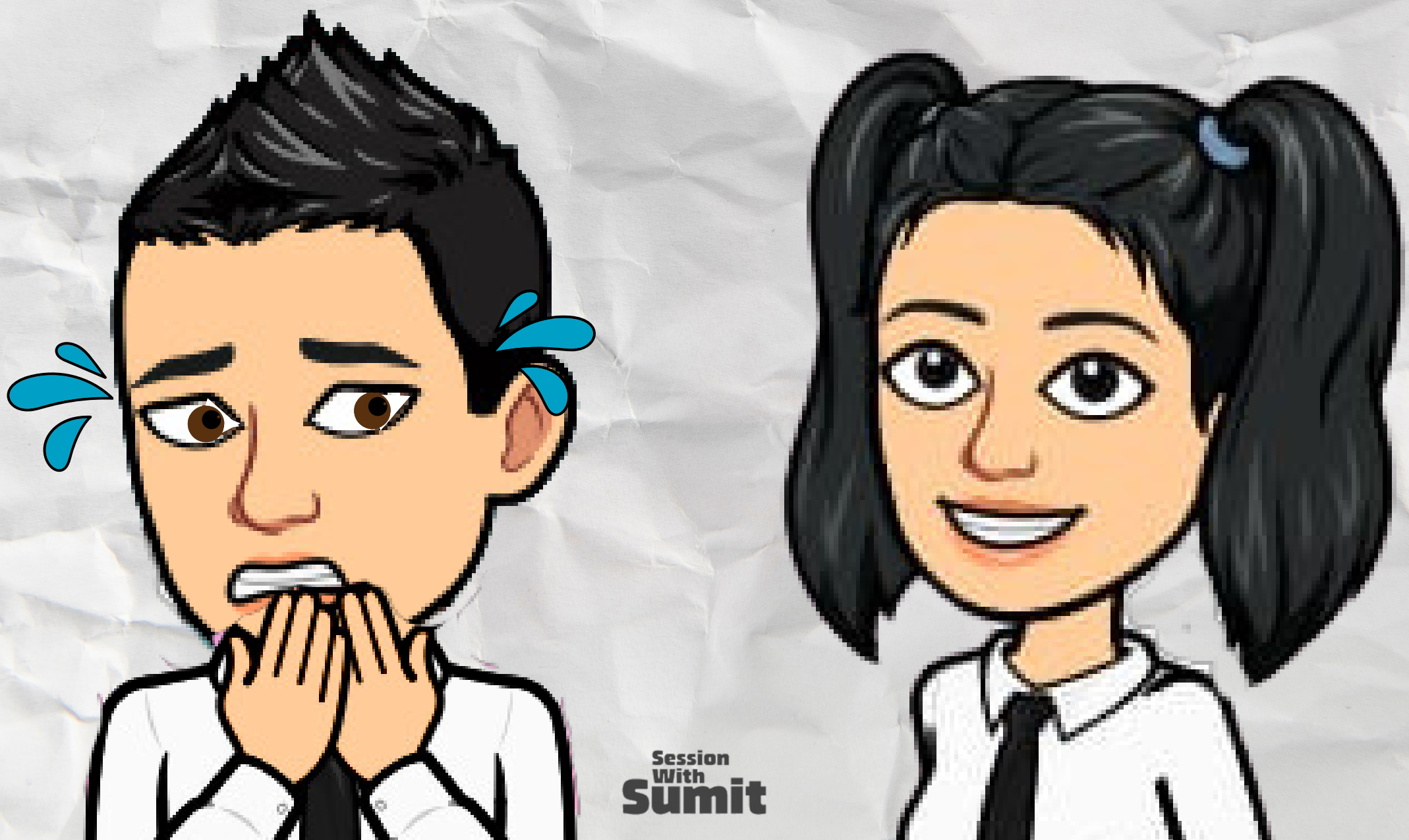
On the examination day

Nandu was shocked to find a paper that was out of the scope of the sample papers. He struggled a lot.

Kishori was fine with the level of the exam and the type of questions. With her knowledge of the underlying concepts, she was able to solve most of the paper without struggling.

Session With Sumit

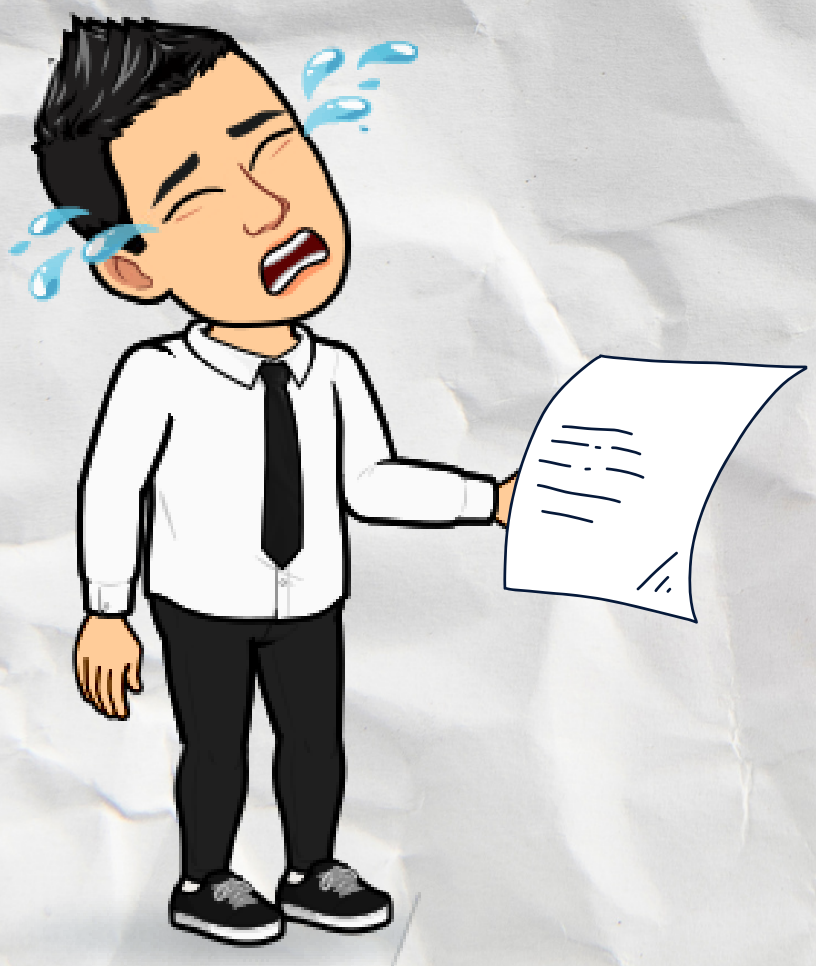
The result day

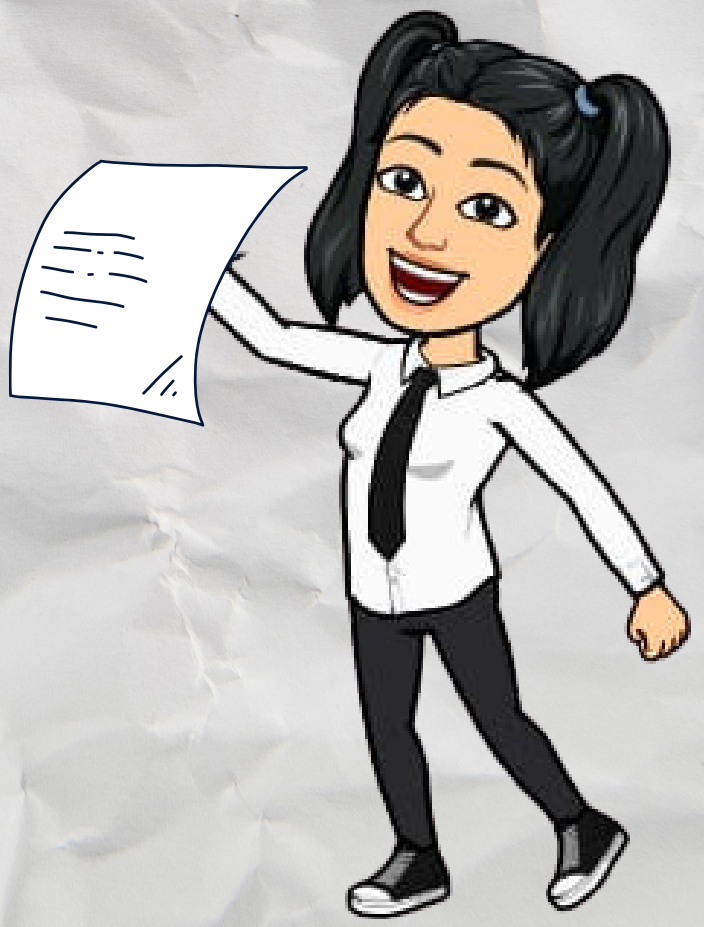


Session
With
Sumit

On The result day

Nadu just passed the exam by scoring 35/100 while Kishori bagged 81/100





Session With Sumit

What Did You Learn From This Story ?

Session With Sumit

Session With Sumit

Let's Summarize

	Home test	Actual exam
Nandu	9/10 = 0.9	35/100= 0.35
Kishori	7.5/10 = 0.75	81/100= 0.81

Here we can easily understand that Nandu performed well during the home test but his performance was very bad on the actual day of the exam.

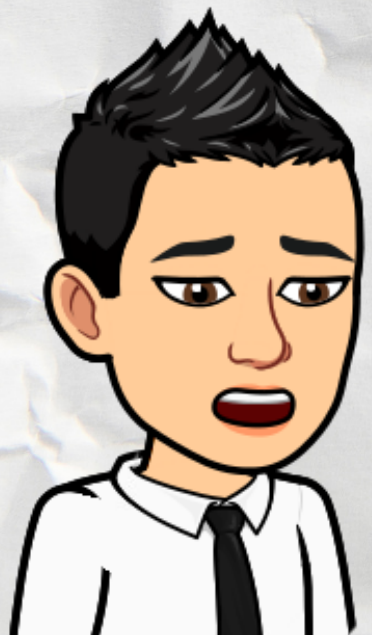
Whereas Kishori performed satisfactorily well on both the home test and the actual exam.

The case that happened with Nandu is known as Overfitting.

Overfitting is the case in which a Machine Learning Model (Nandu) performed very well on the train data (home test) but fails on the test data (Actual exam)

>>>

Whereas we always need a generalized good fit model (Kishori) which can perform well on both train data (home test) as well as test data (Actual exam)



Session With Sumit

But, How To Prevent Overfitting?

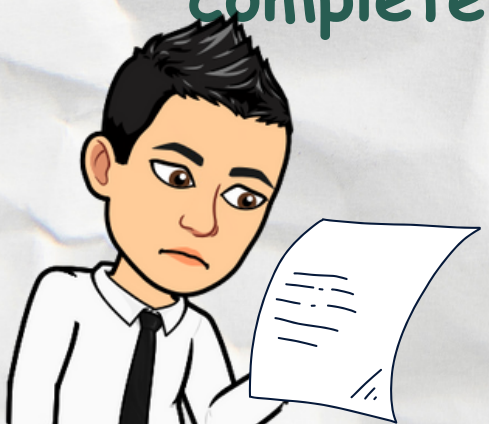
Session With Sumit

Train With More Data

If we look back to the example of Nandu and Kishori, we can understand that since Nandu trained himself on a limited number of questions (1 sample paper), he was easily able to memorize all the questions. In short, he overfitted himself with the sample papers.

While on the other hand, Kishori was learning from books and understood the underlying concept but was not able to memorize the complete book.

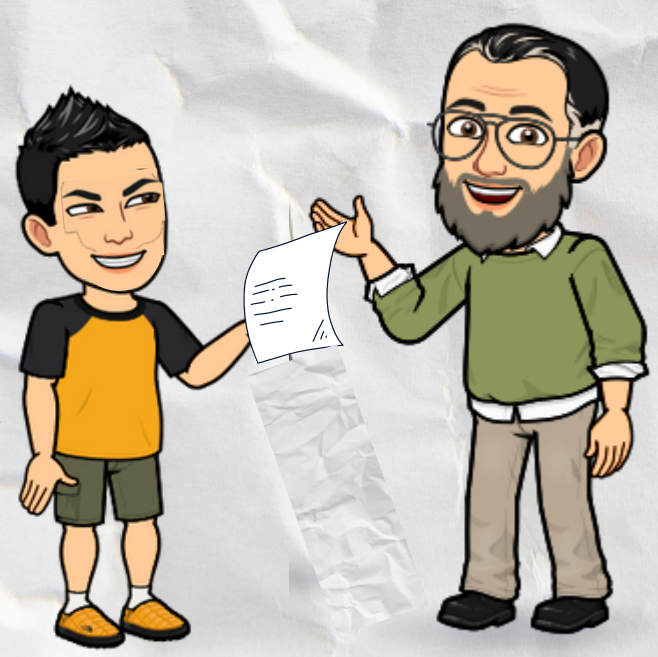
In short, have more training data so that the ML Model (Nandu and Kishori) can't memorize the complete training data which can easily lead to overfitting.

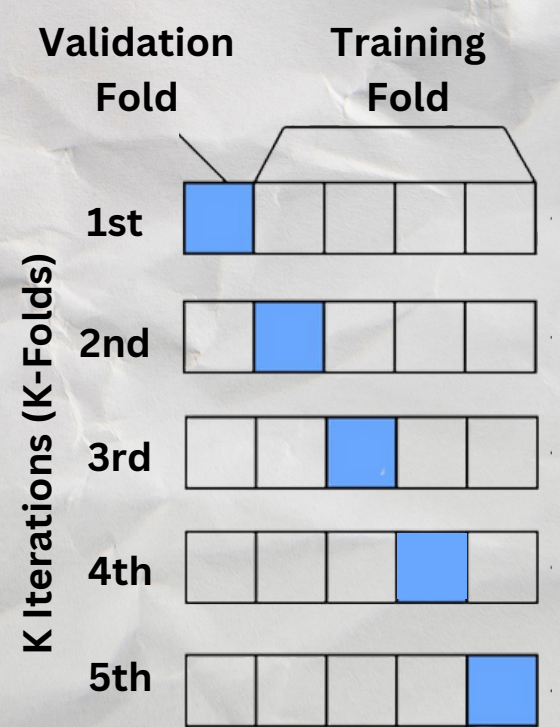


Session With Sumit

Cross-Validation

Remember Nandu handing the same sample papers he prepared to Sumit Kaka for the home test?
Now, what if Sumit Kaka would have used different sample papers? Do you think, Nandu would have performed the same?
No, right?
That's what cross-validation is.
We train the ML Model in one set of data and use another set to test the model.
Cross-Validation is the process in which we test the model on various test samples which lets us understand if the model is a generalized model or not.
Here's the illustration.





Session With Sumit

Session With Sumit

Data augmentation (data)

There are cases in which it's impossible to generate or get more data, in that case, what can we do?

Simple generate artificial data.
There are various ways to generate artificial data, such as

1. Random Oversampling
2. Random Undersampling
3. SMOTE

>>>

Note: We will be covering the following topics in detail in our next post. Please follow us so that you don't miss out on any details.

Session With Sumit

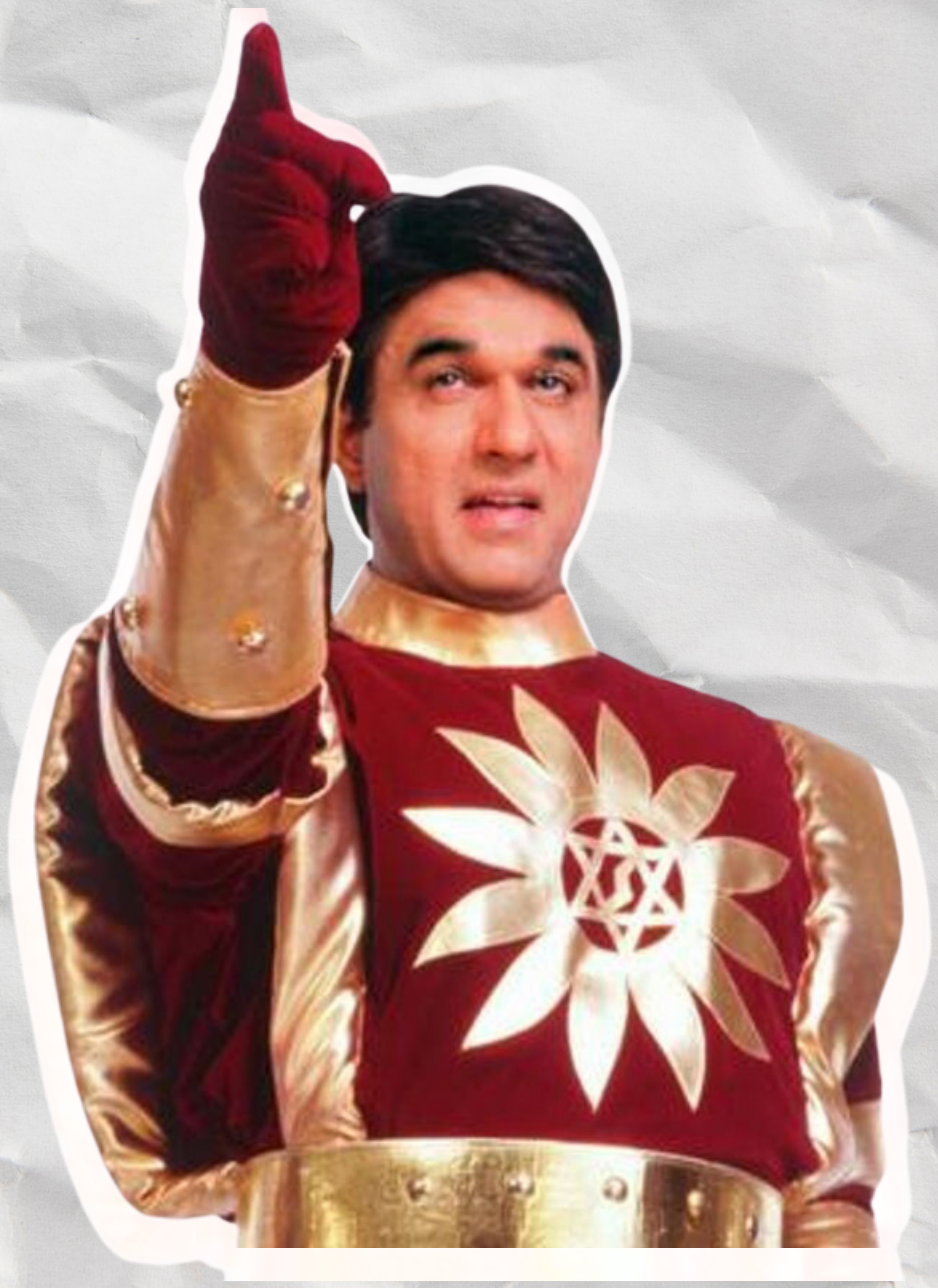
Regularization

Regularization is a technique used in machine learning to prevent the overfitting of a model. It adds a penalty term to the model's objective function, which discourages the model from becoming too complex and helps it generalize better to new data.

Think of it like adding a speed limit to a car to prevent it from going too fast and crashing. Regularization helps keep the model in check and prevents it from "crashing" on new data.

Note: We will be covering the following topics in detail in our next post. Please follow us so that you don't miss out on any details.

If this post helped you to learn the concept of overfitting



Like it, Share it in your network, and Follow Sumit for more.

Session With Sumit

Don't Forget To

SAVE

&

Follow For More

Session With Sumit

This post is dedicated to my cute kittens

Nandu

Kishori

Session With Sumit