OVERFITTING AND UNDERFITTTING

Over fitting:

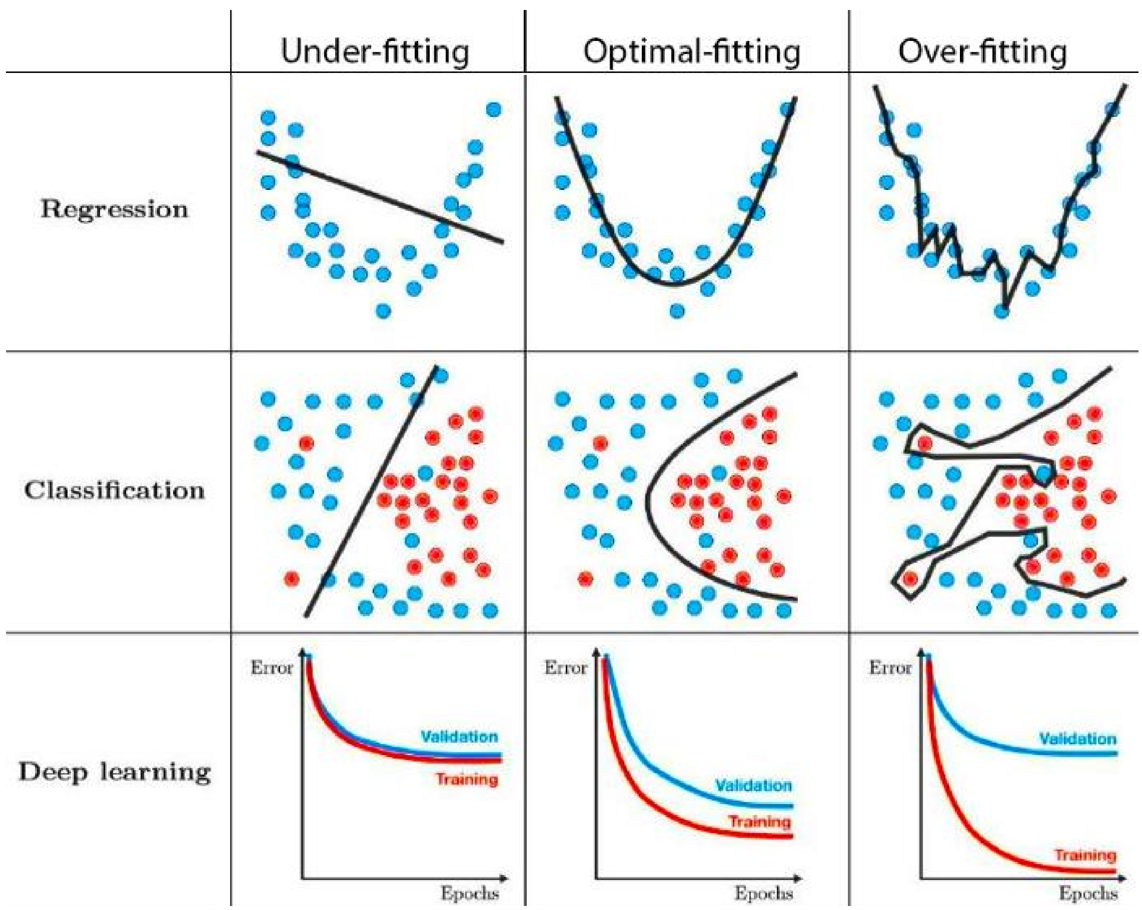
When a model learns the pattern and noise in the data to such extent that it hurts the performance of the model on the new dataset, is termed over fitting. The model fits the data so well that it interprets noise as patterns in the data.

Under fitting:

When the model neither learns from the training dataset nor generalizes well on the test dataset, it is termed as under fitting. This type of problem is not a headache as this can be very easily detected by the performance metrics. If the performance is not good to try other models and you will certainly get good results.

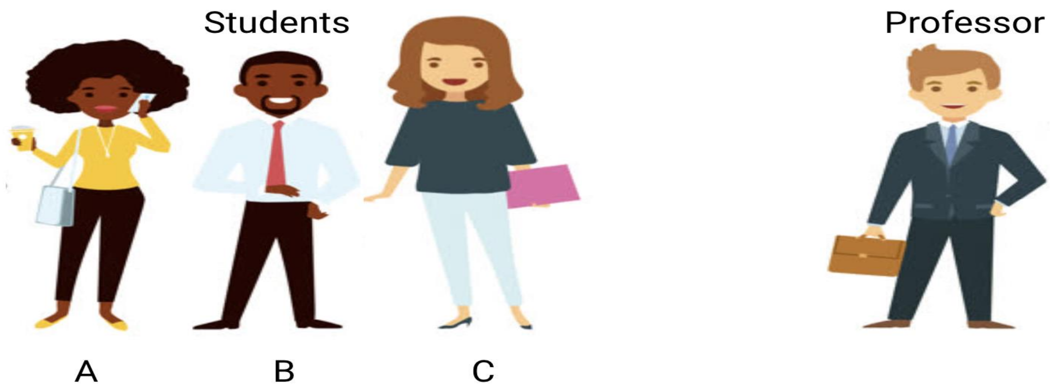
Good Fit:

The spot in the middle of under fitting and over fitting is the good fit.



Let’s Take an Example to Understand Under fitting vs. Over fitting:

- Consider a math class consisting of 3 students and a professor.



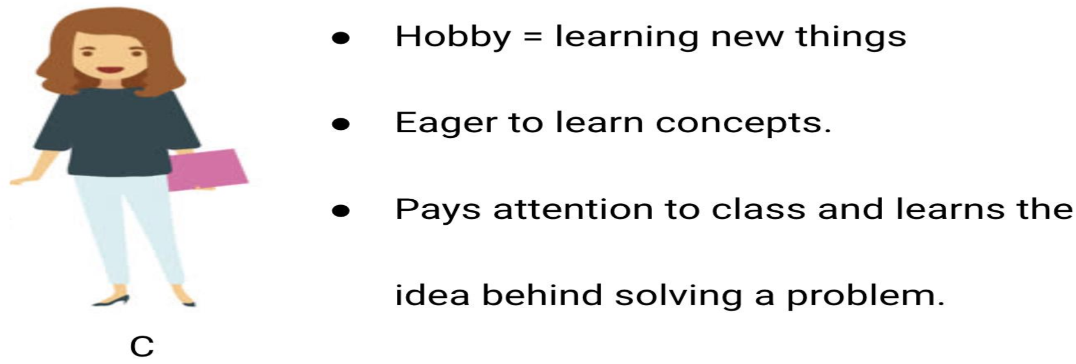
- Now, in any classroom, we can broadly divide the students into 3 categories. We’ll talk about them one-by-one.

- Hobby = chating
- Not interested in class
- Doesn’t pay much attention to professor

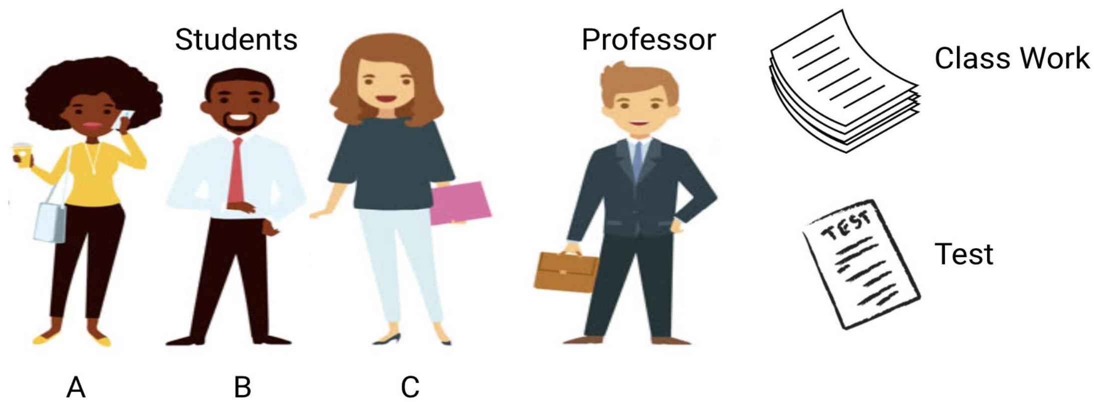
- Let’s say that student A resembles a student who does not like math. She is not interested in what is being taught in the class and therefore does not pay much attention to the professor and the content he is teaching.

- Hobby = to be best in class.
- Mugs up everything professor says.
- Too much attention to the class work.

- Let’s consider student B. He is the most competitive student who focuses on memorizing each and every question being taught in class instead of focusing on the key concepts. Basically, he isn’t interested in learning the problem-solving approach.



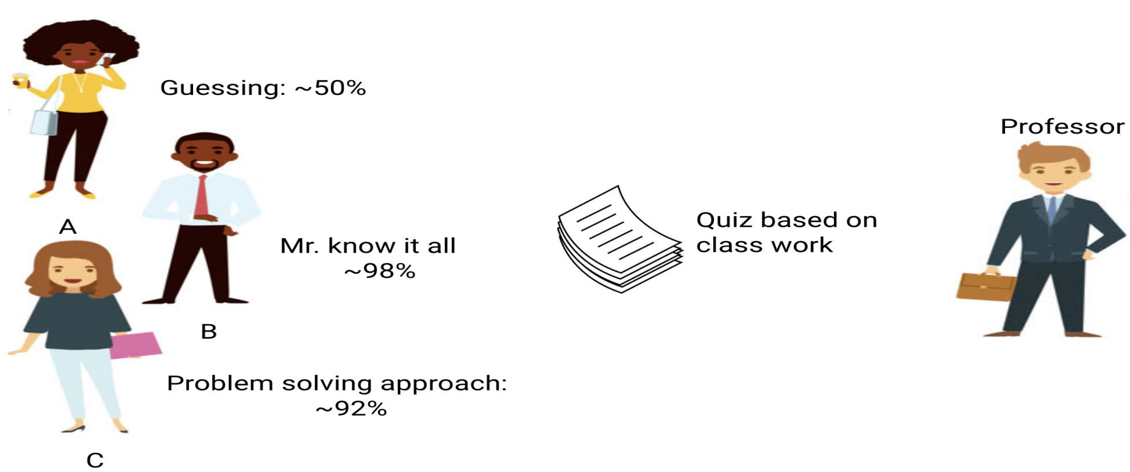
❖ Finally, we have the ideal student C. She is purely interested in learning the key concepts and the problem-solving approach in the math class rather than just memorizing the solutions presented.



❖ The professor first delivers lectures and teaches the students about the problems and how to solve them. At the end of the day, the professor simply takes a quiz based on what he taught in the class.

❖ The obstacle comes in the semester3 tests that the school lays down. This is where new questions (unseen data) comes up. The students haven't seen these questions before and certainly haven't solved them in the classroom. Sounds familiar?

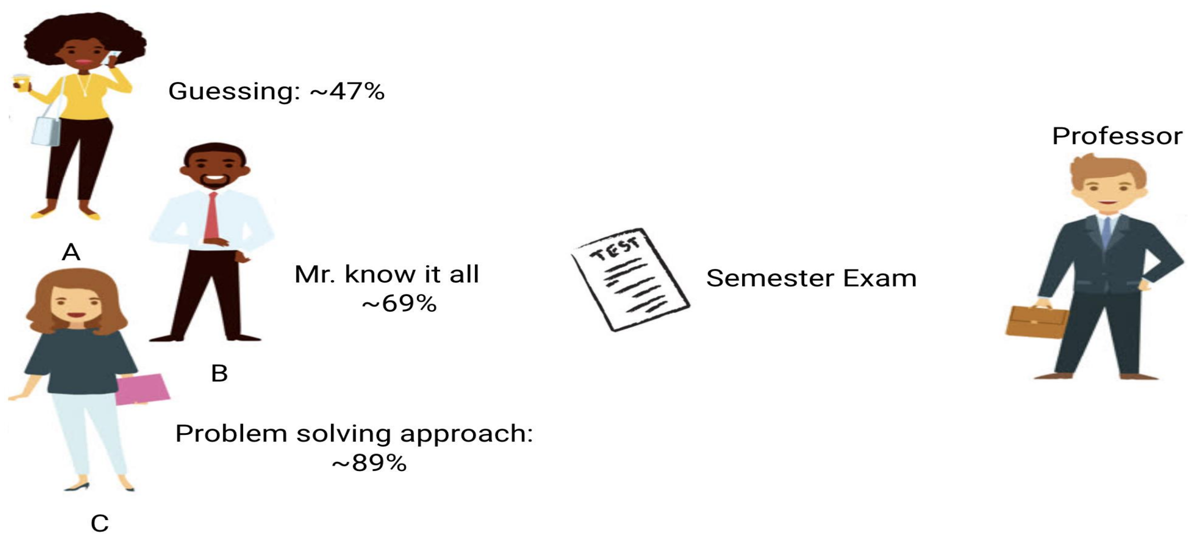
So, let's discuss what happens when the teacher takes a classroom test at the end of the day:



- Student A, who was distracted in his own world, simply guessed the answers and got approximately 50% marks in the test
- On the other hand, the student who memorized each and every question taught in the classroom was able to answer almost every question by memory and therefore obtained 98% marks in the class test
- For student C, she actually solved all the questions using the problem-solving approach she learned in the classroom and scored 92%

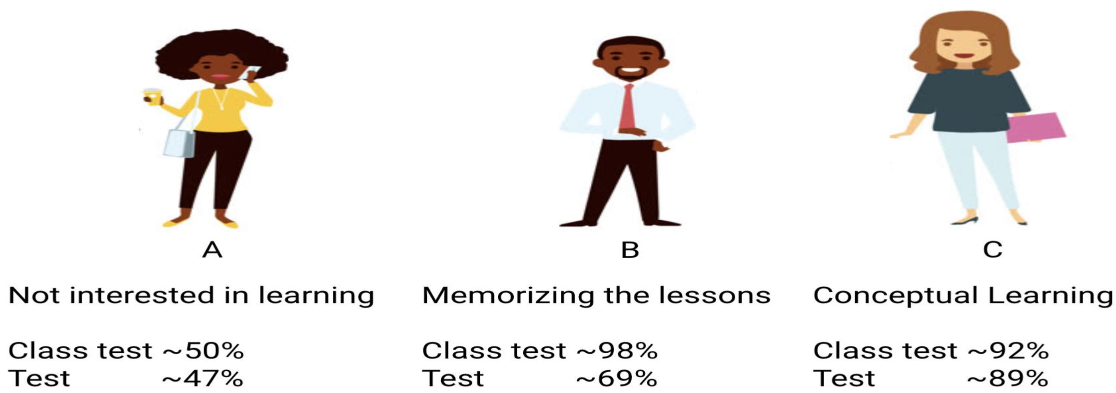
We can clearly infer that the student who simply memorizes everything is scoring better without much difficulty.

Now here's the twist. Let's also look at what happens during the monthly test, when students have to face new unknown questions which are not taught in the class by the teacher.



- In the case of student A, things did not change much and he still randomly answers questions correctly ~50% of the time.
- In the case of Student B, his score dropped significantly. Can you guess why? This is because he always memorized the problems that were taught in the class but this monthly test contained questions which he has never seen before. Therefore, his performance went down significantly
- In the case of Student C, the score remained more or less the same. This is because she focused on learning the problem-solving approach and therefore was able to apply the concepts she learned to solve the unknown questions.

❖ <u>It's similar process for our Machine Learning Process:</u>



➢ As we do in our dataset after splitting training and test dataset is:

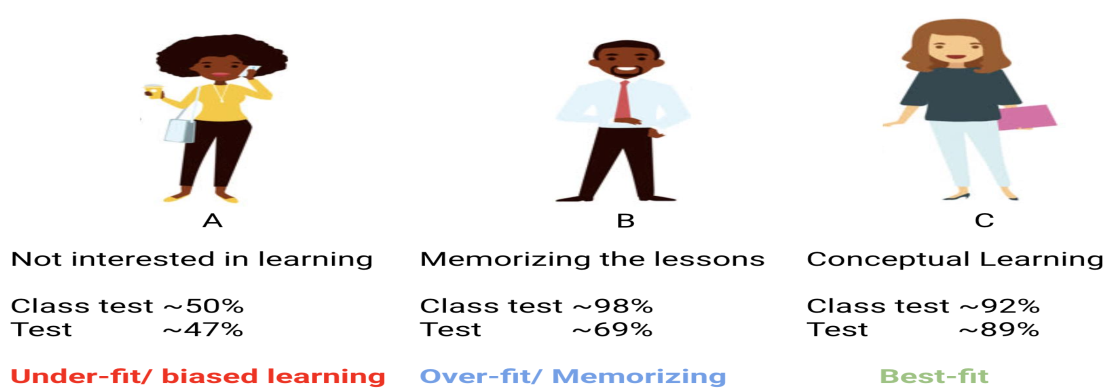
❖ Here we can see that the class work is like the training data what we are learning and test is like our test data in which we don't know what questions will come (unseen).

❖ **This situation where any given model is performing too well on the training data but the performance drops significantly over the test set is called an over fitting model.**

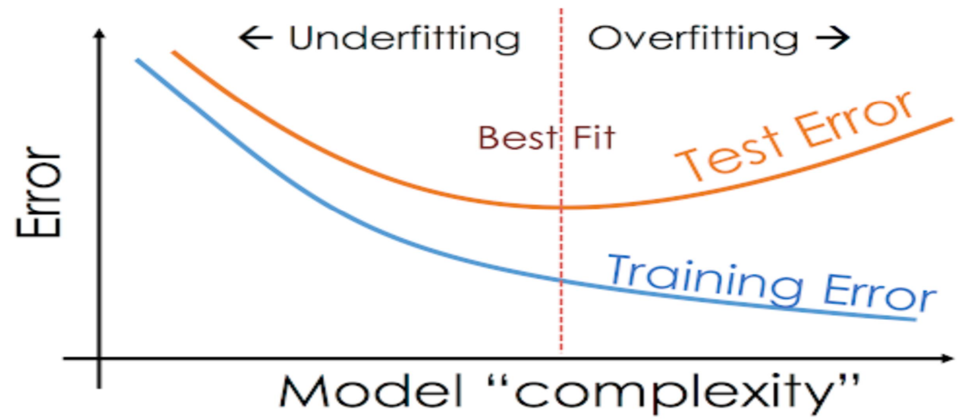
❖ We can see that the 'A' is the perfect example of 'Under Fitting' .

❖ 'B' is the perfect example of 'Over Fitting' .

❖ 'C ' is the perfect example of 'Best Fit'.



For example, non-parametric models like decision trees, KNN, and other tree-based algorithms are very prone to over fitting. These models can learn very complex relations which can result in over fitting. The graph below summarises this concept:



On the other hand, if the model is performing poorly over the test and the train set, then we call that an under fitting model. An example of this situation would be building a linear regression model over non-linear data.

****