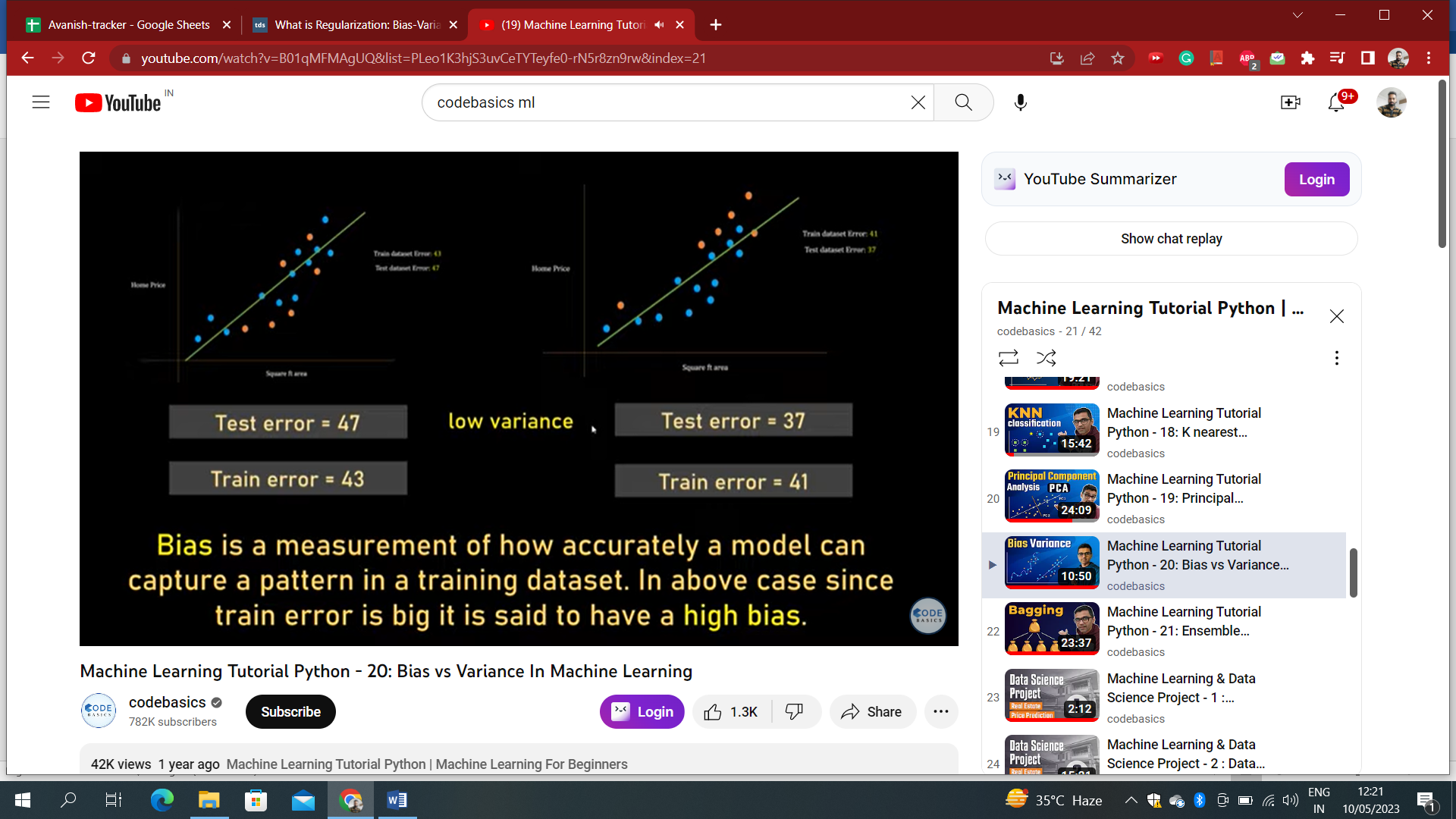
**Under fitting and overfitting**

## **What is under fitting?**

Under fitting occurs when a model can’t capture and generalize the underlying trend of the data. Intuitively, it didn’t fit the data well enough, as a result, it would produce high errors on both **training**and **testing**data (i.e., **high bias**).

The cause of under fitting could be

* the lack of data (e.g., a small amount of training data, unavailability of key explanatory variables) to develop a model.
* the underlying model is NOT capable of capturing the patterns in the data (e.g., using a linear model with nonlinear data)

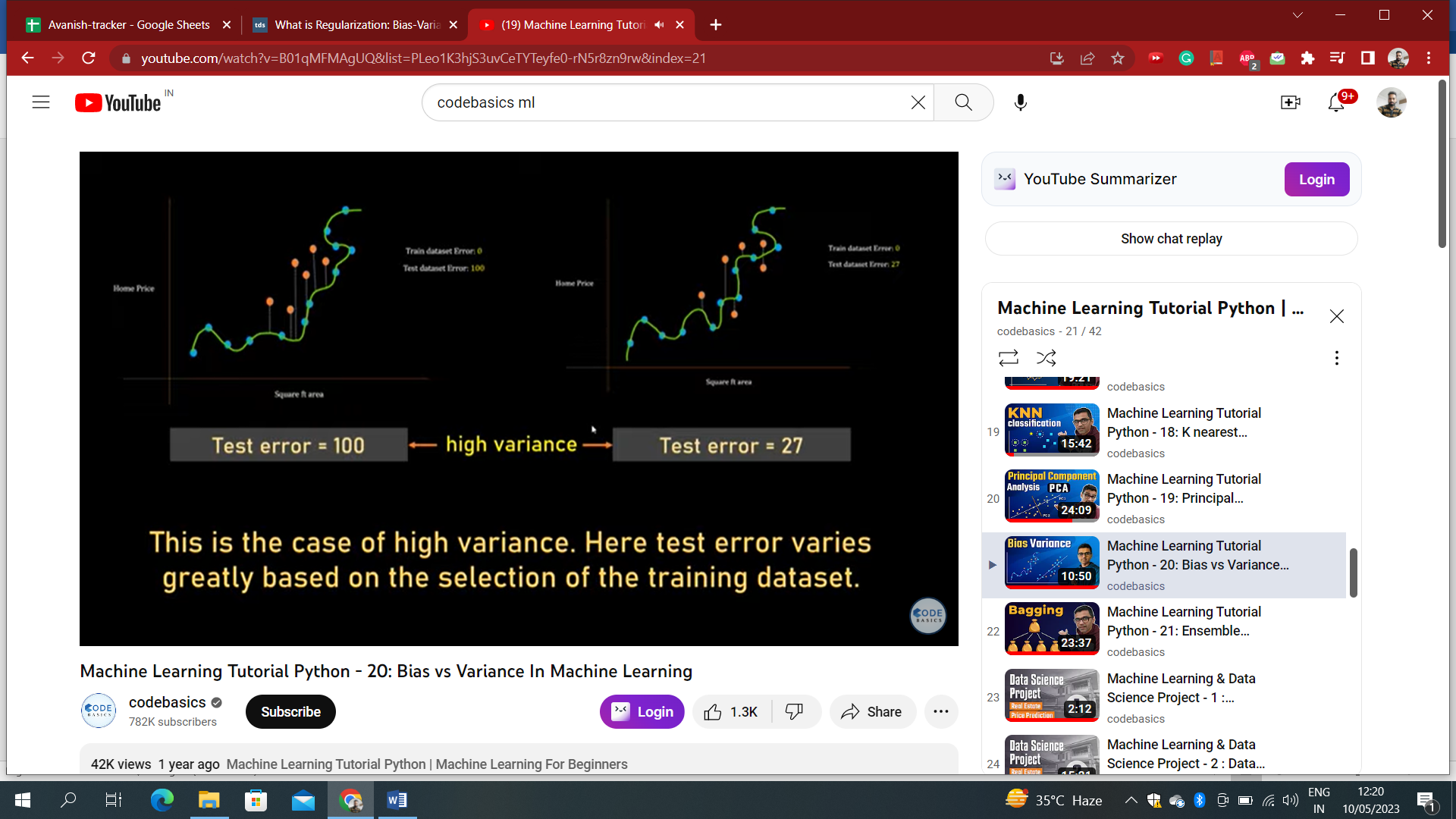


## **What is overfitting?**

On the other hand, overfitting occurs when a model is fitting the training too well and it **starts to model the noise** of the training data. It would have a low error in training data (i.e., **low bias**) but a high error in the testing data (i.e., **high variance**)

The cause of overfitting could be

* due to an excessively complicated model which includes**too many variables**or includes **problematic variables**, such as higher-order polynomial variables, irrelevant variables, variables that are highly correlated, or variables that are affected by the response variable simultaneously.
* **overtraining**of a model. For example, training a decision tree without explicit stopping criteria is prone to overfitting.



## **Balanced Fit**

When the model is balanced fit, it generates low bias and low variance. In this case, both train and test errors are low.

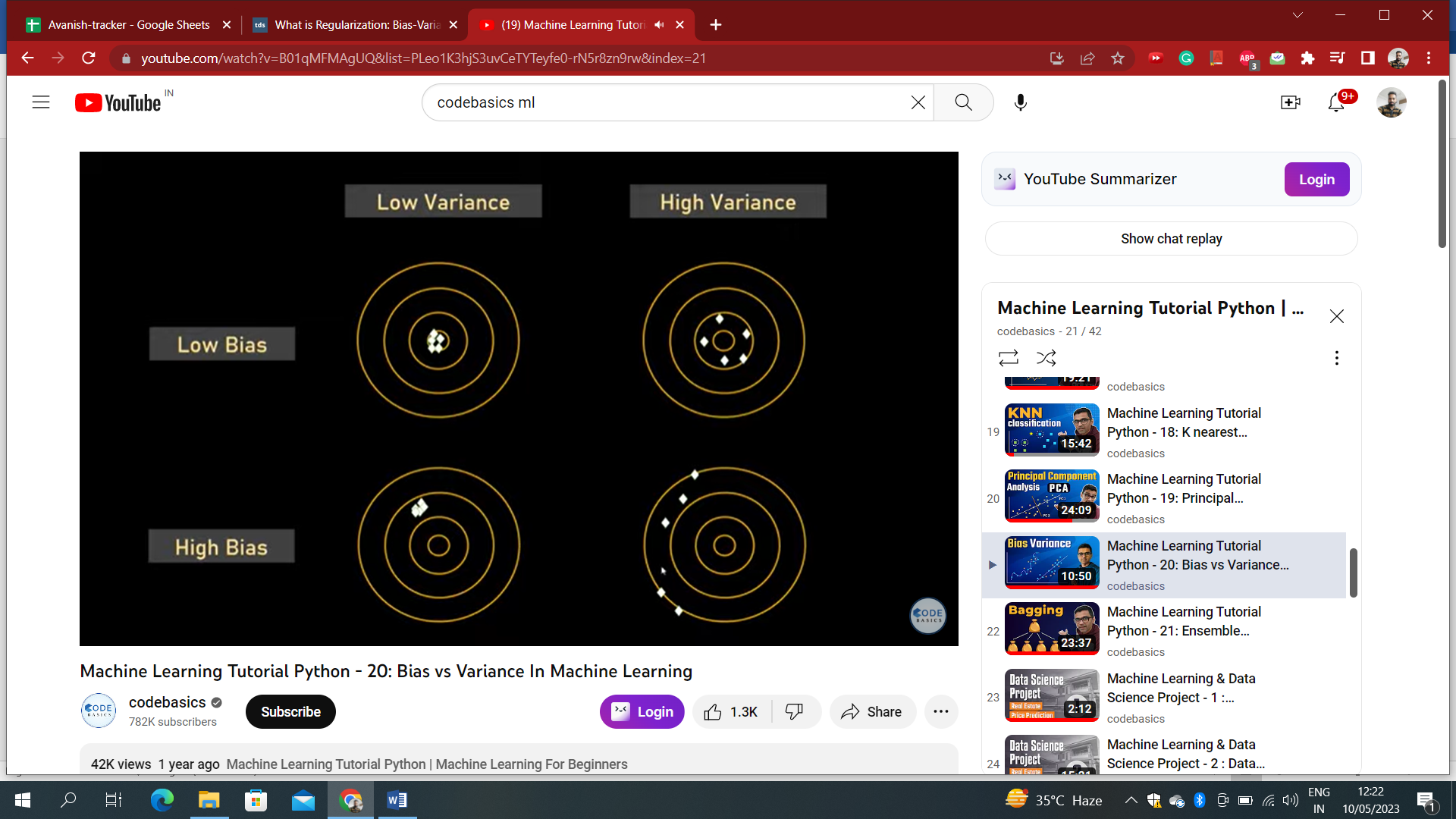
## 

## **Bias-Variance Tradeoff**

Ideally, we would like to reduce both bias and variance of a model. However, it is very difficult and sometimes impossible to achieve. As you’re trying to reduce the prediction error on the training data, the prediction error on the testing data might increase. Bias and variance oftentimes are moving in **opposite directions**.

When developing a model with the training data, it is very easy to fit a complex model to make predictions with small bias, but might not create an optimal result due to the high prediction variance on the testing data.

Alternatively, we can explore the trade-off between bias and variance, where it might be better off **accepting some bias for a reduction of variance**. Consequently, we would have smaller prediction errors on the unseen data.



Central circle represents the truth and white dots representing the predicted value

**Ways to get balanced fit model:**

1. Cross Validation
2. Regularization
3. Dimensionality Reduction
4. Ensemble Techniques