**Imbalanced Class handling**

**Balanced dataset-** there is roughly equal number of classes in the target variable.

**Imbalanced dataset-** there is more unequal number of classes in the target variable.

Binary Classification problem-

Example- In a credit card fraud detection dataset;

Total Observations = 1000

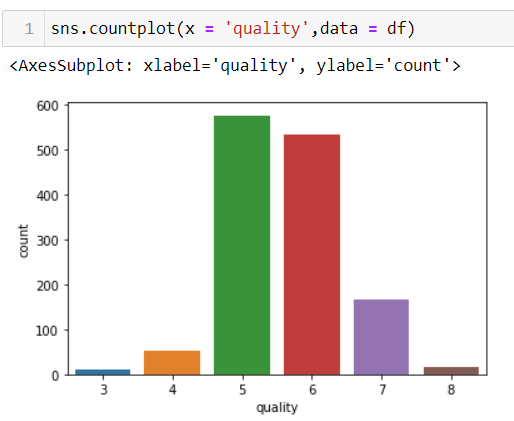
Fraudulent Observations = 20

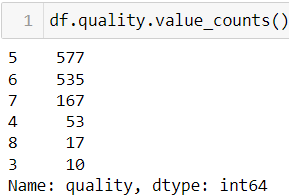
Non Fraudulent Observations = 980

Multiclass Classification problem-

Unequal number of classes in the target variable.

For example, wine quality prediction model in which quality is rated 0 to 10 (dataset credit: UCI ml repository).





**Challenges faced with imbalanced dataset:**

1. Machine learning algorithms may get biased towards the majority class and thus tend to predict output as the majority class.
2. The features of the minority class are treated as noise and are often ignored.
3. Imbalanced dataset results in false accuracy score due to there is a high probability of misclassification of the minority class as compared to the majority class. Thus, **ROC** or **Recall** is used as evaluation matrix for model performance.

**Imbalanced dataset handling:**

This can be done in two ways:

1. By balancing classes in the training data (preferred over second method due to it’s wider application)
2. By improving classification algorithms

**Class Balancing Techniques-**

The main goal of class balancing is to either increase the frequency of the minority class or decrease the frequency of the majority class or a combination of both i.e. resampling them. This is done to get approximately the same number of instances for both classes. Let's look at a few resampling techniques:

**1. Oversampling:**

1. Random oversampling
2. Synthetic Minority Oversampling Technique(SMOTE)
3. Adaptive Synthetic Sampling(ADASYN)

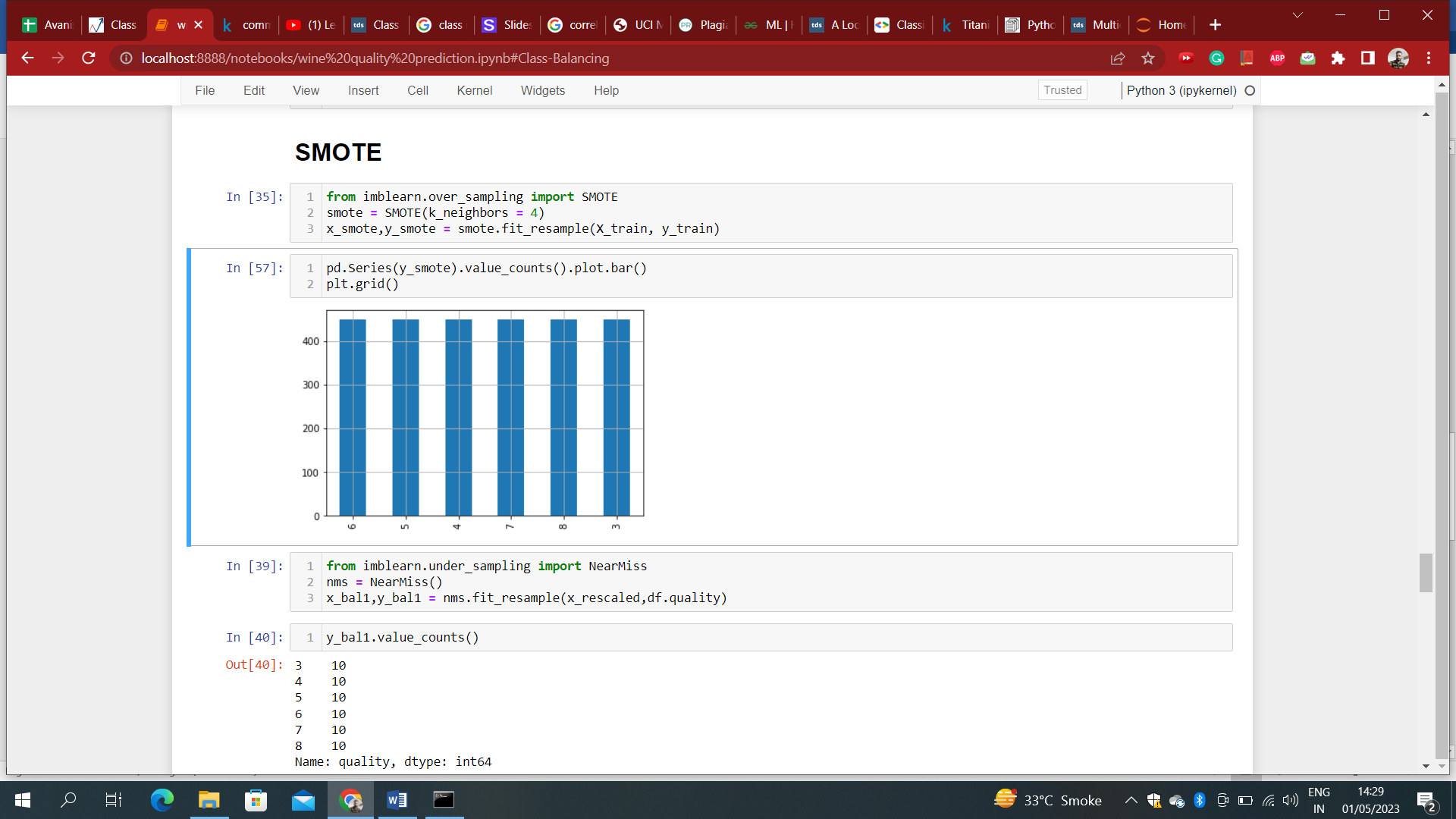
**2. Under sampling:**

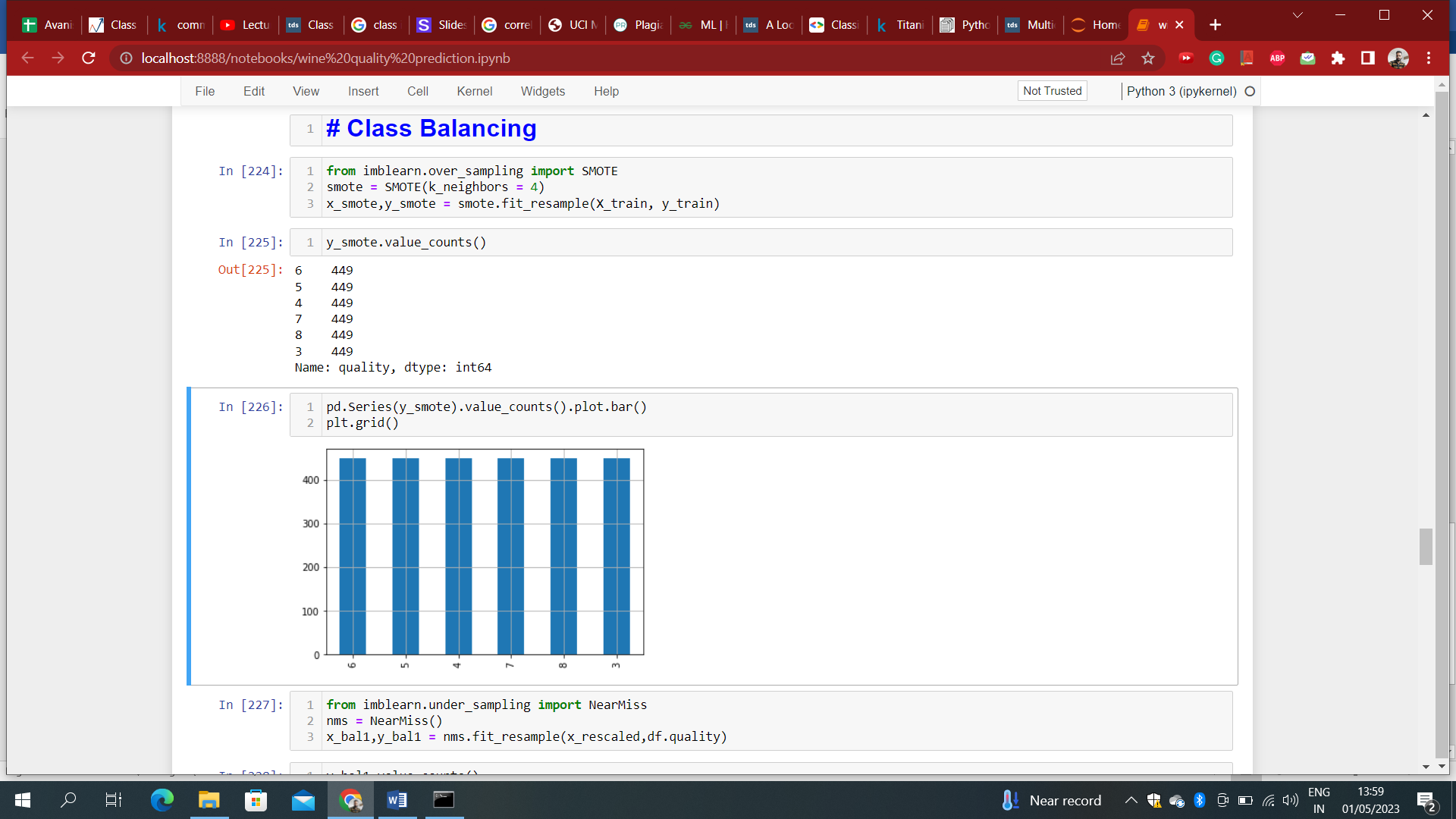
1. Random under sampling
2. Near Miss Under sampling
3. Tomek Links under sampling

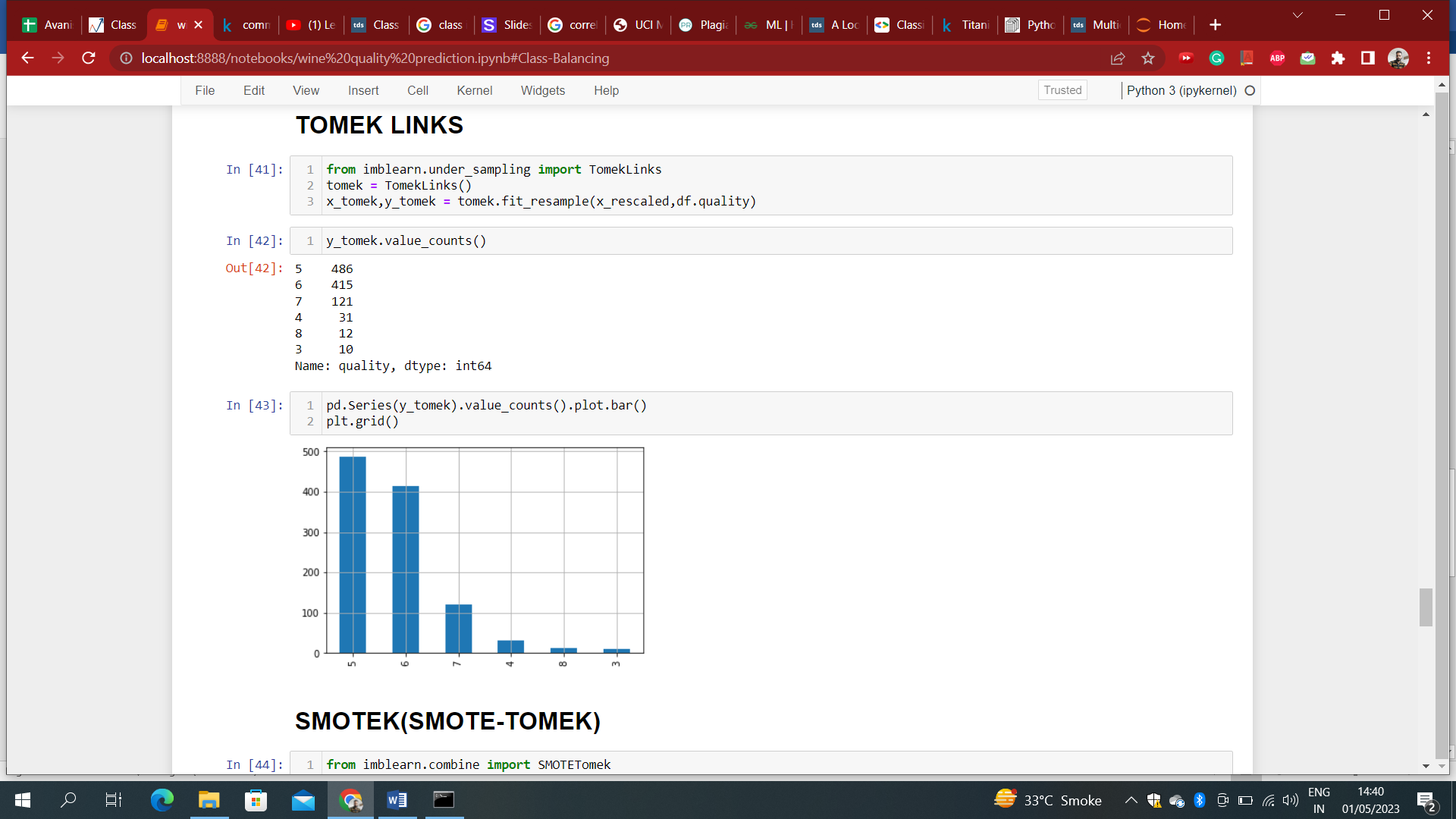
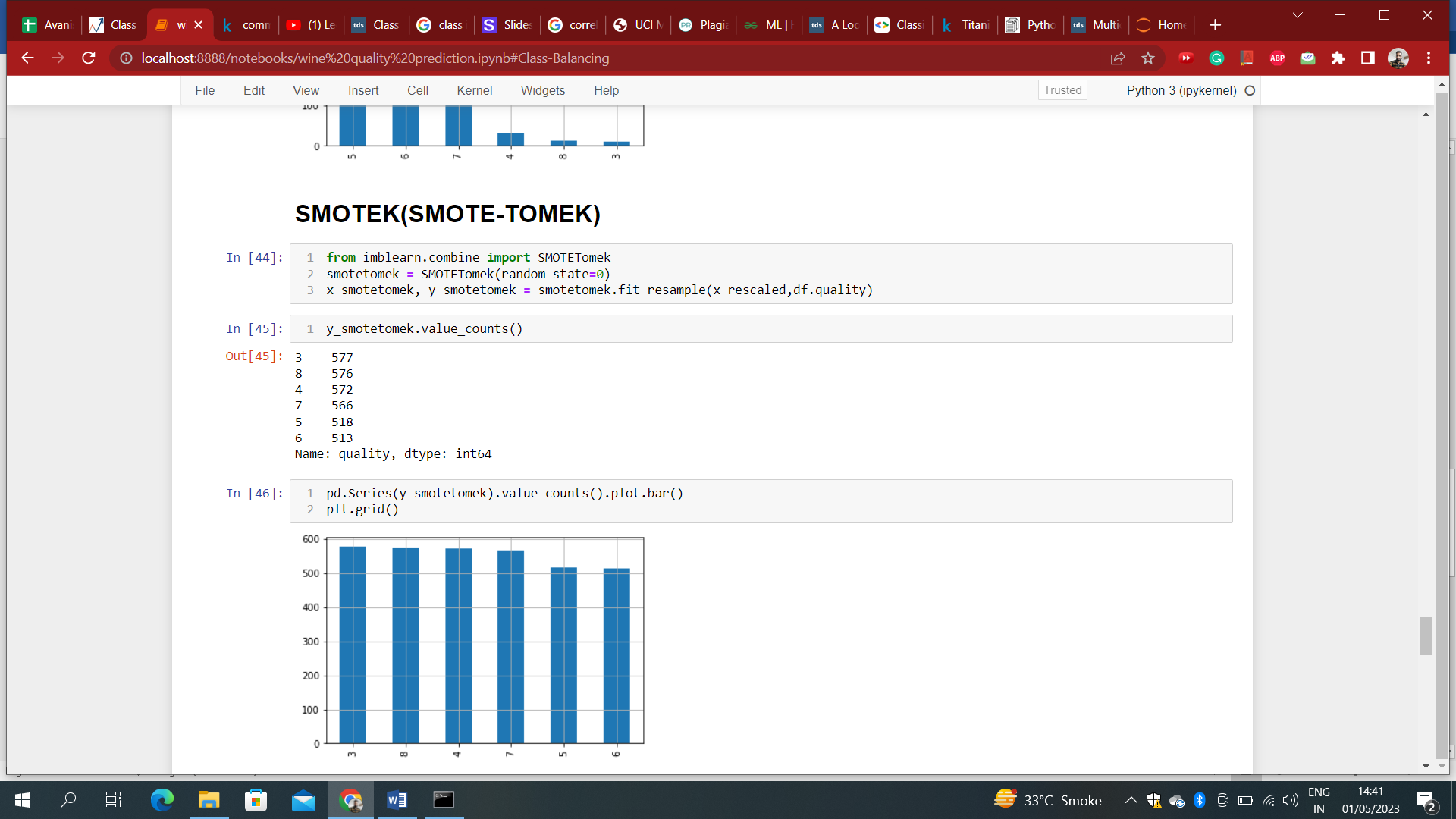
**3. Combination of oversampling and under sampling:**

1. SMOTEK( or **SMOTE-TOMEK)**

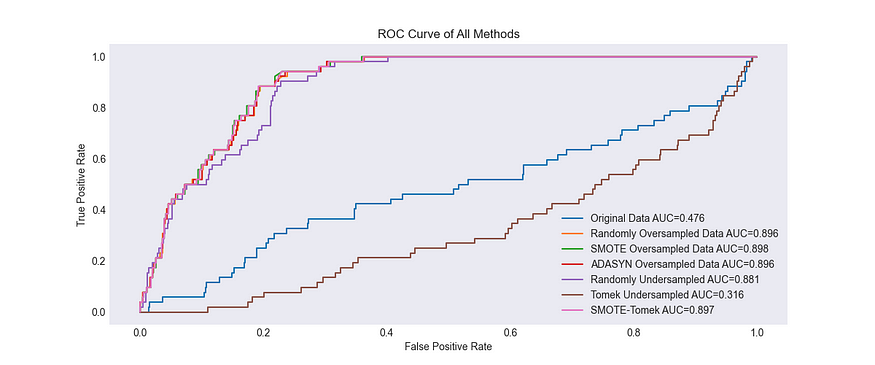
**In the below picture, we could see that how the target variable is balanced using SMOTE (an oversampling method), Tomek Links(an under sampling method) and SMOTEK(SMOTE-TOMEK) which is a combination of both i.e. first oversampled using SMOTE and then under sampled using TOMEK.**





For the grand finale, we will compare all the techniques that we have noted above. Viola, SMOTE-TOMEK performed the best.



**Conclusion:**

Overall, you can use oversampling, under sampling or a combination of both to deal with data imbalance. If you have the computational resources, it is often better to use a combination of over- and under-sampling; Oversampling is a good strategy when you have few data points; while under sampling is good when there are potentially many similar data points.