# Overview of what each thing is doing

### Load and Process Data

Load data from the CSV provided by Dr. Rueda

Transpose the data (rows need to be columns and vice versa to allow for transcripts/genes to be the features)

Remove features that only contain 0 values

Add target variable – Grade Group (based on Gleason score and primary/secondary patterns):

1 – Gleason score <=6

2 – Gleason score == 7 with primary pattern == 3

3 – Gleason score == 7 with primary pattern == 4

4 – Gleason score == 8

5 – Gleason score == 9 / 10

Identify feature matrix (Genes) and target matrix (Grade Group)

### Isolation Forest

Used to identify “outlier” samples

A set of random trees are used to “isolate” each sample within the population.

Path length == height of tree == number of decisions required to isolate the sample.

Samples which can be isolated easily (path length noticeably short on average) are deemed to be outliers.

We remove these outliers from our dataset.

### Select K Best

Here we reduce the total number of features to 15000 based on their chi2 score

This is a filter method that will make using more advance feature selection faster / more feasible

### Standard Scaler

Here we fit the data to a normal distribution.

Note: for each sample (x), the new value (z) will be z = (x – u) / s where u is the mean and s is the standard deviation

### Variance Threshold

Here we remove any feature with 0 variance

If all the samples have the same value for a particular feature than that feature will be removed.

### SMOTEENN

Objective here is to balance the majority classes and the minority classes

Performs both over-sampling (SMOTE) and cleaning/under-sampling (ENN – edited nearest neighbours)

**Synthetic Minority Oversampling Technique (SMOTE)**

Originated in 2002 – link for bibliography (<https://arxiv.org/abs/1106.1813> )

Synthesize new samples for the minority class, thus resolving the class imbalance and yielding “new information”

Quoted from (<https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/> ):

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

Advantage is that new samples are different than old ones and also plausible.

Disadvantage is that majority class is not considered (can lead ambiguity with new samples if there is strong overlap between majority class and minority class)

Smote is combined with random undersampling of the minority class

Edited Nearest Neighbour (ENN)

Removes samples from the majority class that are most likely to confuse a classifier (thereby making for easier predictions, and also helping to balance the classes)

From a link (<https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family> ):

2-SMOTEENN: Just like Tomek, Edited Nearest Neighbor removes any example whose class label differs from the class of at least two of its three nearest neighbors. The ENN method removes the instances of the majority class whose prediction made by KNN method is different from the majority class. ENN method can remove both the noisy examples as borderline examples, providing a smoother decision surface.

### Linear Discriminant Analysis

Used for dimensionality reduction (reducing the number of features to 3)

This is potentially a concern?

### Classifiers

Then we perform classification

When the classifiers are run the first time they are run on raw data (X)

When they are run the second time they are performed on data that has had some of the transformations applied (see the pipeline section below)

## Pipeline for First Set of Tests

Run classifiers (there is no preprocessing beyond what is done in “load\_data”

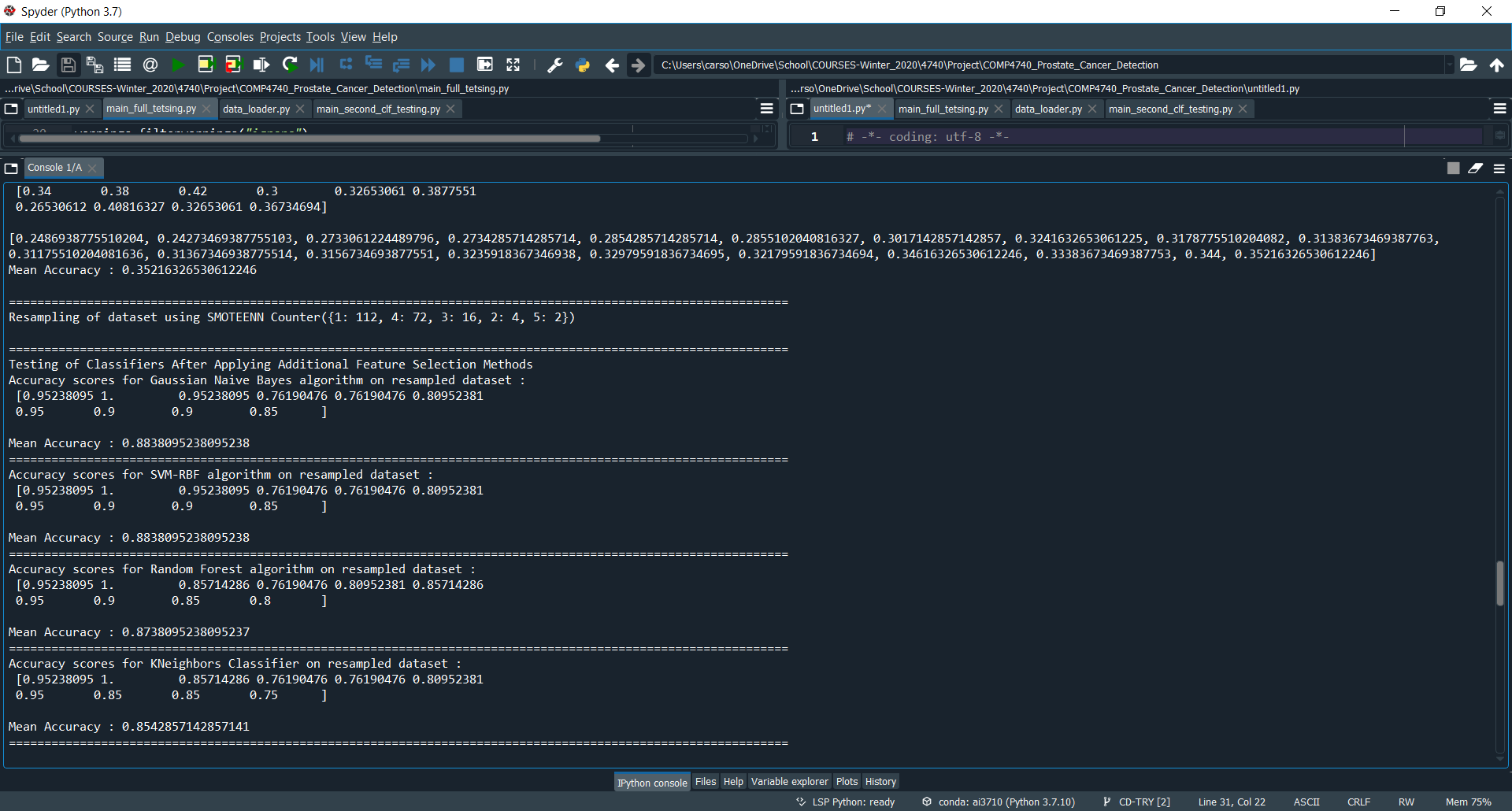
## Pipeline for Second Set of Tests

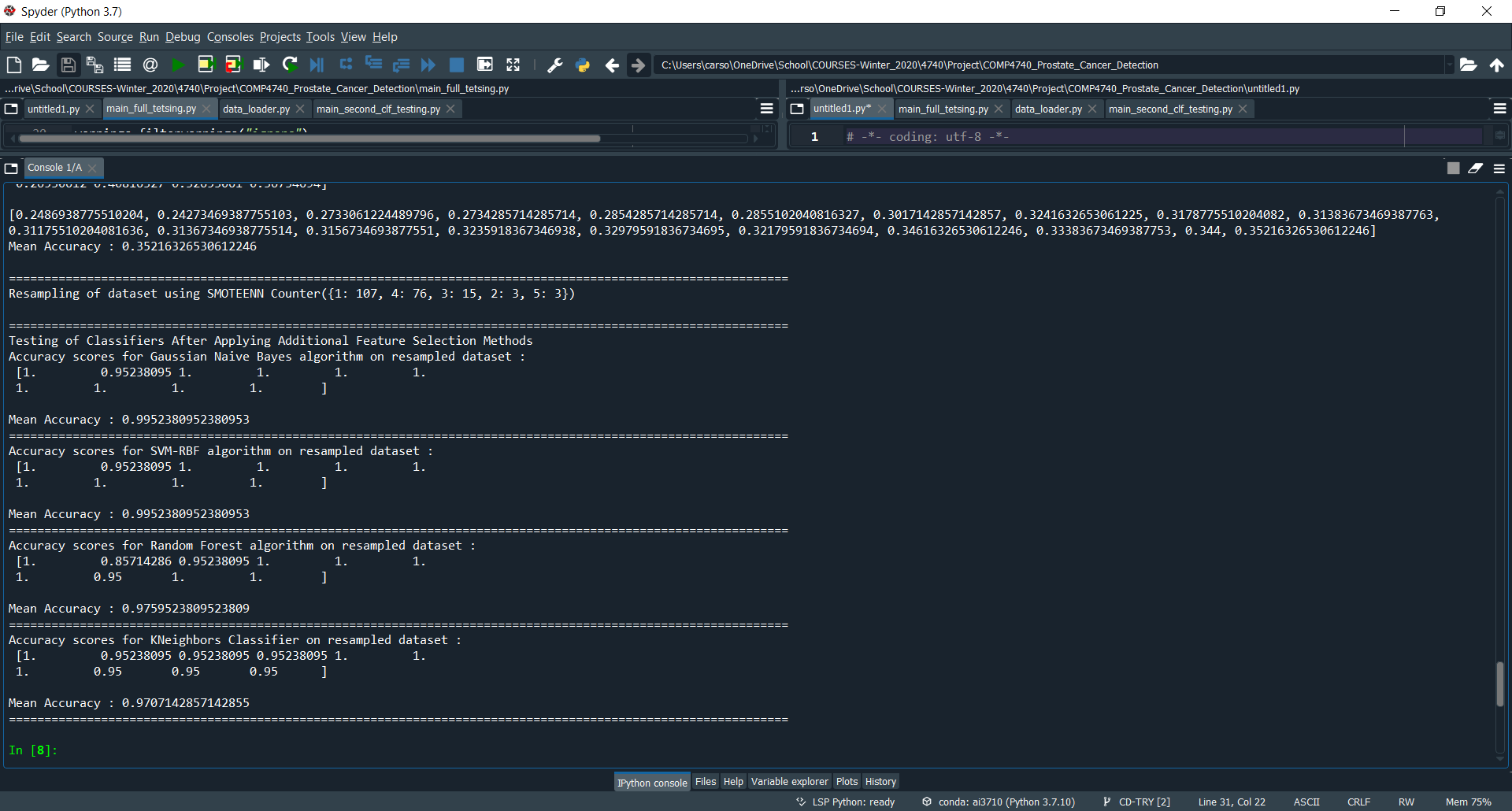
Remove outliers with isolation forest

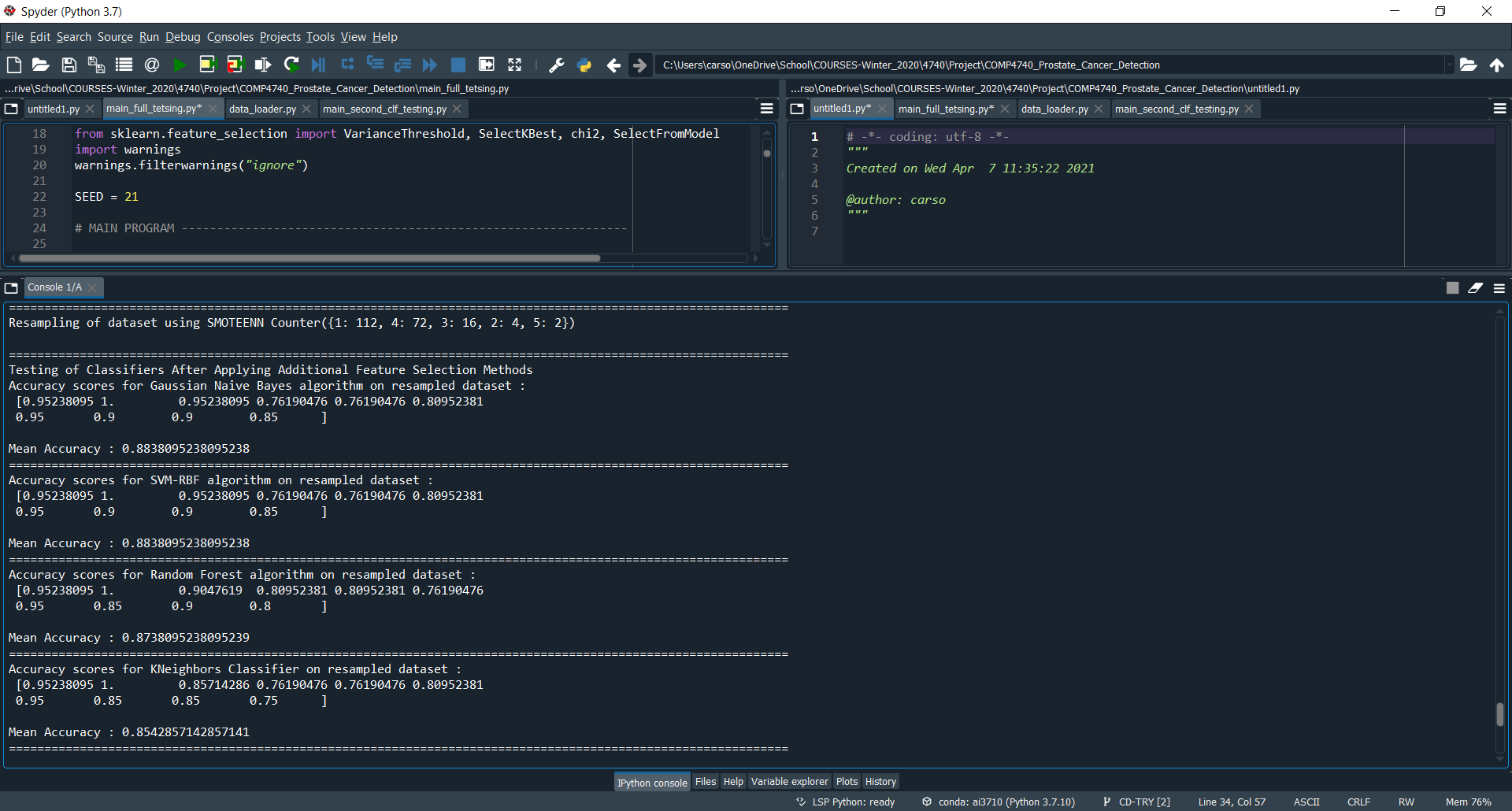
Do resampling with SMOTEEN

Do dimensionality reduction with LDA

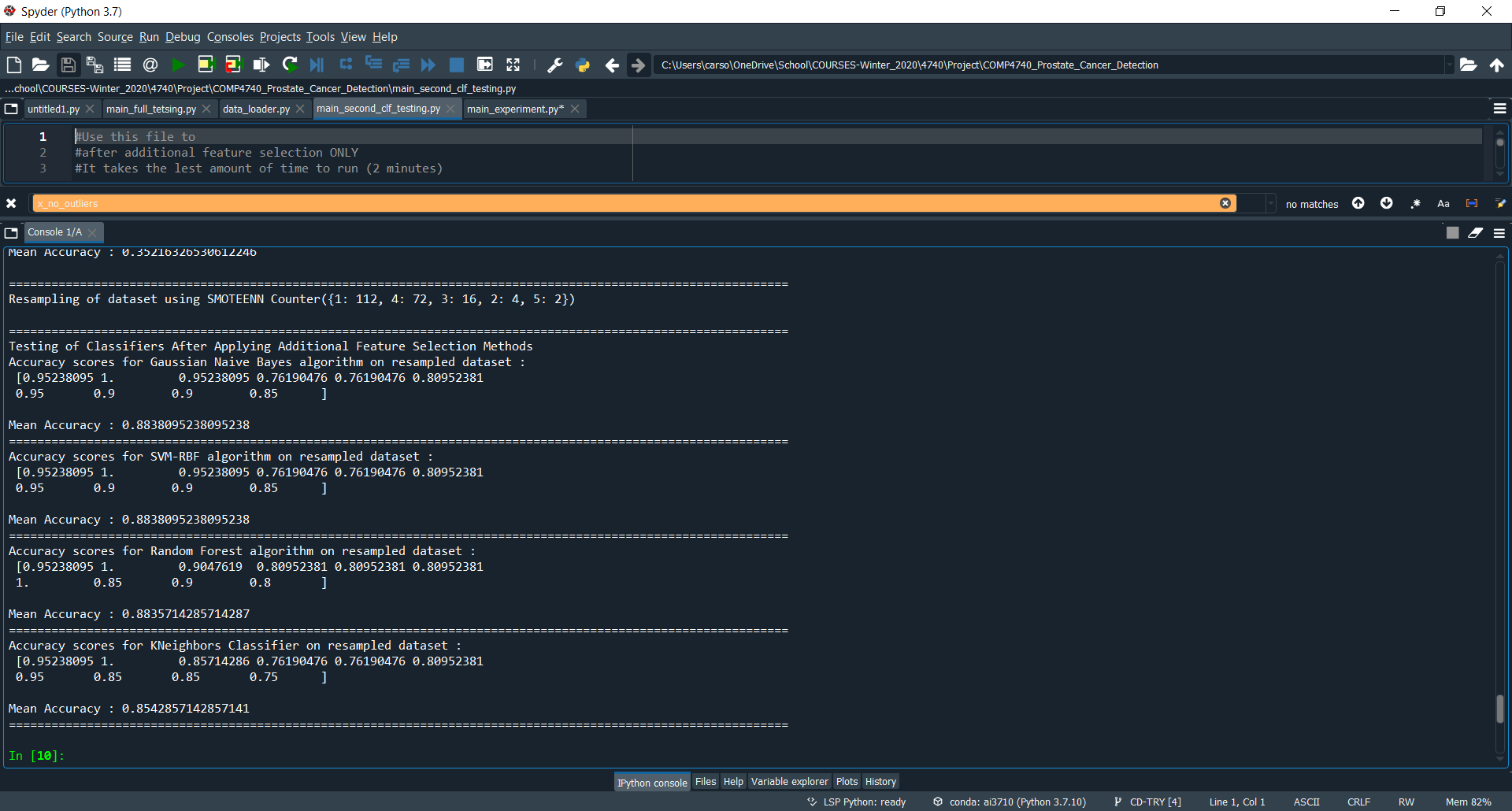
Run classifiers

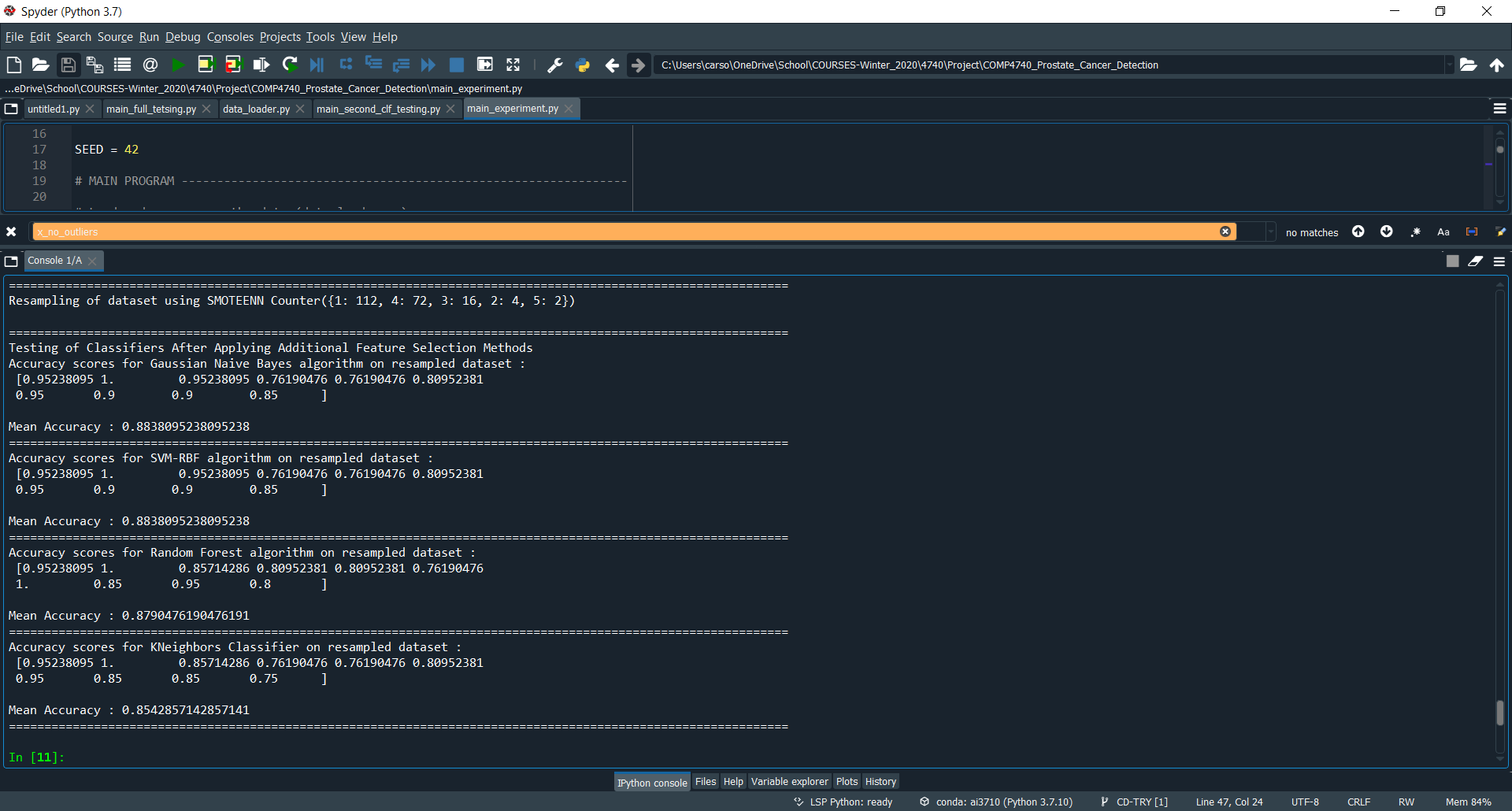




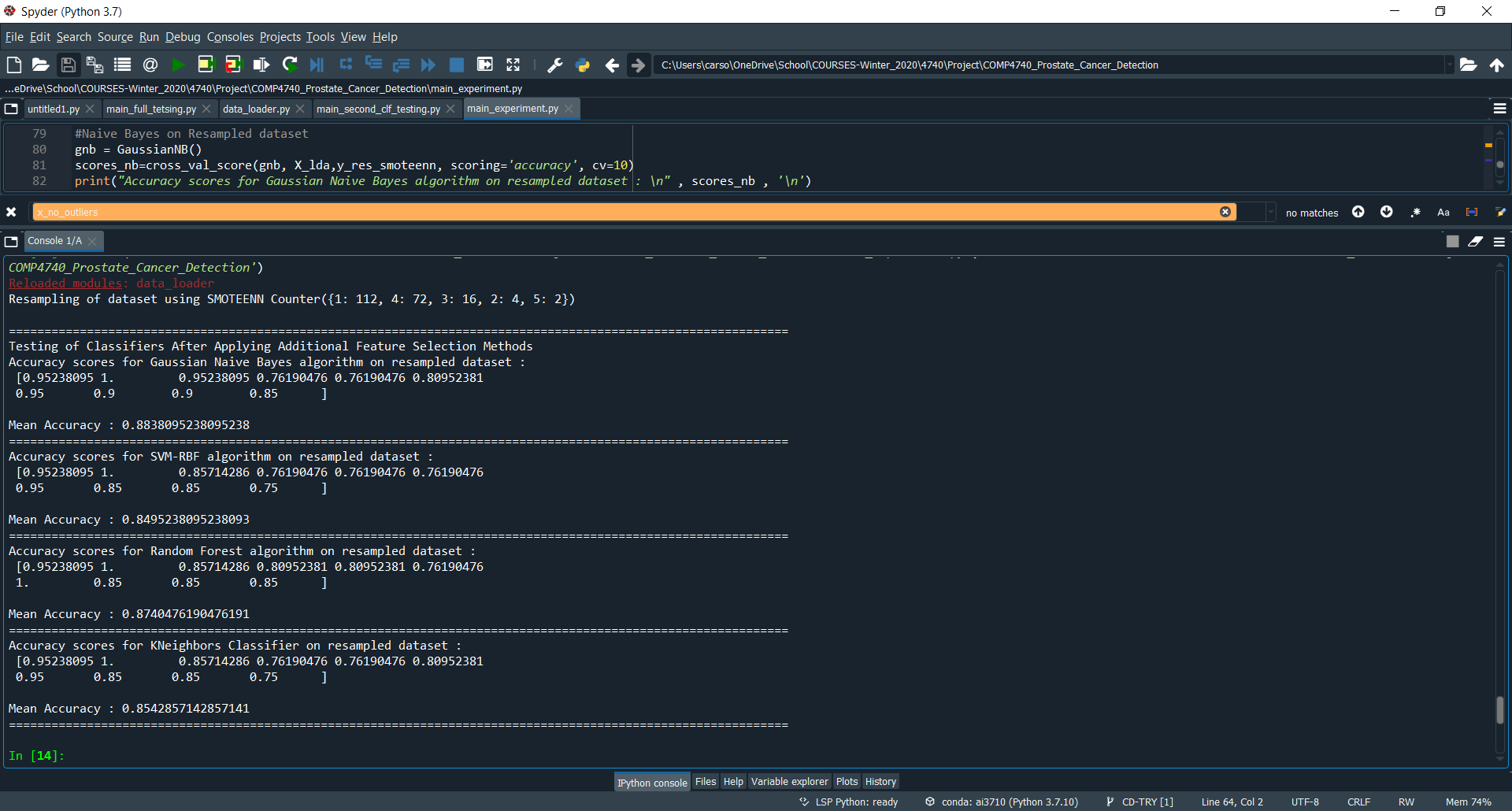


With fixed seeds everywhere:

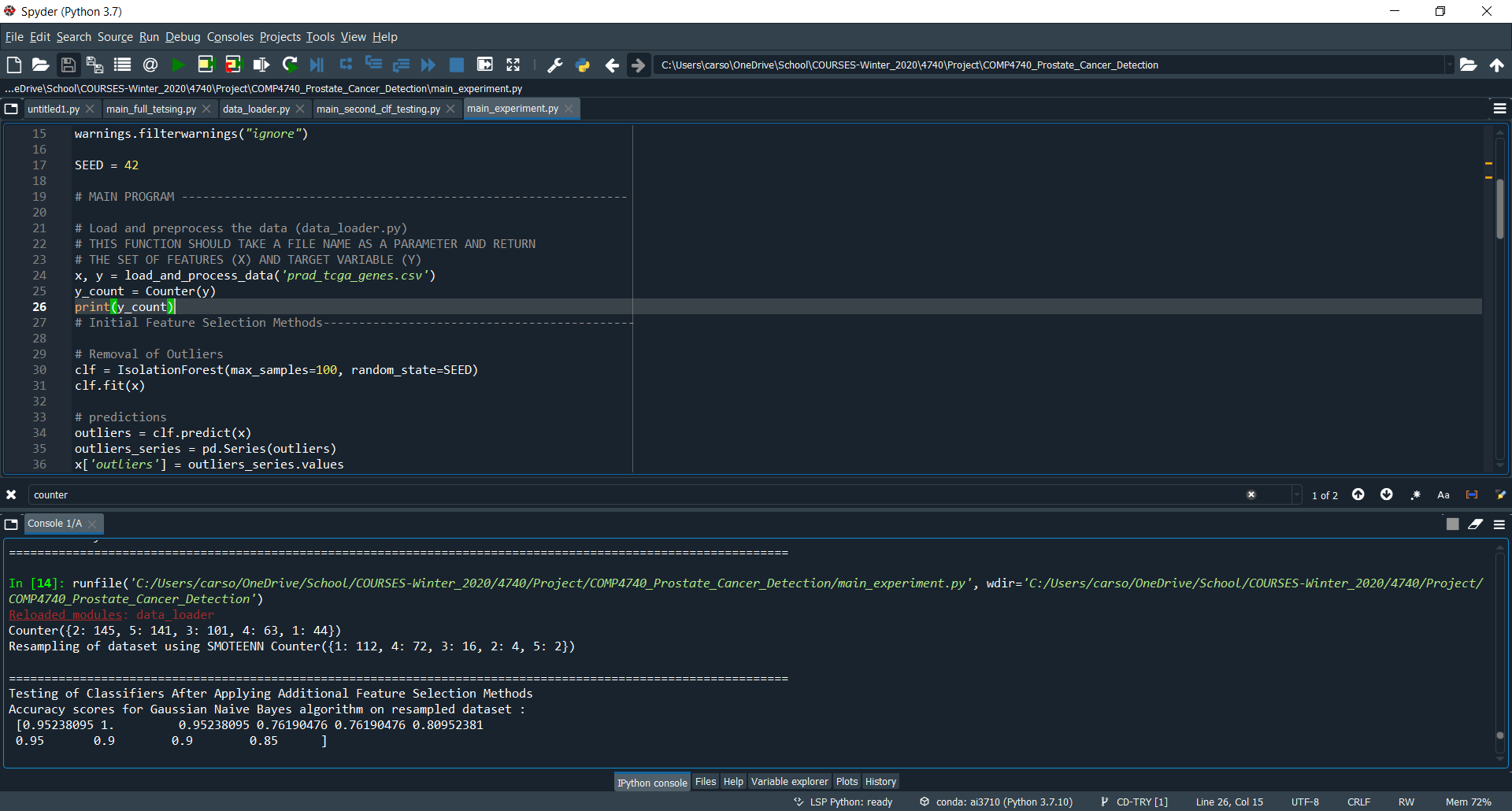




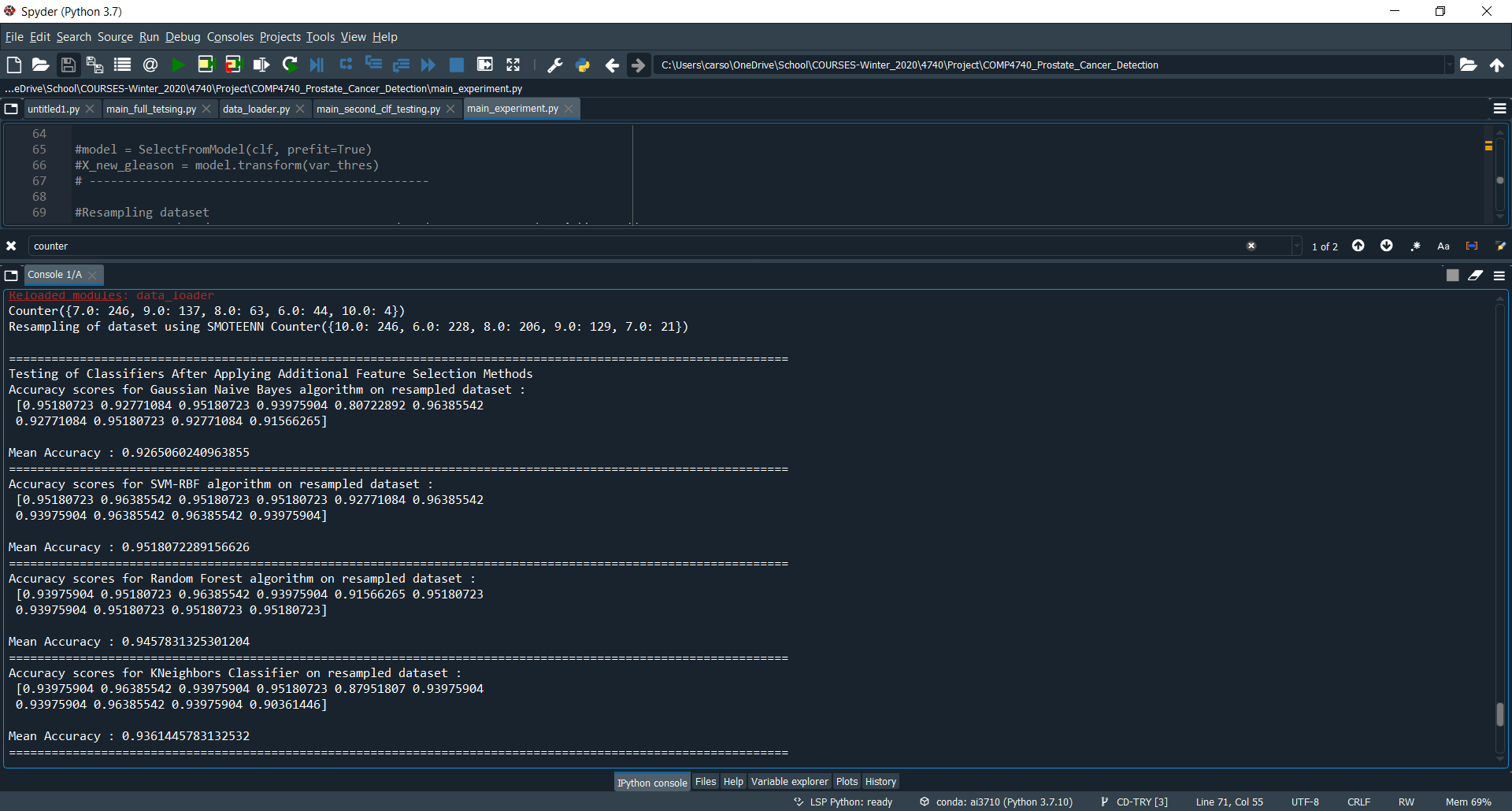
Output where elements not in the pipeline are removed and SVM set properly:



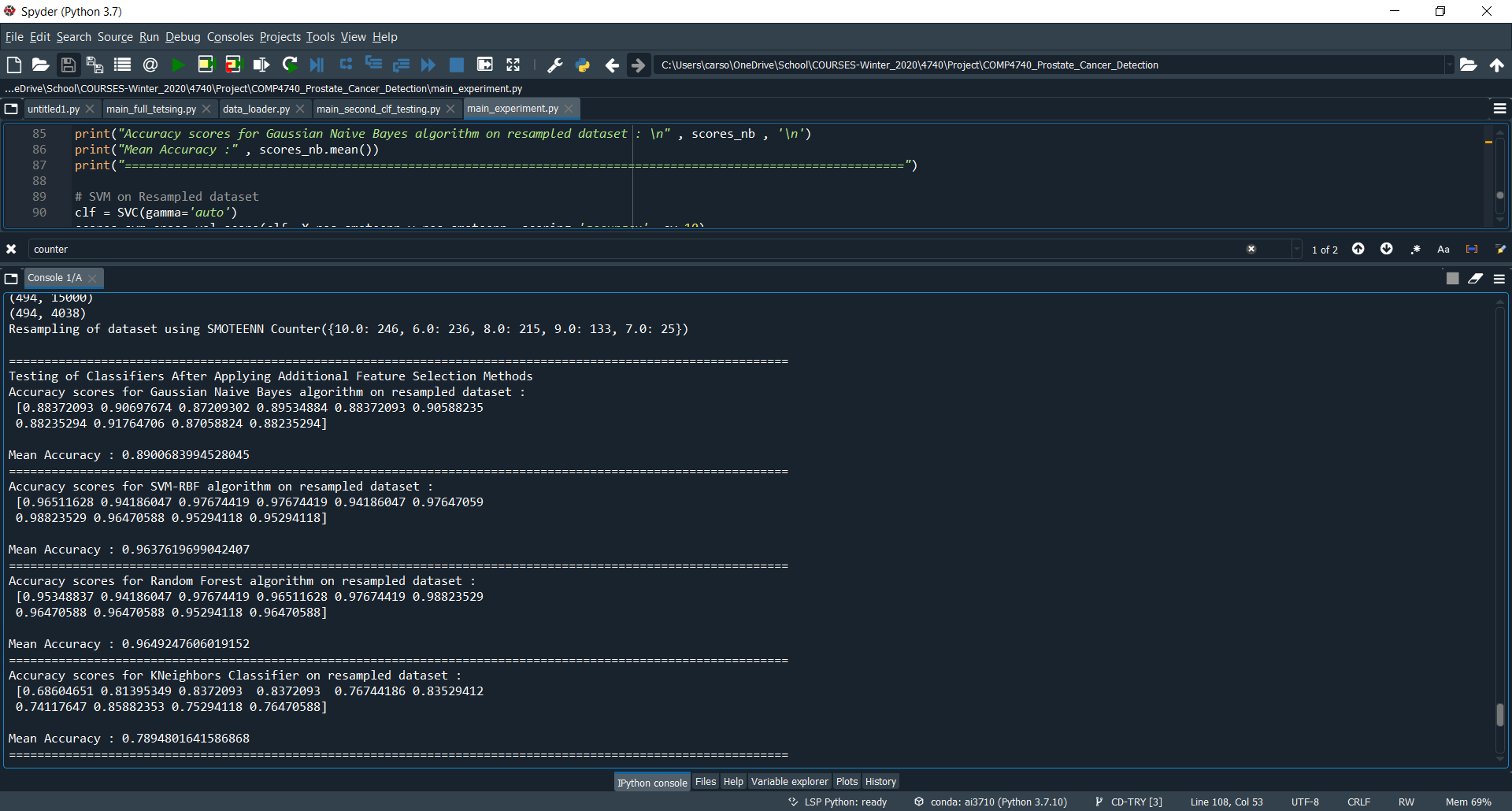
Number of samples before and after SMOTE (using grade):



Test using GLEASON\_SCORE instead of GRADE\_GROUP (and no isolation forest)



Results using everything and turning off LDA



Results with everything and turning off LDA and fixing the output for KNN

