Adithya B

**CURNEU TASK-1 FRUITS DATASET**

**PROBLEM STATEMENT:**

A dataset labelled based on fruit height, width, mass and colour score is given in fruits.xlsx. A classifier based on k Nearest Neighbour (KNN) algorithm is to be crafted for classification.

• Generate scatter plots for various combination of parameters and do the feature engineering meaning thereby which parameters of best suited to build the classifier.

• Split the data into test and training split.

• Building a classifier using KNN from scratch.

• Figure out the best value of k with highest r\_score.

• Run at least three test cases on the parameter and assess the fruit using the classifier.

• Only use python

**SOLUTION:**

The dataset has 6 attributes . They are :

🡪fruit\_label

🡪fruit\_name

🡪mass

🡪width

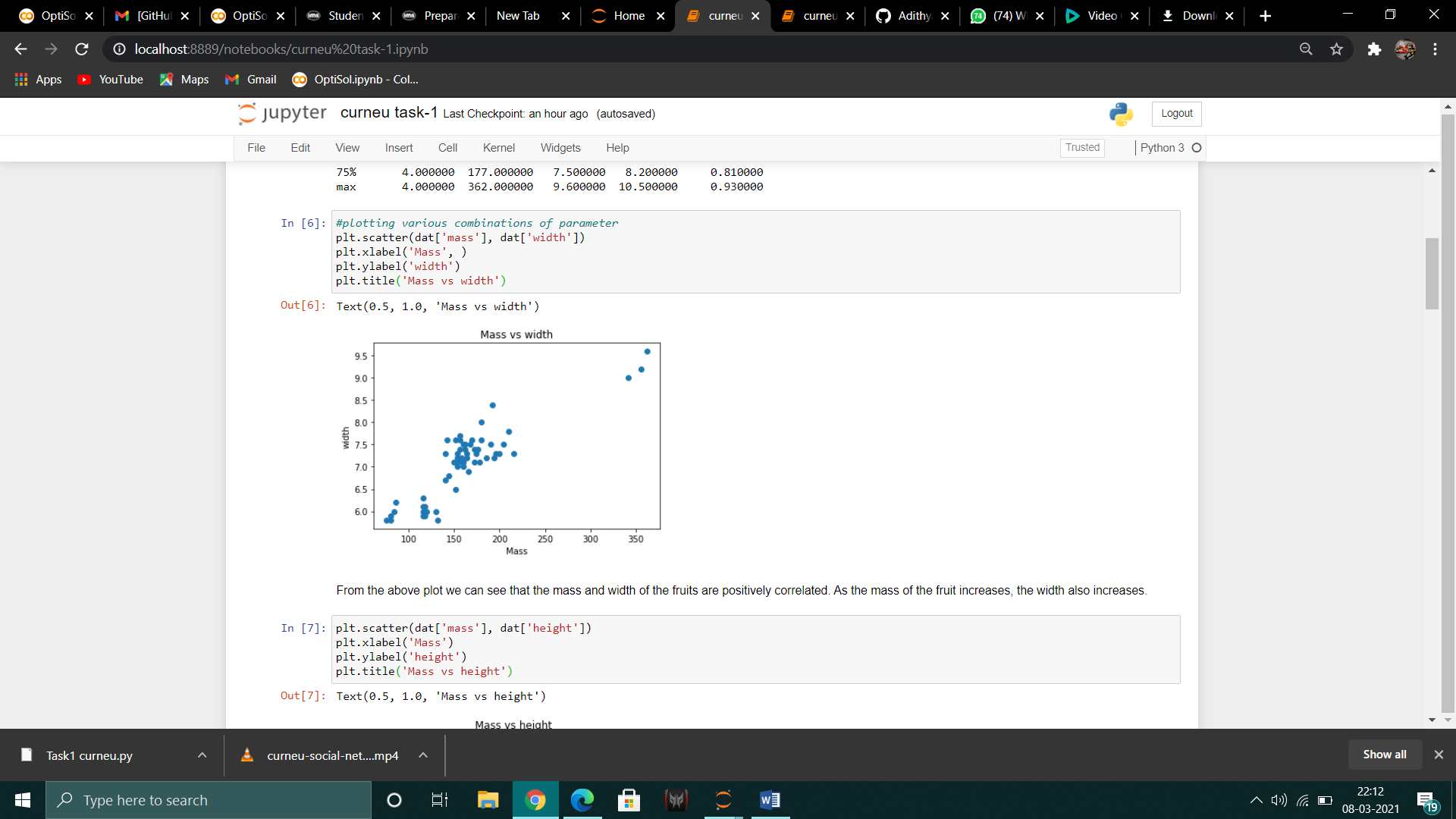
🡪height

🡪color\_score

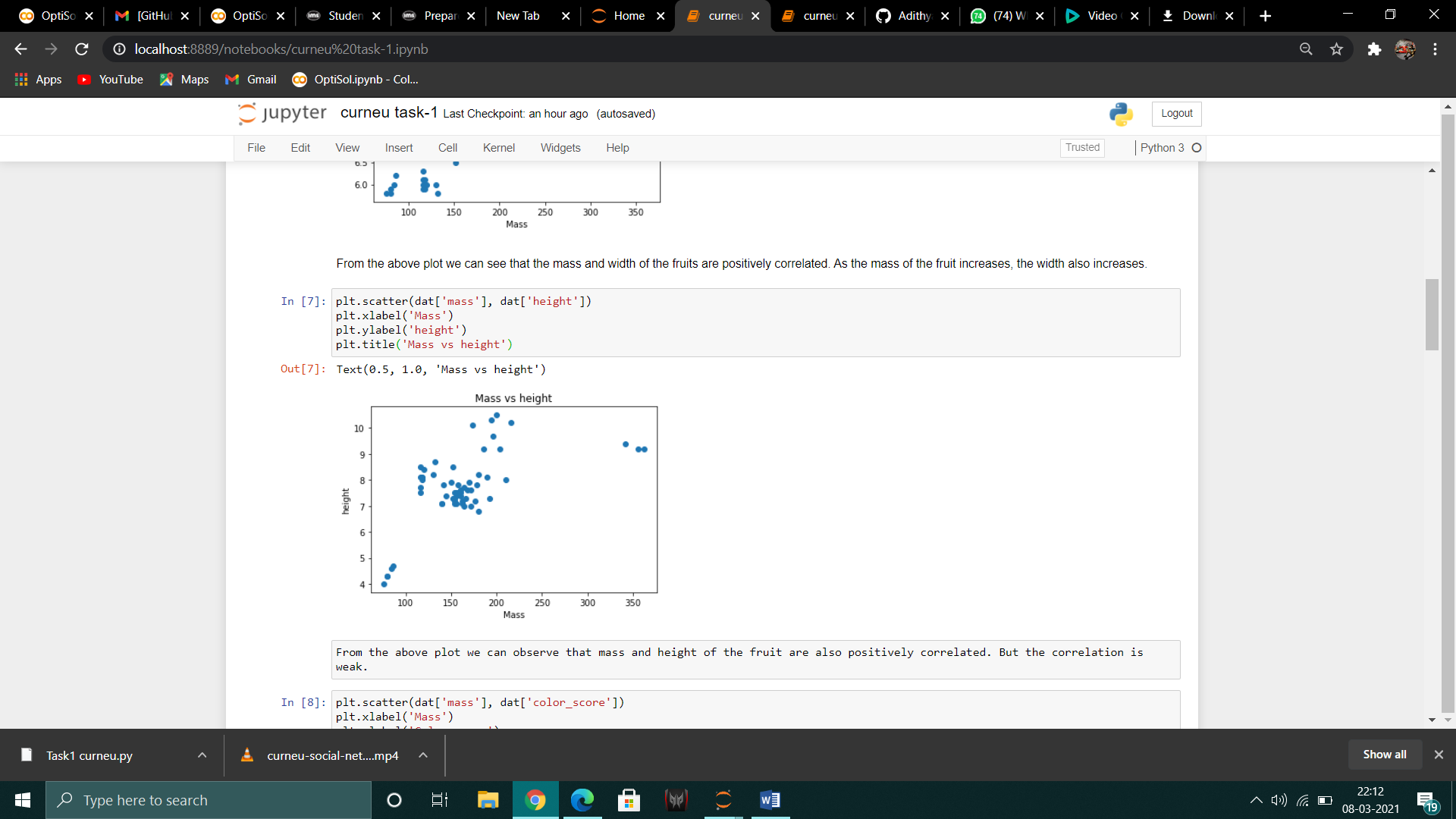
Out of the above 6 variables, fruit label is our target variable. We can remove the fruit name column since it is already encoded in the target variable column.

Going further we make scatter plots for all the variables in the dataset to check for correlation between them.

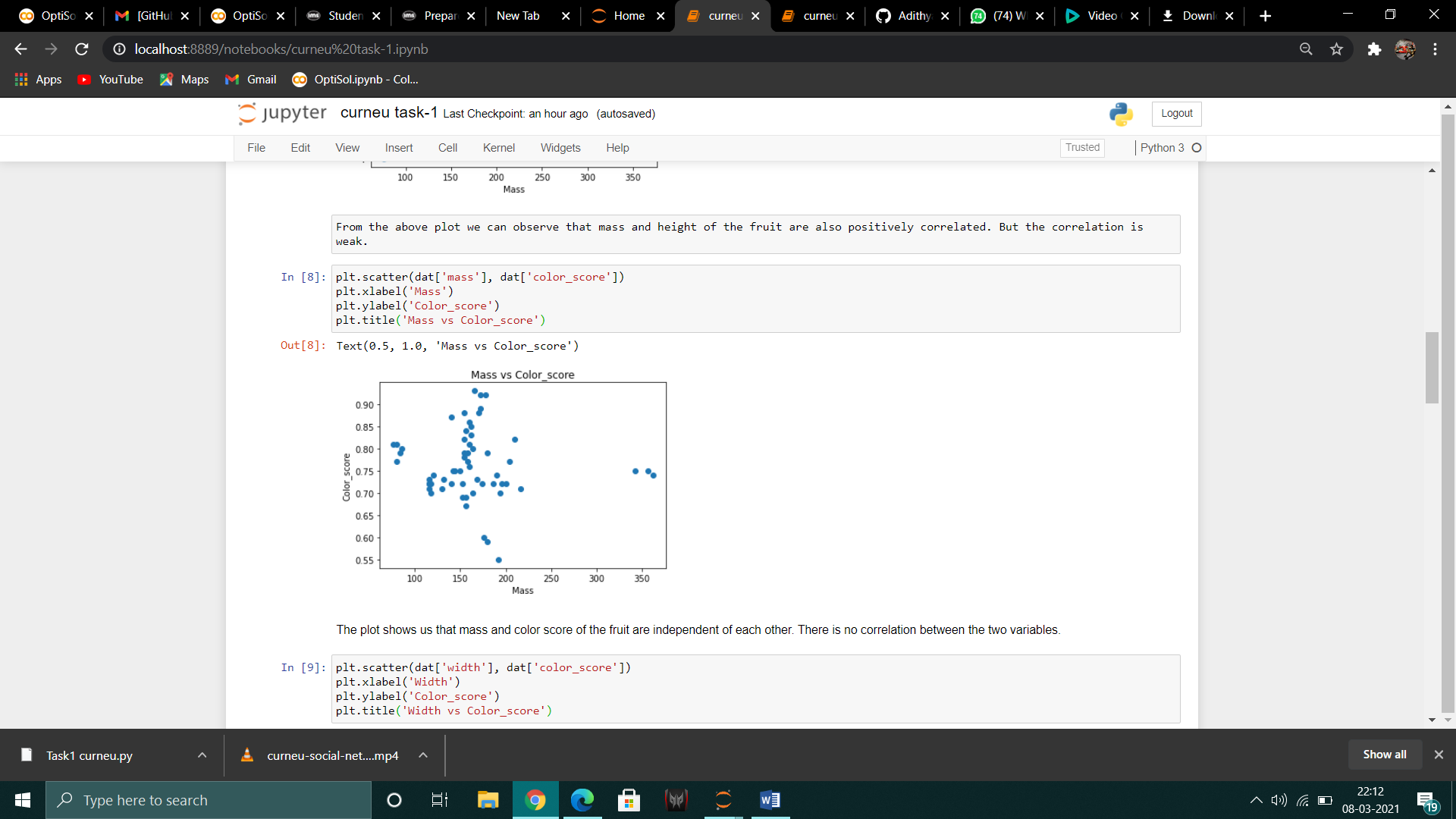
The results of the plots are given below :



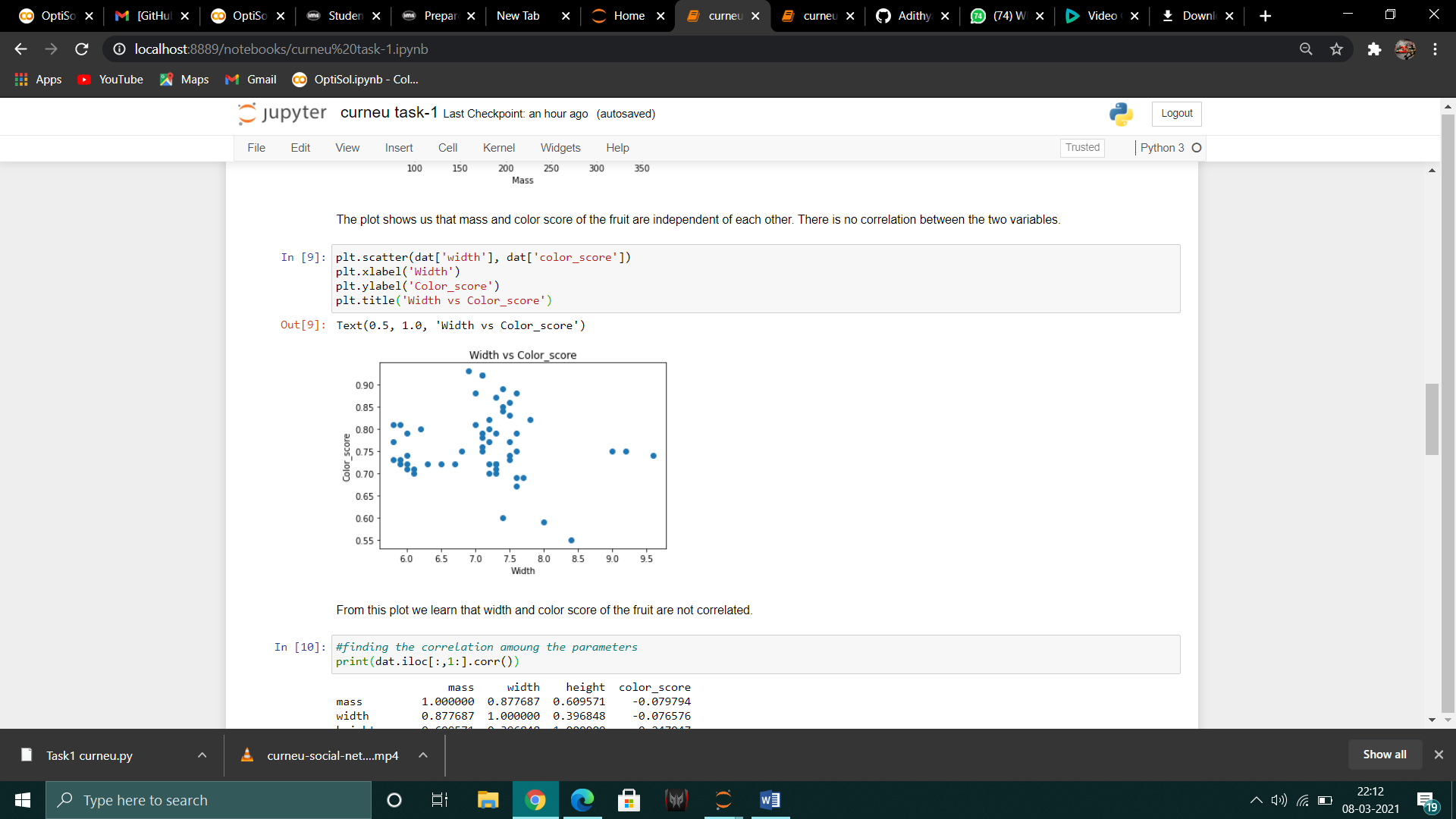
From the above plot we can see that the mass and width of the fruits are positively correlated. As the mass of the fruit increases, the width also increases.



From the above plot we can observe that mass and height of the fruit are also positively correlated. But the correlation is weak.

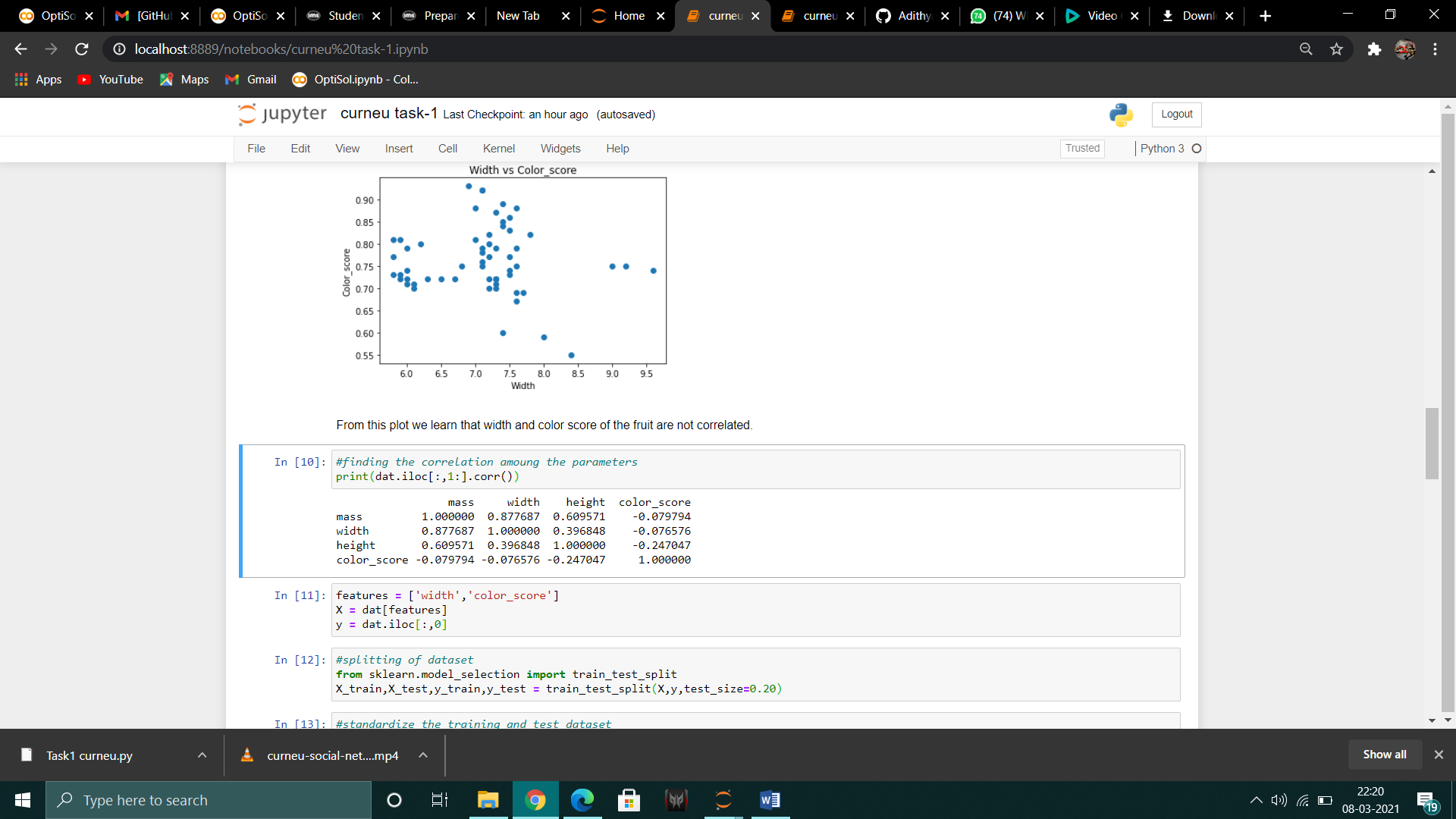


The plot shows us that mass and color score of the fruit are independent of each other. There is no correlation between the two variables.



From this plot we learn that width and color score of the fruit are not correlated.

Next we find the correlation values among the variables and select the independent variables that have less correlation between them so that it doesn’t affect the performance of the model.



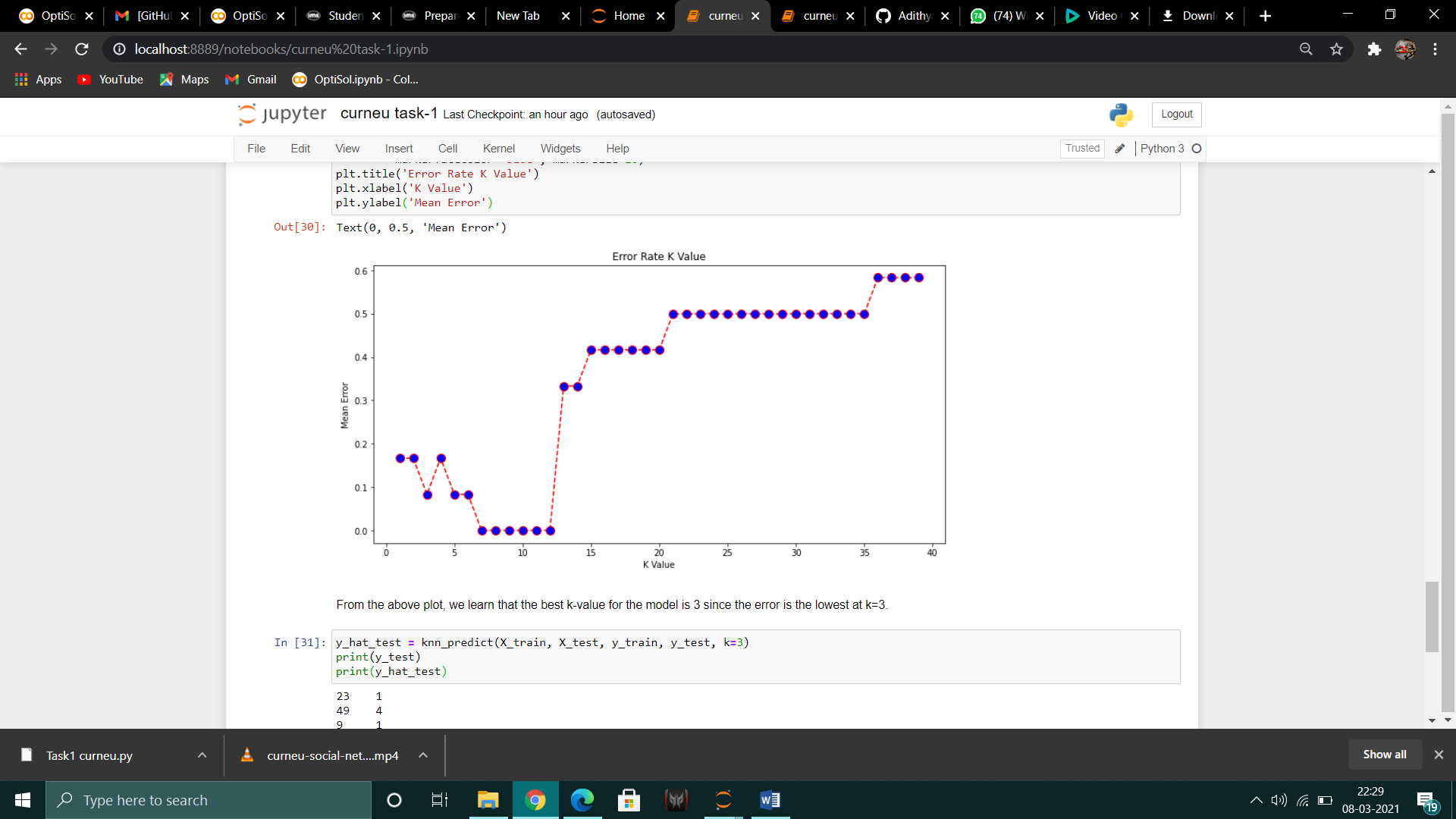
Based on the results of the correlation values, we take only two attributes as independent variables (width and color\_score).

Now the data is split into training and test set.

After splitting, the data values are scaled using the standard scalar library.

Finally the KNN model is coded from scratch and the best k-value for the model is found by plotting the k values from 1 to 40 in the x-axis and and corresponding error for each.

From this graph the k-value which gives minimum error is chosen. In our case it is found to be 3.



The model is now fit to the dataset and is trained.

Then the model was tested and the accuracy was 91% and r-squared value is .802 .