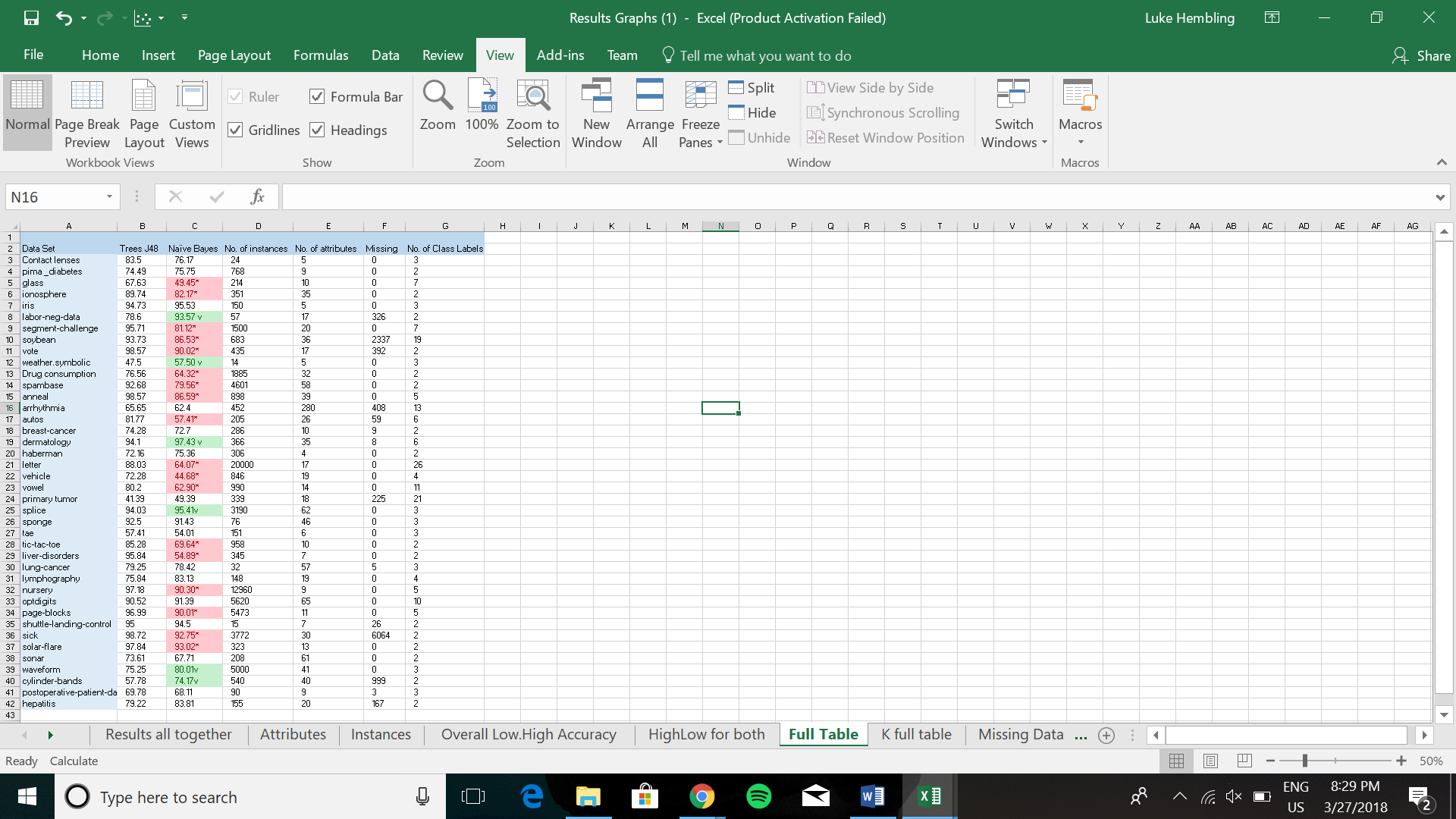
**Practical Data Analytics and Mining – Second Submission**

**777158**

# Task I

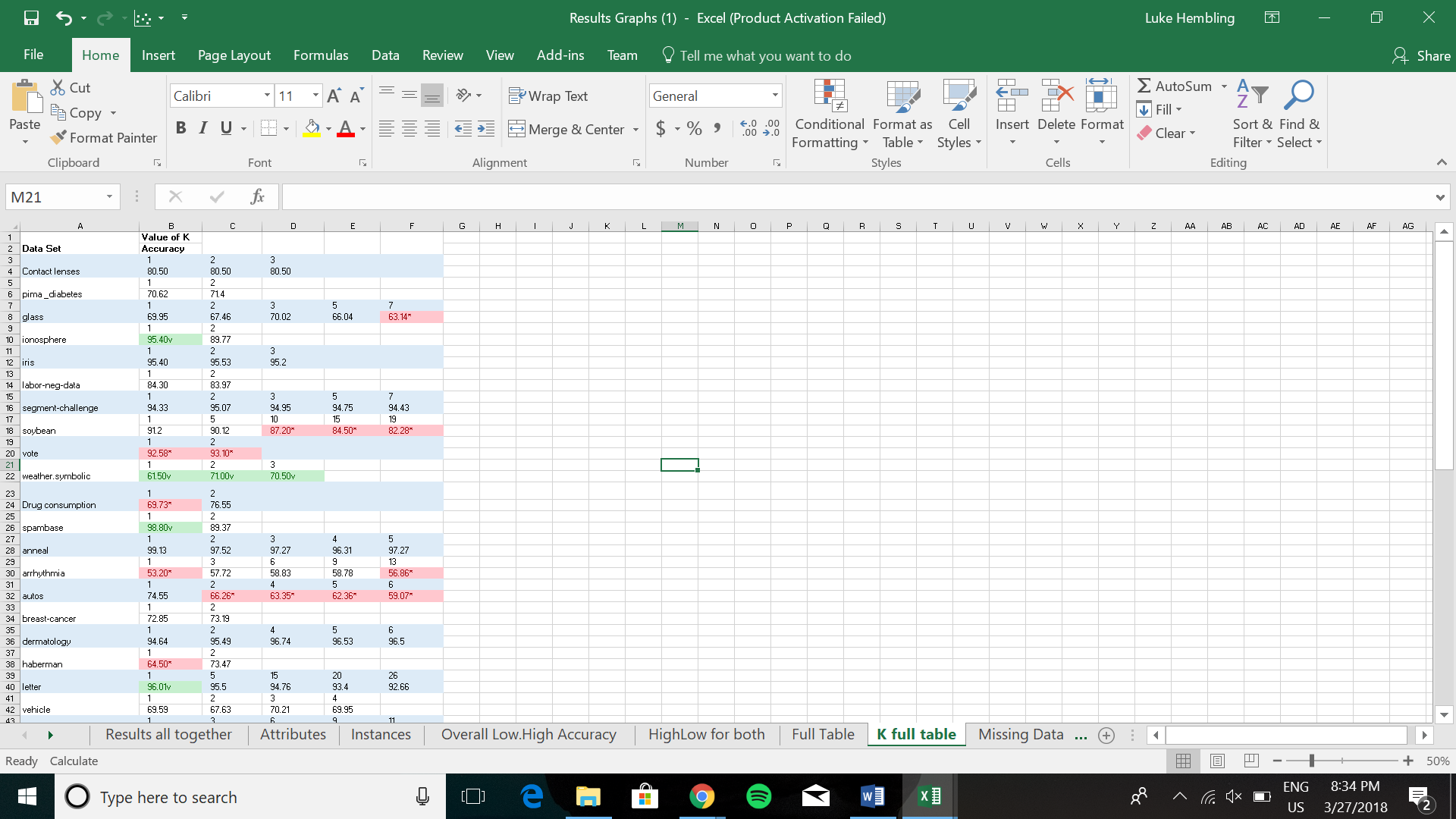
## Data Set

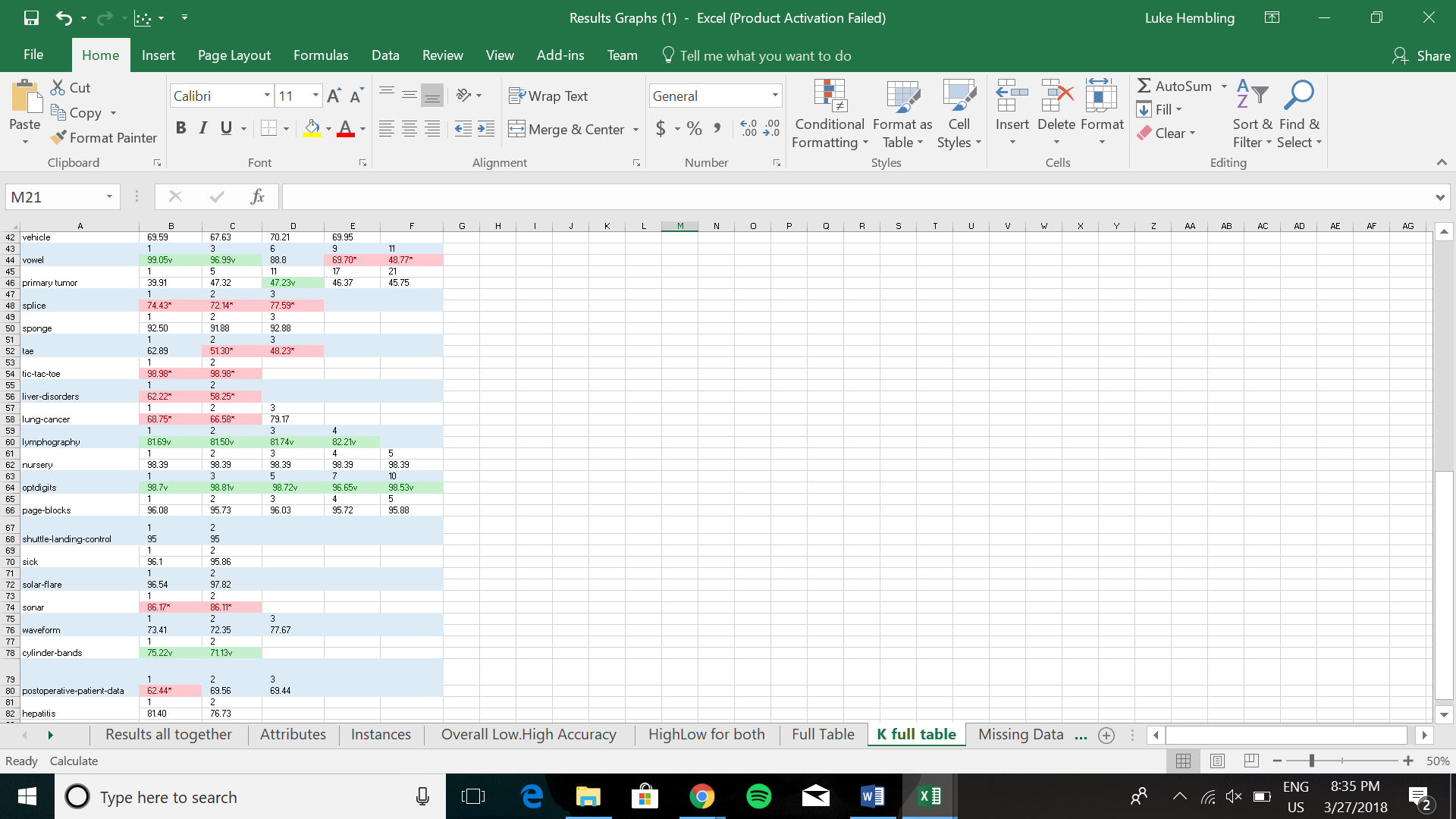


(\*) significantly worse than base classifier (J48 Trees)

(v) significantly better than base classifier (J48 Trees)

## K-NN (IBk) Accuracy





(\*) significantly worse than base classifier (J48 Trees)

(v) significantly better than base classifier (J48 Trees)

## Justification

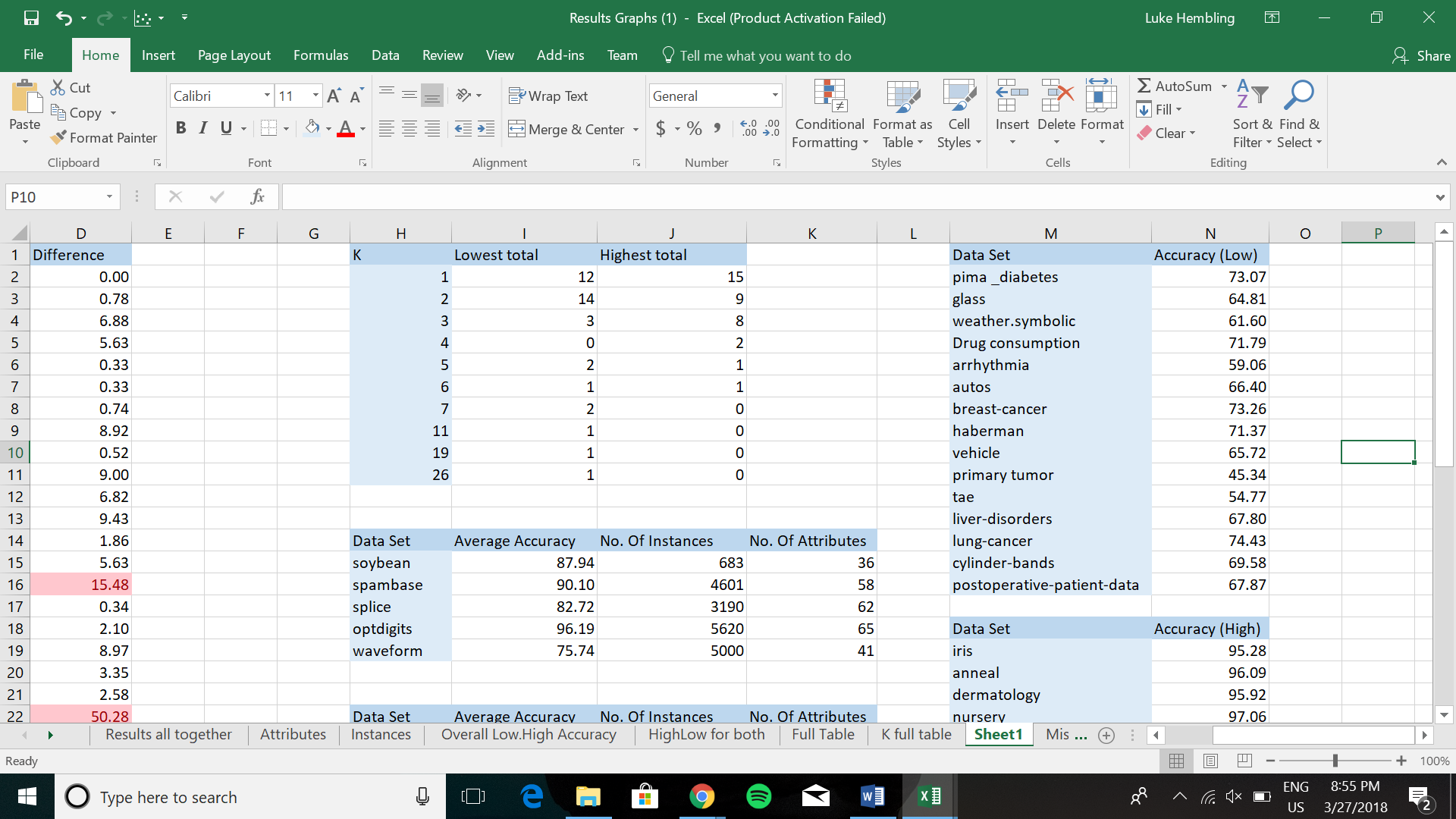
The datasets were chosen because of their wide range of properties such as the data sets with low and high instances, attributes, as well as having missing values in their data sets. This allows me to test for a wider range of accuracies to further understand the trends that may occur in each algorithm. As there are a total of 40 datasets being tested, the results can be more accurate and a conclusion may be reached if several data sets provide evidence for a trend, however no conclusion may also be reached there are no correlations. The accuracies varied between all the different data sets, and that may be represented in the following graphs that compare the trees and naïve method to each other for each data set.

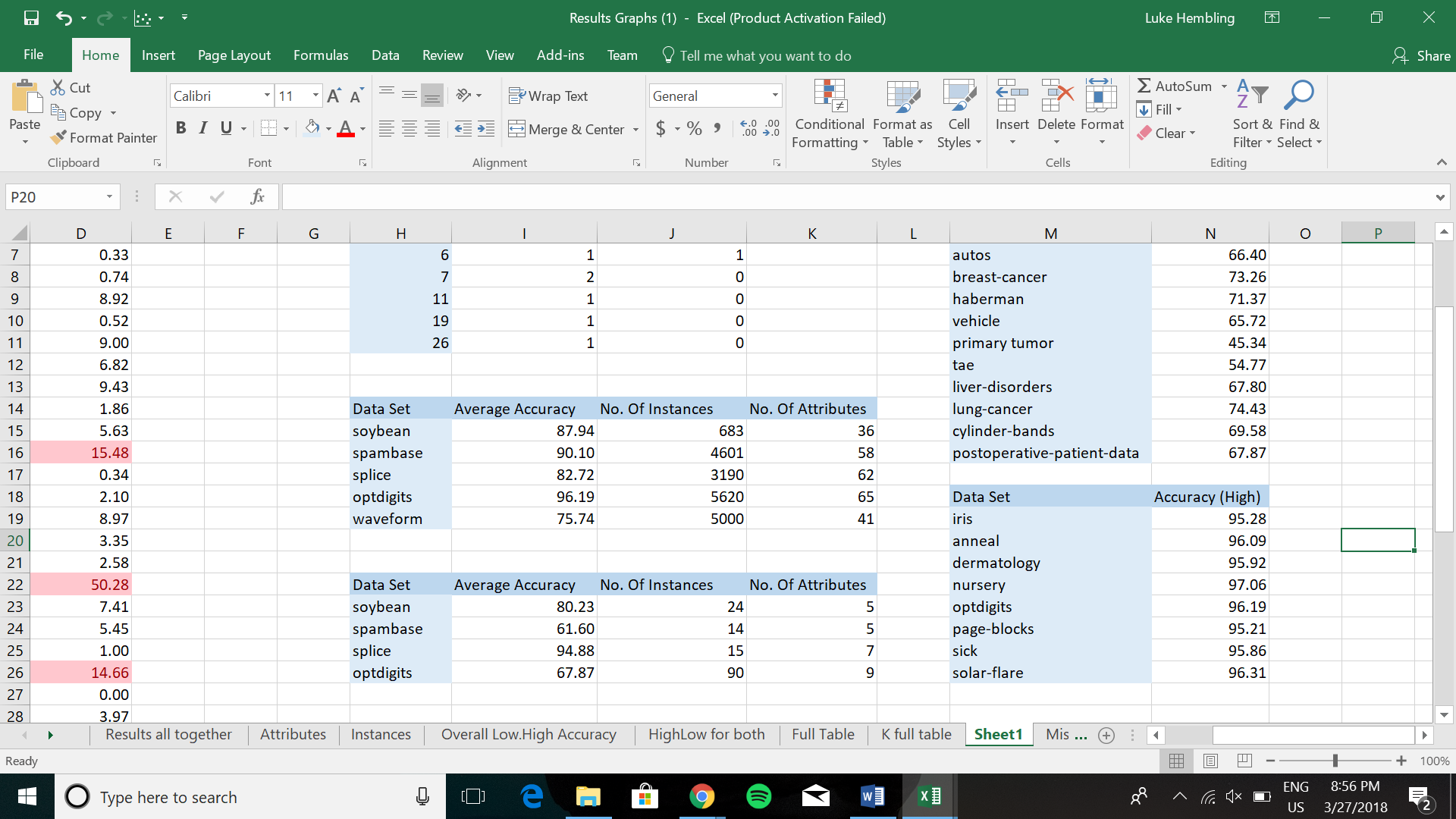
## Analysis of Results

There is a total of 40 data sets where the techniques have been applied. The techniques applied are Decision trees (J48), Naive Bayes, and K-NN (IBk). The values of K varied between 1 and the number of class labels within each set, going up in intervals to get 5 results. With this I have generated the accuracy of each technique with each individual data set. The classification process should find a model that separates the data into different classes, with the accuracy showing the percentage of correctly classified instances. Each technique uses a different classification algorithm, which classifies the data differently, and this may possibly be further affected by dataset properties such as the number of instances or attributes, and the number of missing values in the dataset. The base classifier for these sets is the decision trees algorithm, and the other algorithm accuracies are compared to this one for the significantly worse or better accuracies.

We can see that the accuracies favoured J48 slightly, as it sometimes showed that trees was more accurate however we can see in some cases the same for naïve bayes, but J48 seems to be more accurate in a majority of the data sets.

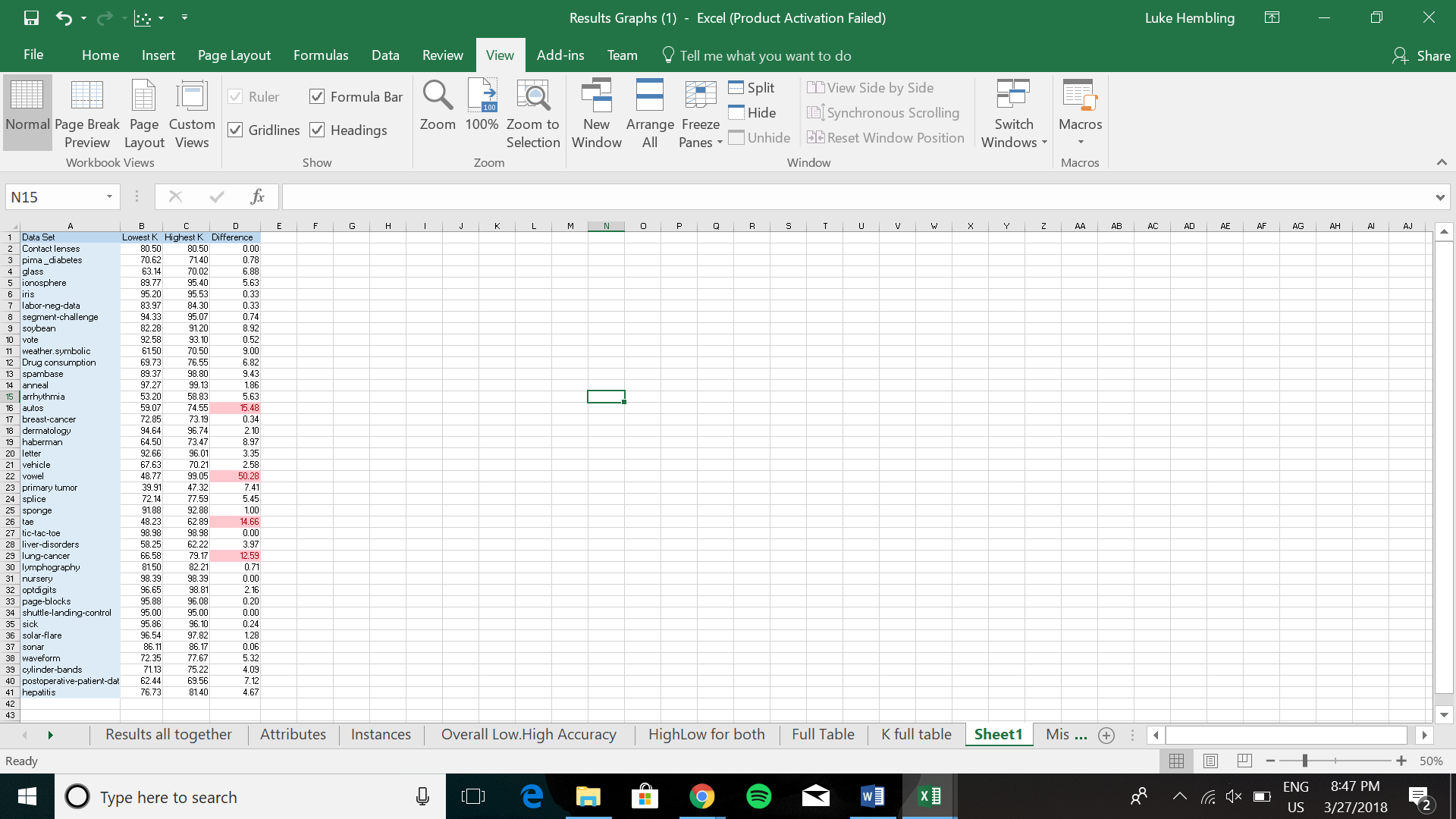
From the results, we can see the overall accuracies of each dataset, with the significant accuracies being indicated by a \* or v. From this, I have calculated the average accuracies for each data sets, and these datasets have the lowest and highest accuracies which will be useful to see the effect of dataset properties on each dataset.





## K-Values

For the K-NN accuracies, different results were produced because of the value of K being changed, the highest and lowest accuracies are shown for each data set below:



The accuracy stayed the same for four data sets (contact lenses, tic-tac-toe, nursery, shuttle-landing-control), where of the k values tested, there was no change in accuracy. Nursery had five different k values tested, whilst the other three had only two or three due to there being only two or three class labels. Also from these results, we see that vowel had the biggest different in accuracy, with the highest accuracy being for K=1, and lowest being K=11.

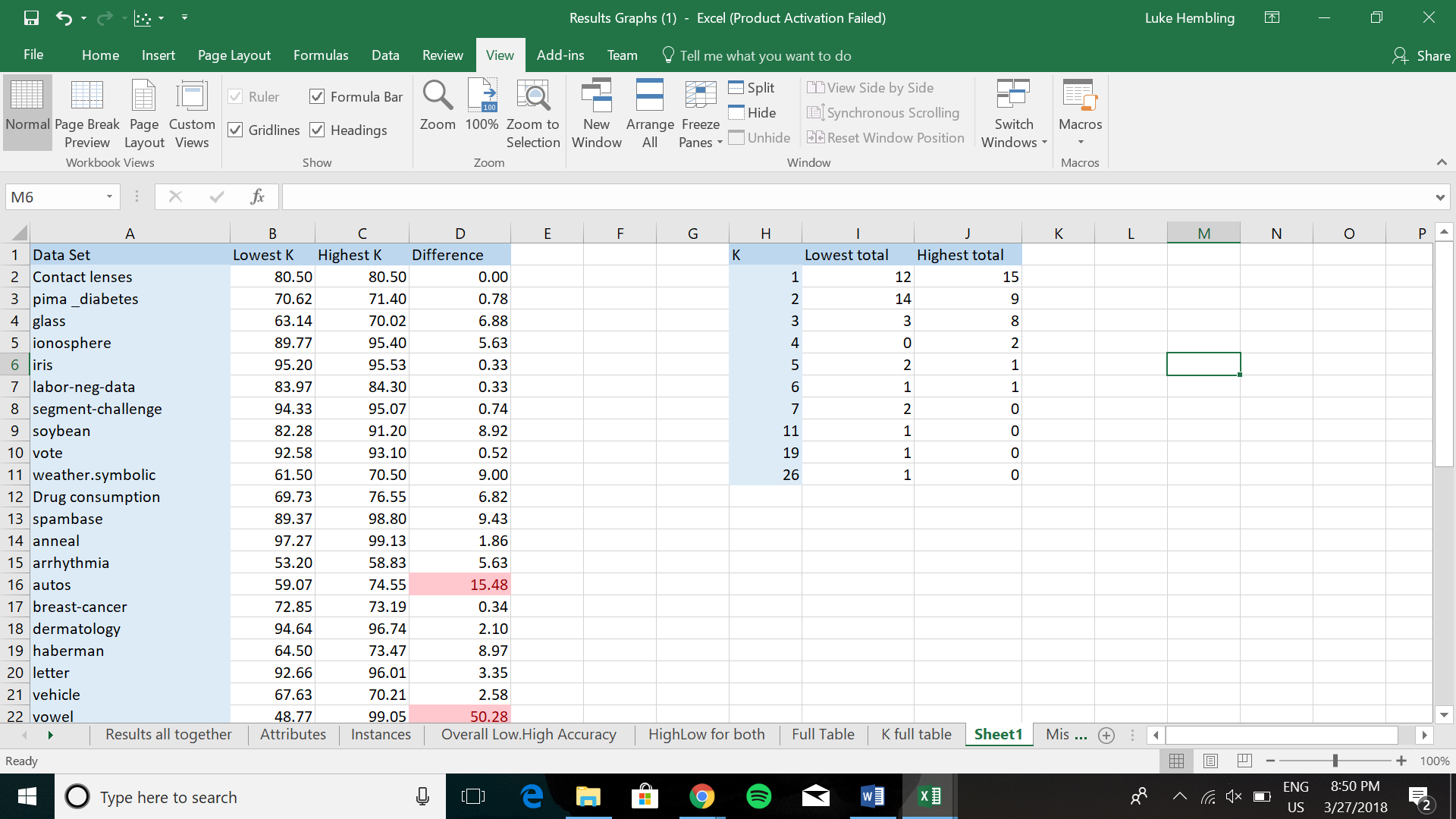
For these three data sets with a substantial difference in K accuracy, the following line graphs show the trends with a difference in K value:

The general trend is that accuracy went down as the K value went up. However, there are also datasets as follows:

In these graphs, the lowest value of K wasn’t the most accurate K, however it wasn’t a continuous rise as K went up as at 5, 2 and 2 the peak was reached. Looking at a table with a large difference in K value:

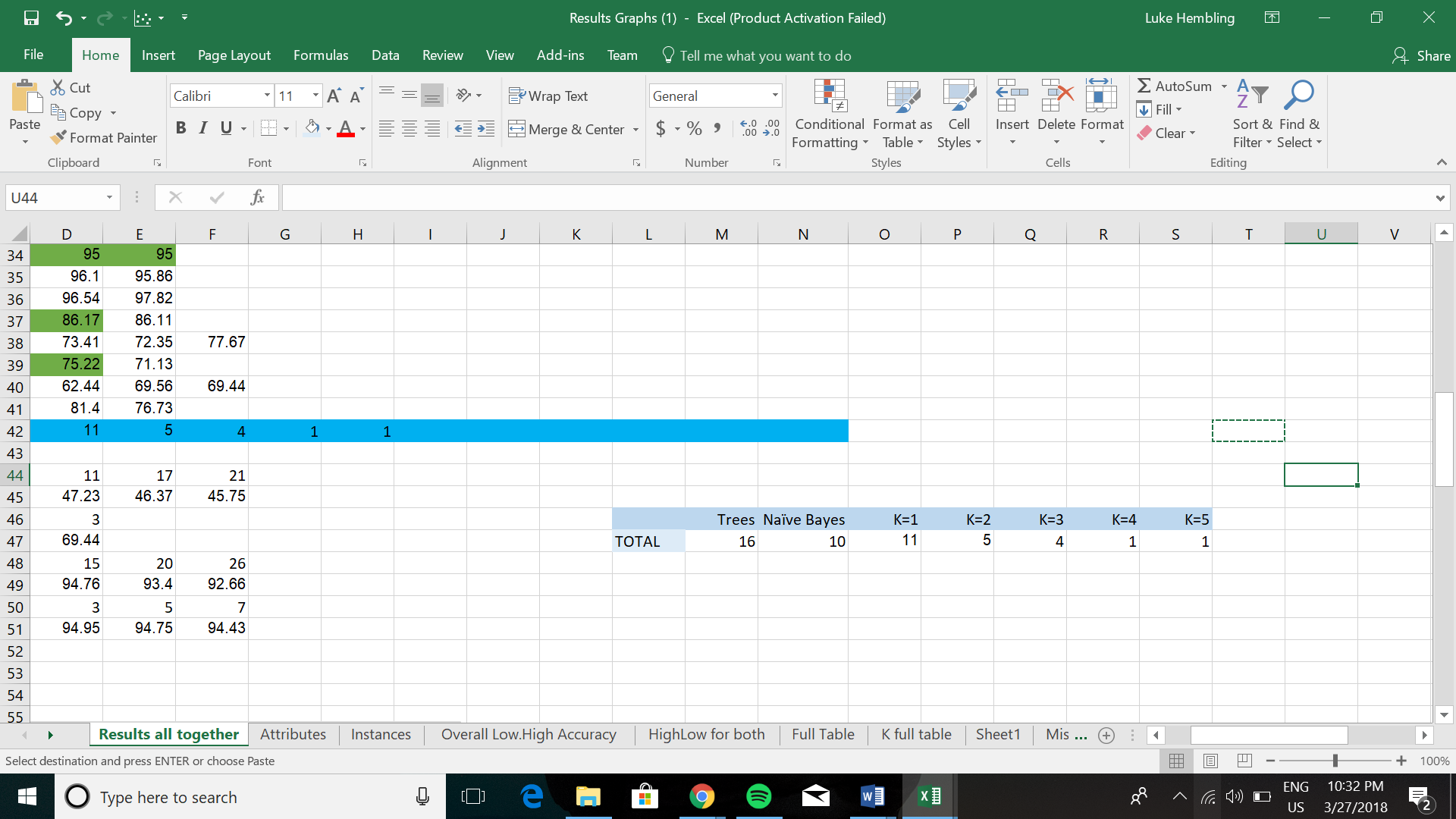
We see a similar trend to as seen previously, where the lowest K was the most accurate.

This table shows for which K value had the best or worst results for all the datasets, the best and worst being the lowest and highest accuracies. Any K Values not included didn’t have any of either.



So from this table we can see that a K value of 1 to 3 was more common to have the best or worst result, however a K value of 3 has a significantly more highest accuracy than lowest.

Overall, with all the classification algorithms involved, the following table shows which of them had the best accuracies:



From these results, we can see that Trees consistently has the highest number of best accuracies of all these methods.

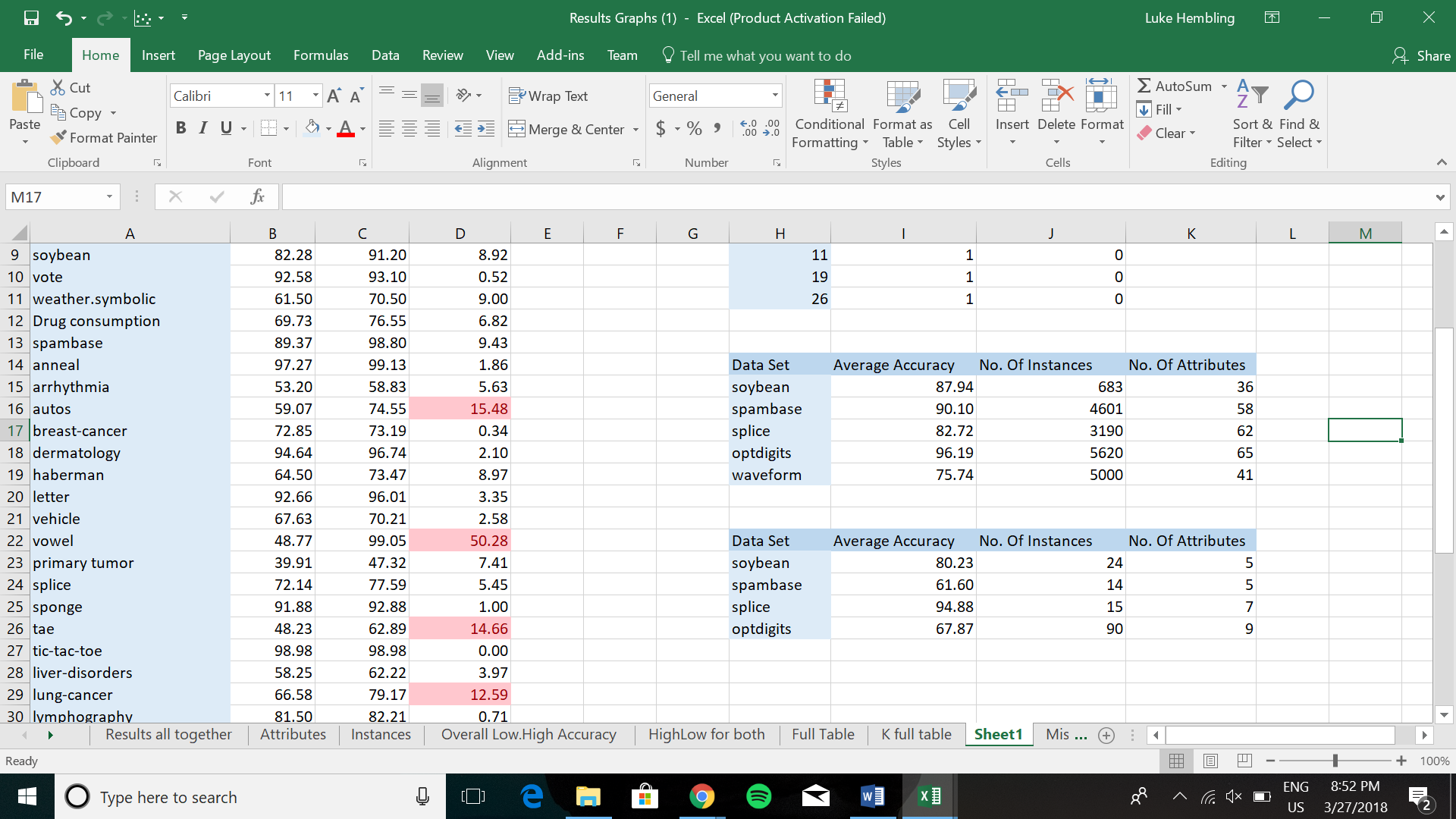
## Attributes

Tables may also be created from generating an average accuracy per dataset seeing if there is a correlation between the dataset property and the accuracy. These graphs show that relationship between accuracy and attributes, in intervals dividing the attributes into three categories of low to high number of attributes.

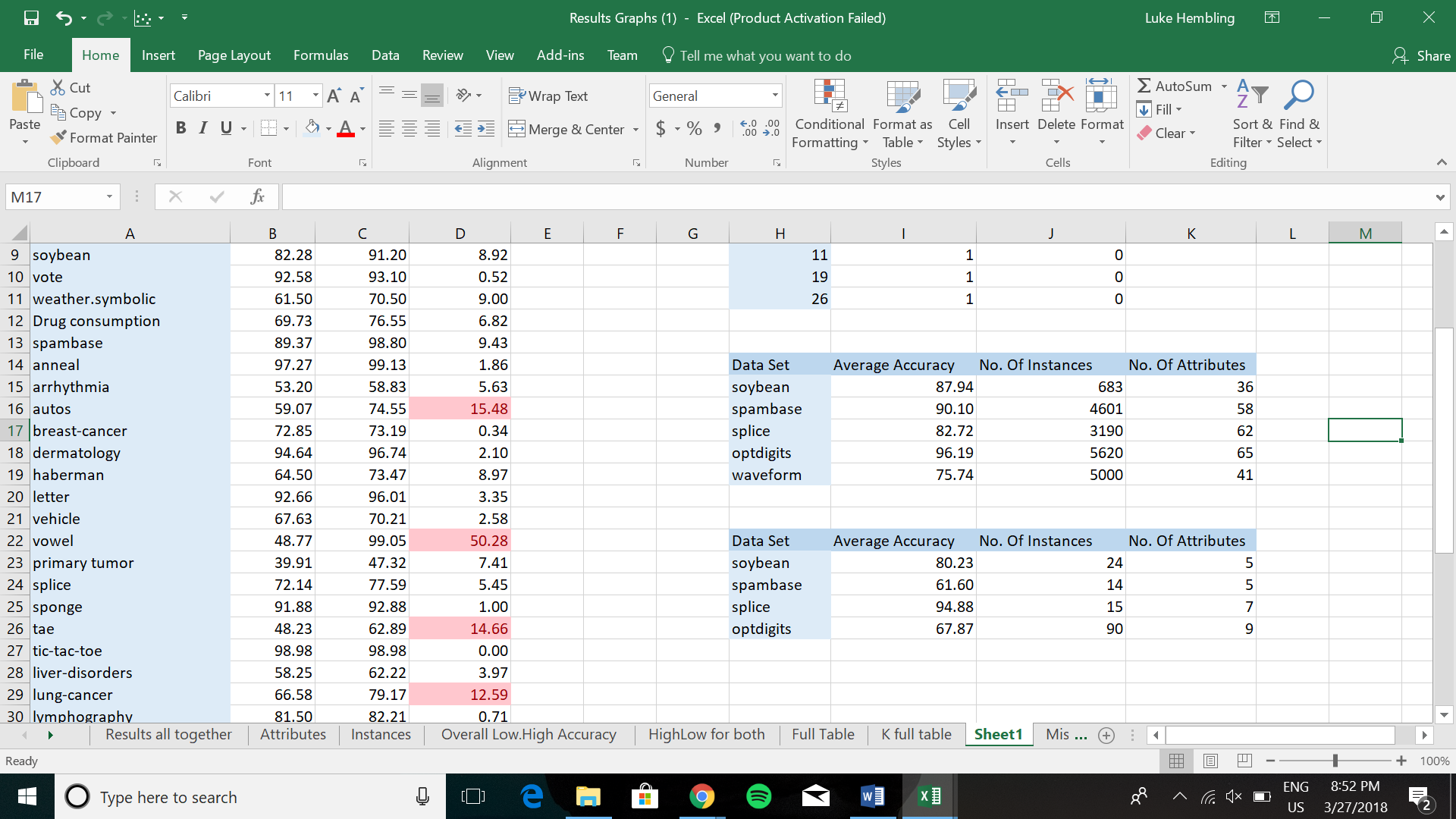
From these graphs, I didn’t really see any correlation between the number of attributes and accuracy as the results stayed varied throughout all the tables. With the other dataset property, instances, we can also look at the number of instances per dataset in relation to the average accuracy of each dataset.

## Instances

Like the previous graphs about attributes, there was no real correlation between the number of instances and the accuracy as well. We may also look at the average accuracies for datasets that have a high number of instances and attributes. This is a graph for data sets that have a high instance and high attributes:

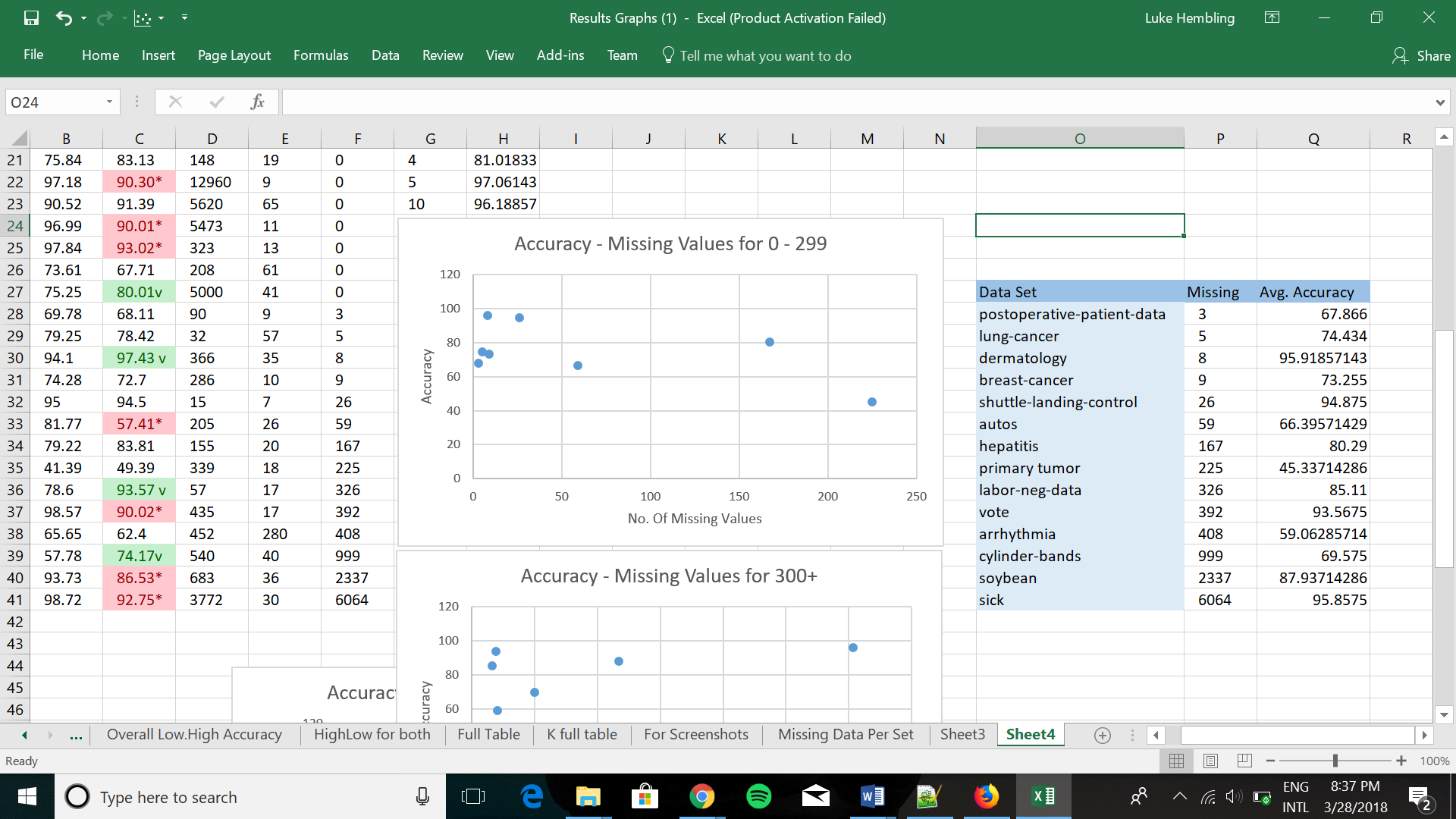


Also, all the ones with low instance/attribute values:



Overall from these tables, we can see that with high instance and attribute numbers, there were high accuracies in comparison to low. Two of the accuracies are below 70% for low values, whilst two were above 90% for high values.

## Missing Values



There were 14 data sets with missing values, ranging from 3 to 6064. The following graph displays if there is a trend between accuracy and the number of missing values (increased number of missing values with every data set going to the right).

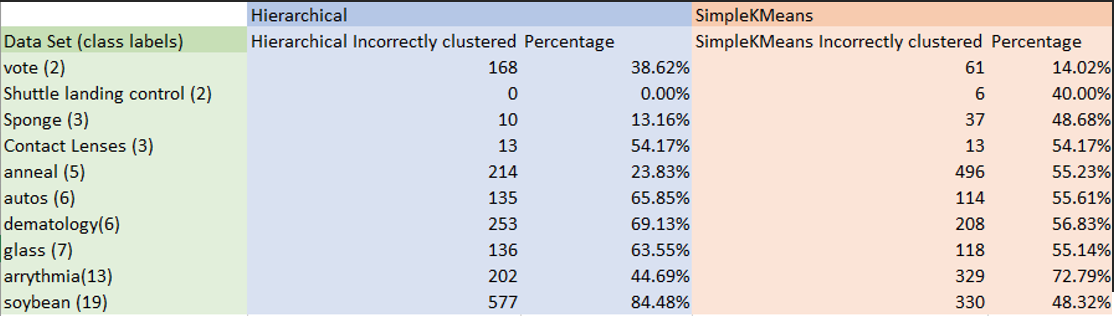
## Conclusion

To conclude, the most consistent classification algorithm I found was J48, as seen from the accuracy values, it had the highest number of best accuracies. I found that the most common trend for the K Value is for the accuracy to go down as the K value goes up, however there were cases where the lowest one (K=1) wasn’t that most accurate one as accuracy went up, and went down after a certain K value such as 3 for example. The missing values didn’t have any significant effect on the accuracy, however the importance of the missing values is still valid, as a dataset with relation to medical issues such as breast cancer or hepatitis would need to be quite accurate.

Regarding the specific data sets where I found something to be irregular, the vowel data set differentiated by 50 in accuracy between the classification algorithms, even with a low number of instances and attributes, as well as no missing values. Furthermore, the data set nursery stayed consistent through all the K values, keeping the same accuracy.

# Task II

## Results



The techniques applied to these datasets are Hierarchical and SimpleKMeans, two methods of clustering that group similar objects. For these data sets, I have chosen to the data sets based on the number class labels to cover a wider range of results, which may indicate relationships or trends present in the data sets with the two clustering techniques. I have taken the number of incorrectly clustered instances and the percentages from both methods after simulations on WEKA.

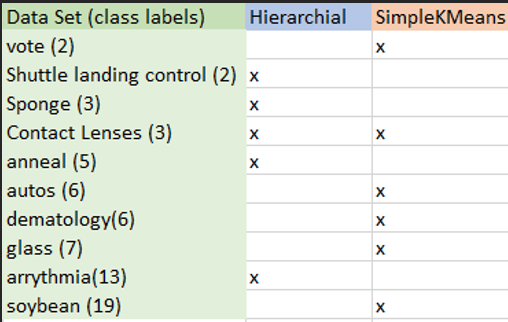
The following graph is a visual representation of the data above:

## Analysis

The performance of these methods are quite similar overall, there doesn’t appear to be any obvious choice as SimpleKMeans only performed better on only one data set in comparison to Hierarchical. When looking at the relationship between the incorrectly clustered instances and the number of class labels, we can see from this figure that it is more likely to get incorrectly clustered instances with a higher number of class labels, which is evident by the low number of incorrect clustering instances through to 5. We can further see this in the following graphs:

All these graphs follow a trend of more class labels meaning more incorrect instances (higher percentage).

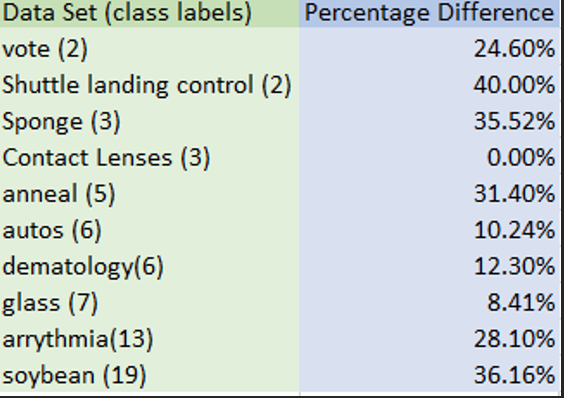
Furthermore, in terms of percentages which can be seen here:



\*x represents a lower percentage which indicates that it is preferred.

I found that the lower number of class labels did better with hierarchical, and SimpleKMeans did better with a higher number of class labels.

The percentage difference between the two methods varied, sometimes have little to no difference with around 0-8%, whereas other datasets were clustered very differently between the two methods, differing by a percentage of 40%. The following table shows each data sets percentage difference between the two techniques:

From the data, we can look at the data set shuttle landing control, the hierarchical technique correctly clustered all the instances, in comparison to the SimpleKMeans technique where there were 6.

Also, the contact lenses class, where the incorrectly clustered instances and percentages were the same between the two techniques, values of 13 and 54.175%.

The soybean data set with the hierarchical technique had a percentage of 84.48% of incorrectly clustered instances, which is a large portion of the overall data. Compared to the SimpleKMeans, which had 48.32% that were incorrect, which is a large difference in values.

We can also compare the data sets with the same number of class labels, which were 2, 3 and 6.

The shuttle landing control class doesn’t have many incorrect in comparison to the vote class, so we can see that even though these two classes have the same number of class labels, the incorrect clusters varies.

The following two charts furthermore shows this:

## Conclusion

The choice of the technique is decided by its ability to correctly cluster the instances, so from the data that is taken, the lower number is the better choice of the two. In conclusion, I found that the data sets were affected differently by the two methods, and that the higher number of class labels meant more incorrectly clustered instances. Furthermore, I found that the Hierarchical technique correctly classified the instances better for a lower number of class labels, between 2-5, and 6-19 was better classified by SimpleKMeans.