**Did the Author Walk the Dogs Today?**

**Description**

Daily pedometer data from Steave

**Format**

A data frame with 223 observations on the following 7 variables.

StepCount

Number of steps taken in the day

Kcal

Calories burned (according to pedometer)

Miles

Miles walked

Weather

cold, rain, or shine

Day

Day of week (F=Friday, M=Monday, R=Thursday, S=Saturday, T=Tuesday, U=Sunday, W=Wednesday)

Walk

Were the dogs walked? (1=yes or 0=no)

Steps

Steps in units of 1,000 (so StepCount/1000)

**Details**

One of the authors recorded daily pedometer data, the weather, and whether or not he walked the dogs.

**Referencesa**

Cameron, A.C. and Trivedi, P.K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, **1**, 29–53.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

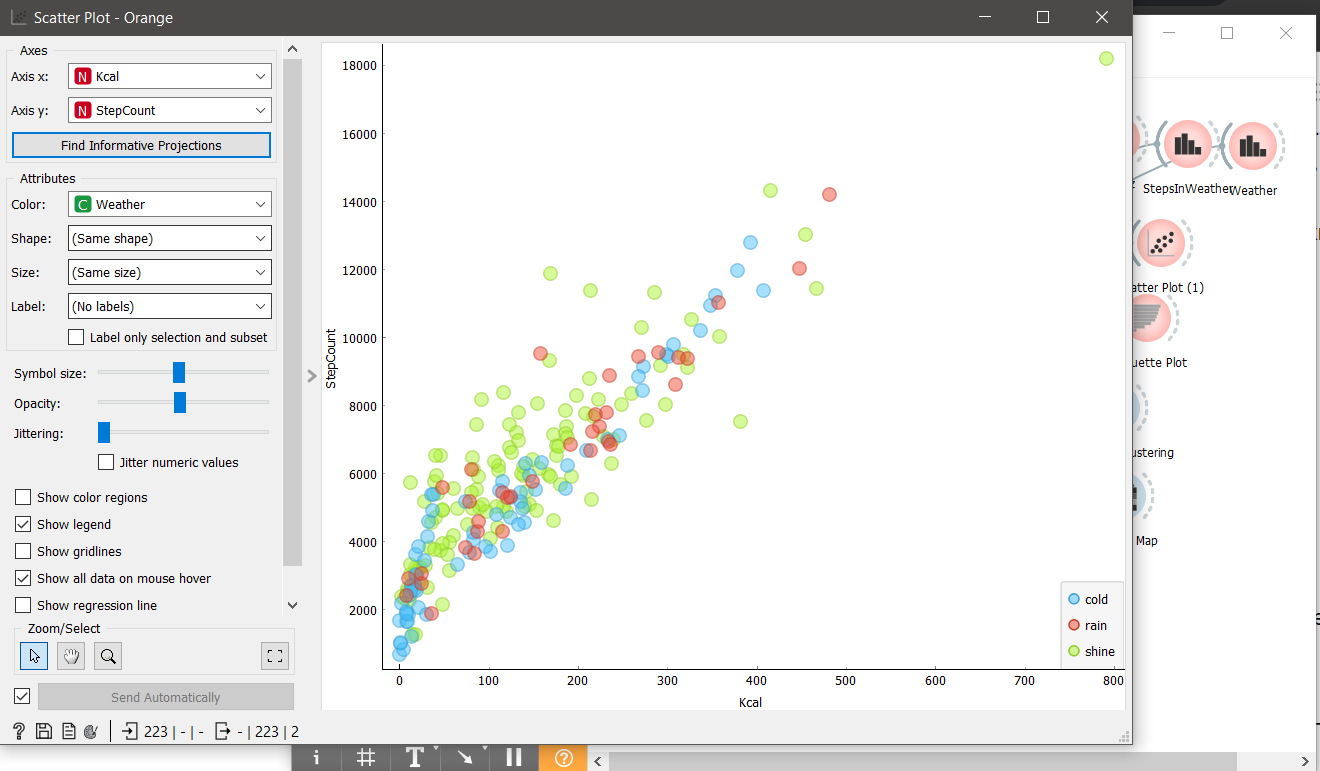
Mullahy, J. (1997). Heterogeneity, Excess Zeros, and the Structure of Count Data Models. , **12**

**Source**

[**https://vincentarelbundock.github.io/Rdatasets/articles/data.html**](https://vincentarelbundock.github.io/Rdatasets/articles/data.html)

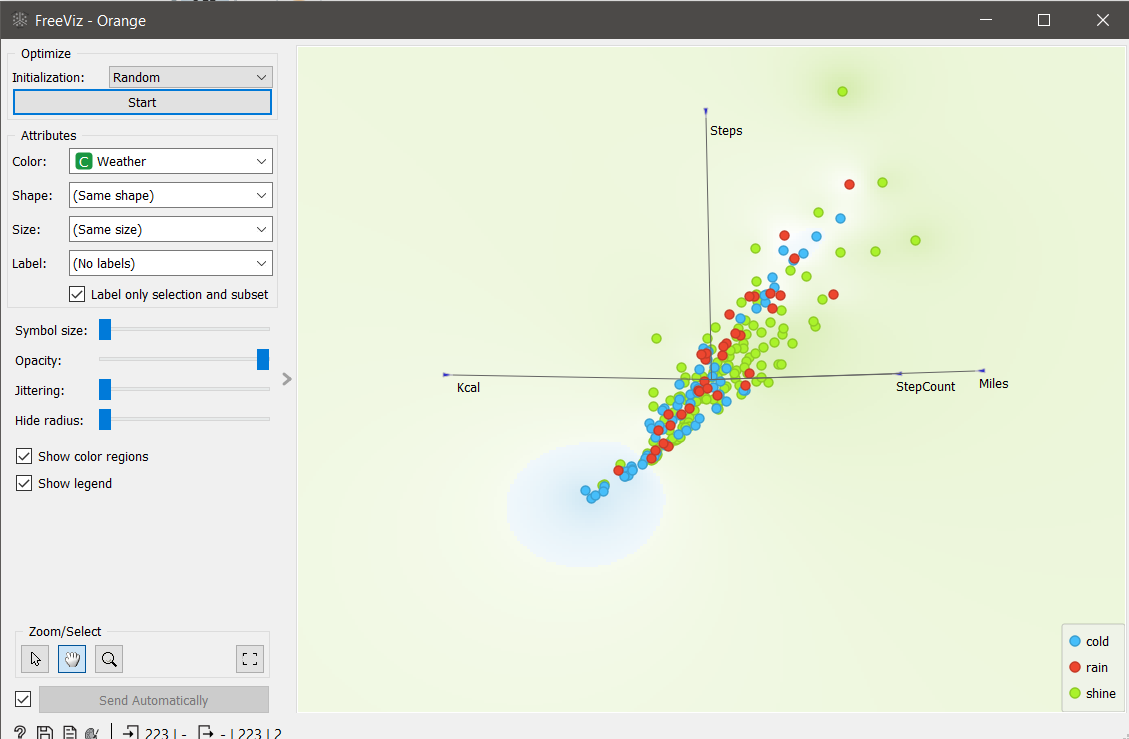
**I)**

**a)**you must create at least two 2- or 3-dimensional scatter plots illustrating the separability of classes in your dataset based on different features; the student should avoid using the data object ID as a variable in the scatterplot;



Here I showed 2 features that show how many steps were taken and how many calories were burned in these steps, and their colors show what weather the walk was in.

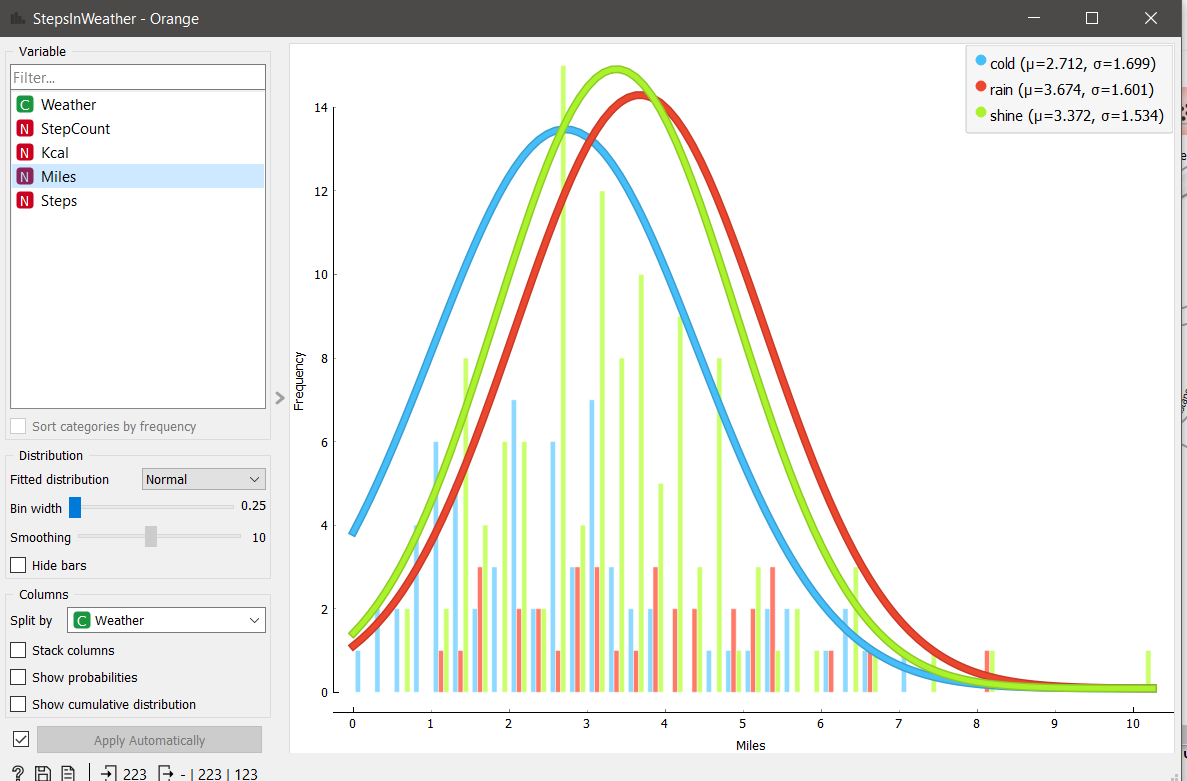
Thus, you can understand in what weather a person walked more.



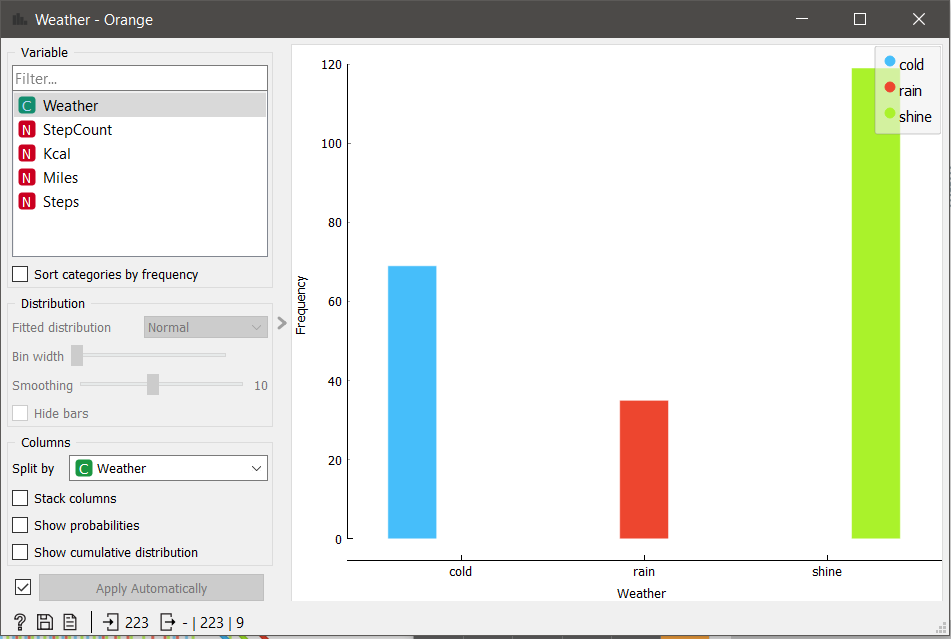
In this scatter plot, I used all the classes and by putting the illustration in this form, we can draw a line to see that the lower ones themselves are days in which there are few steps, miles traveled, calories burned, etc. And the top one which shows that it prevails over everyone in all classes

**b)** you must create at least 2 histograms showing the separation of classes for the features of interest;

**c)** you must show 2 distributions for the features of interest;

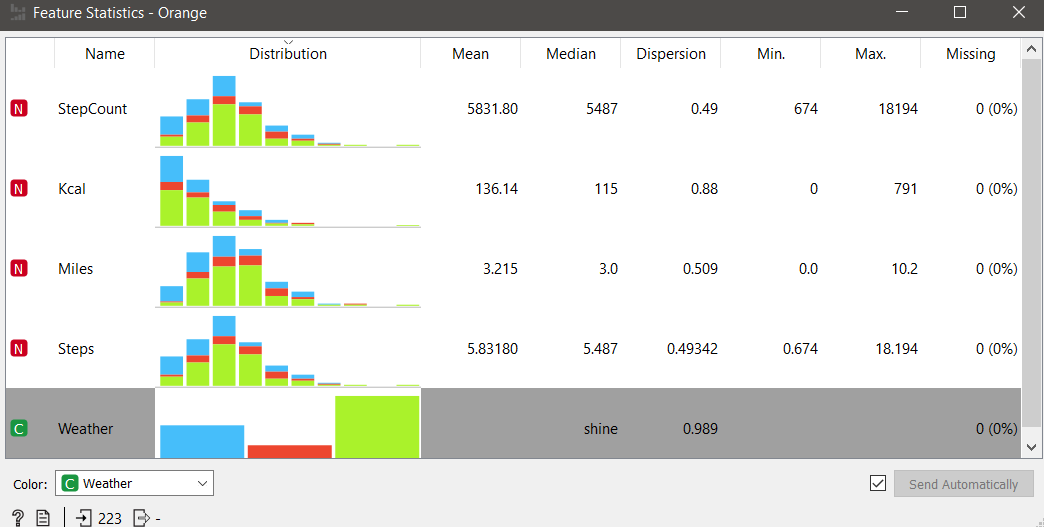


In this histogram, you can understand in what weather he often walked long distances. Here we see that during good weather he walked a lot of miles on average 3-5 miles, he walked the least in cold weather, where the average shows between 2-3 miles



Here we can also see the number of walks in different weather conditions and see the difference in walks.

**d)** you must calculate statistics on your data (at least the central tendency and the dispersion of the feature values)



In this statistic, the central tendency and the dispersion of the feature values were calculated.

**Whether classes in your dataset are balanced, or is one class (several classes) prevailing?**

As I described above, I can say that my classes are different and not balanced. Due to this, using the freeviz tool, it showed how they are spread out and how they are scattered, basically the weather class (shine) in working with the step class takes precedence over other classes and the values of these classes, a minimum also appears here. Such as a weather class like cold and steps . Thereby dividing and making poverty.

**Does the visual representation of the data allow the structure of the data to be seen? It is a question of whether data objects belonging to different classes are clearly separable.**

It shows the structure of the data, and in point analysis, you can understand and examine each element, i.e. the structure of the class and the structure of the element of the class, and how a category like weather affects these classes. All this is thoroughly shown in a point plot and in a multidimensional scatter plot.

**How many data groupings can be identified by studying the visual representation of the data? It is a question of whether there are any separable groupings of data if the data objects of different classes merge.**

I can’t say exactly how much it can be, but one thing is for sure, that a group can at least consist of 3 categories or groups, and the more groups, the deeper the analysis becomes and you can even find similarities in different categories that can fall into one group, What will I show next...

**Are the identified data groupings close to each other or far from each other?**

It depends on the number of groups or categories. As I said above, the more groups, the more accurate the analysis, and each time the intersection of different categories becomes smaller and smaller, and vice versa, when there are fewer and fewer categories or groups, the number of intersections of different categories increases.

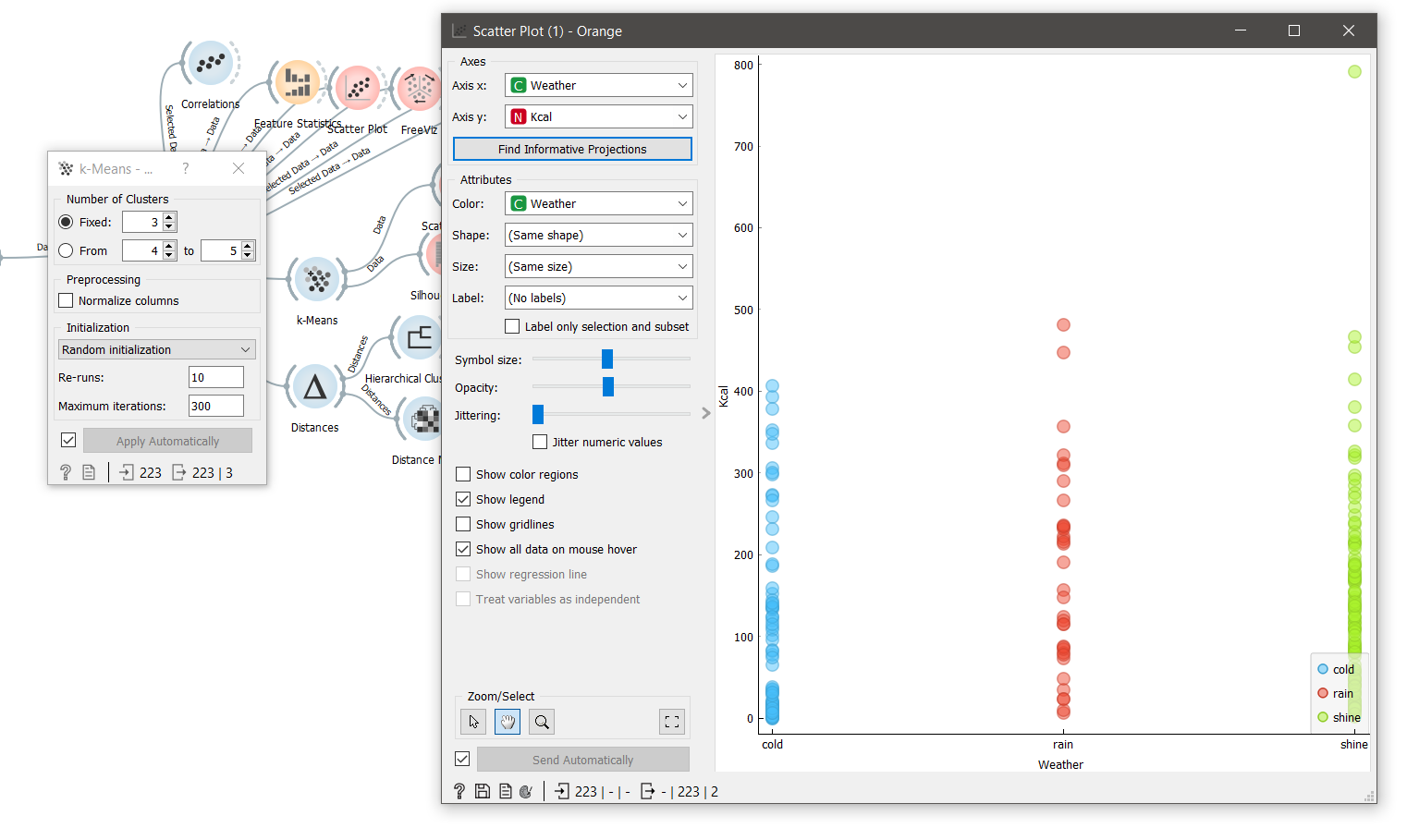
**Conclusions coming from the analysis of statistical calculations (central tendency and dispersion)**

In my case, the variance is not so large, which means that the data is not too scattered and can be grouped. And also variance is opposed to location or central tendency

My central trend is in favor of large numbers and the main holders of this trend are the category with shine as shown in the statistics above.

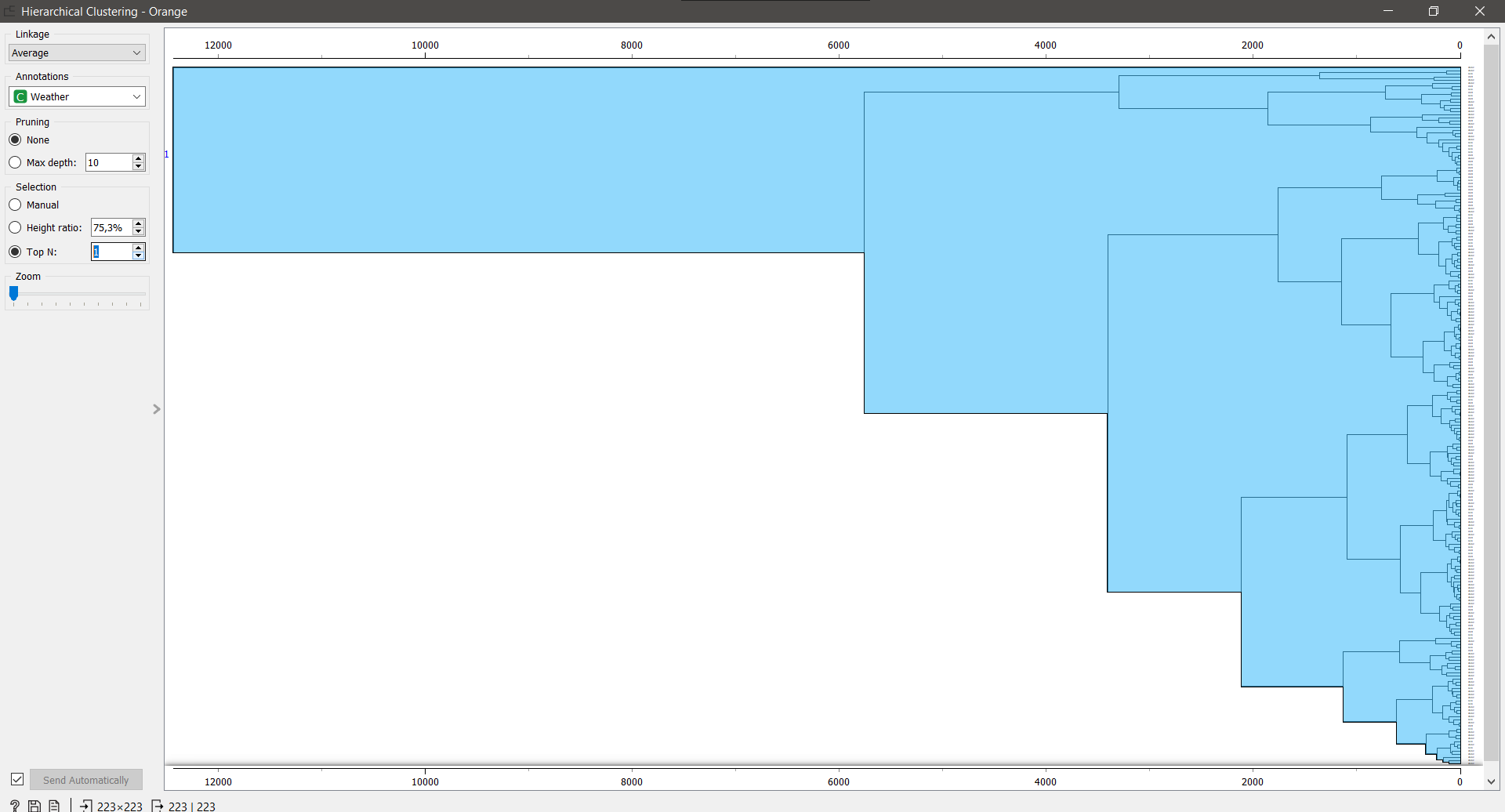
**II)**

1. **Apply two methods of unsupervised learning considered in class: (1) Hierarchical clustering and (2) K-Means.**

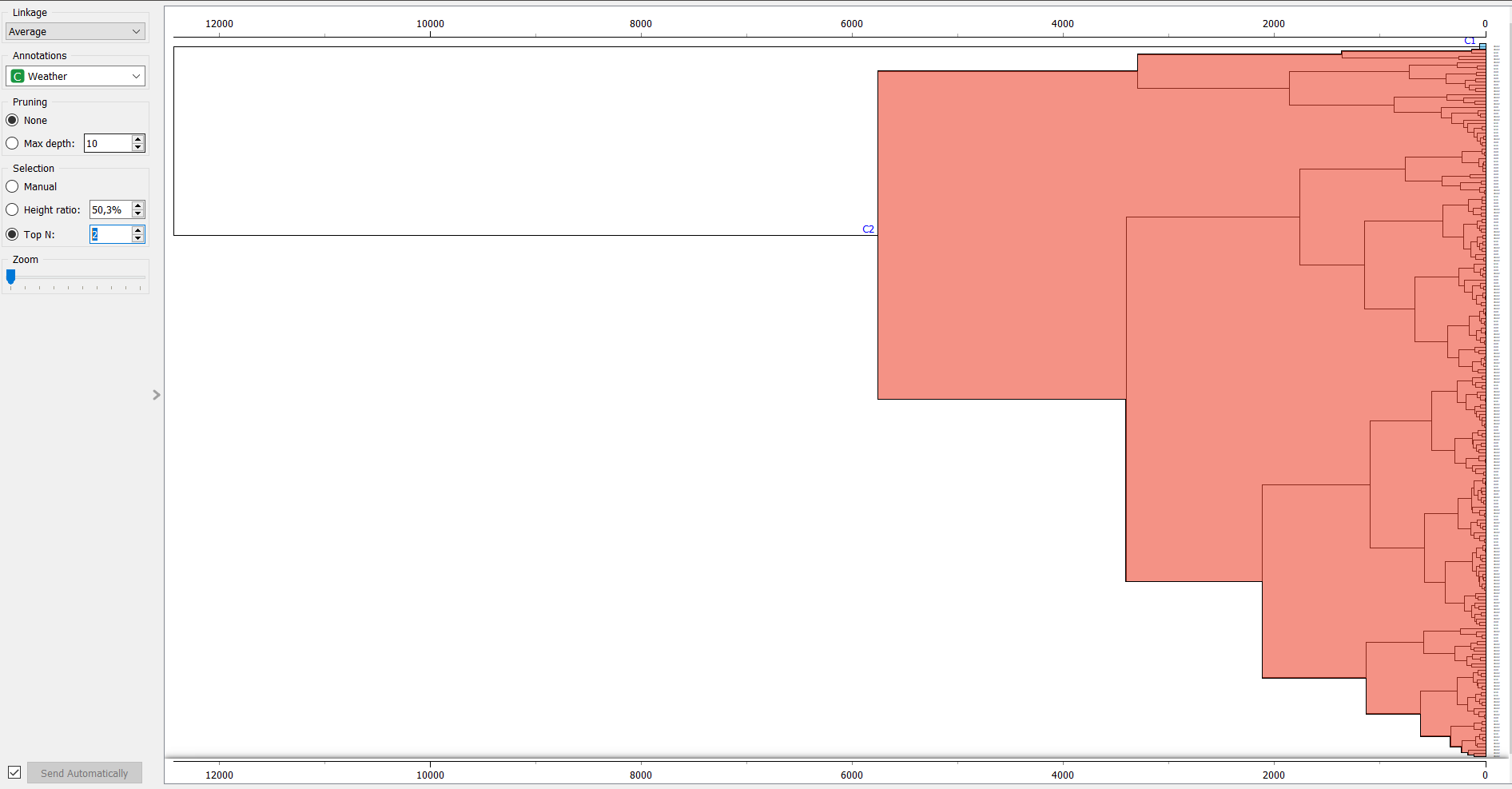
****

Using K-means and a scatter plot, I showed what would be better for the weather if it was divided into 3 categories, where and on the scatter I showed in which weather more calories were burned, on average rain and shine weather are the same, but there is a big jump during shines, which overtakes and shows good results.

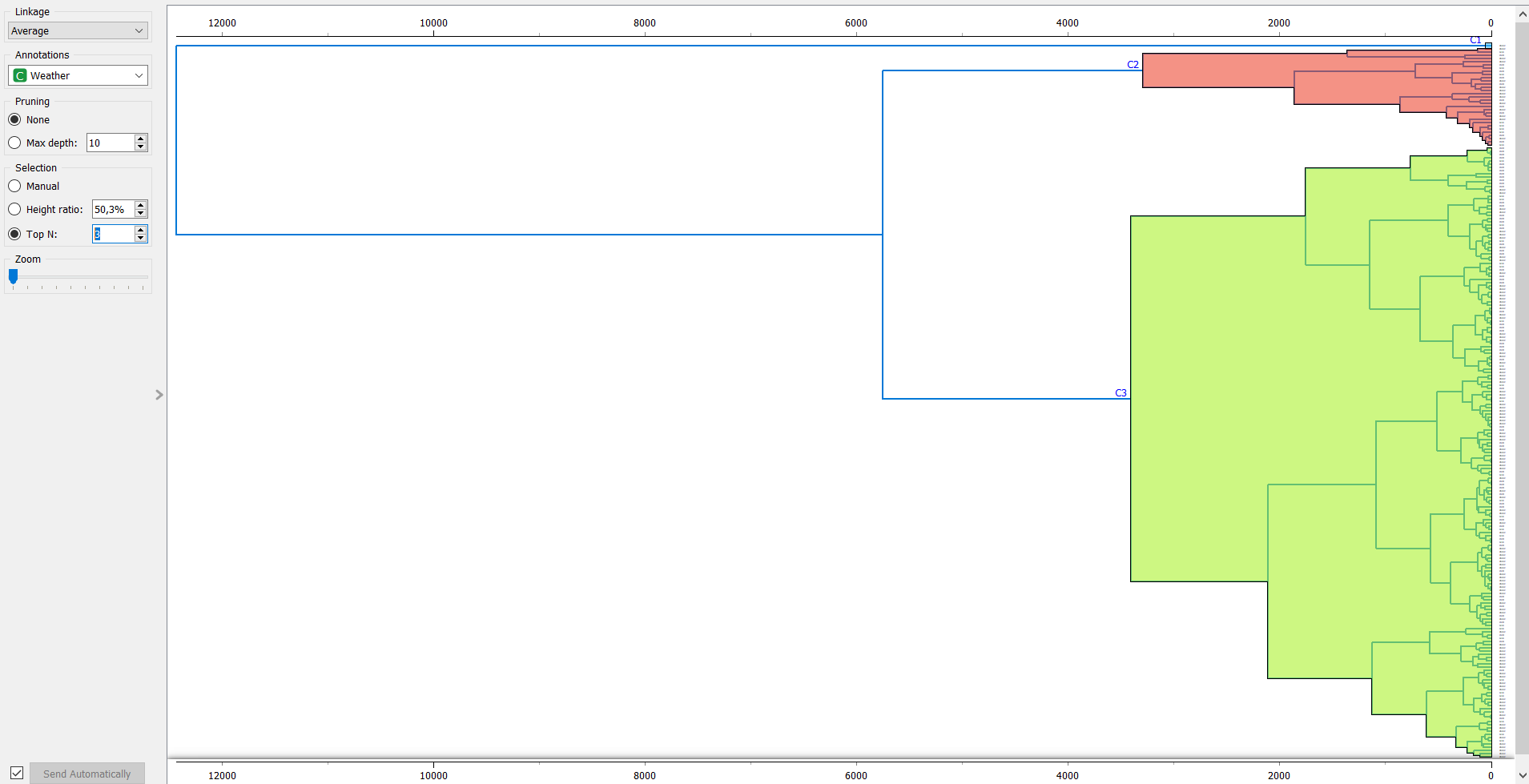
**2)** **Perform at least 3 experiments with Hierarchical clustering, freely changing the values of hyperparameters, and analysing the operation of the algorithm;**

****

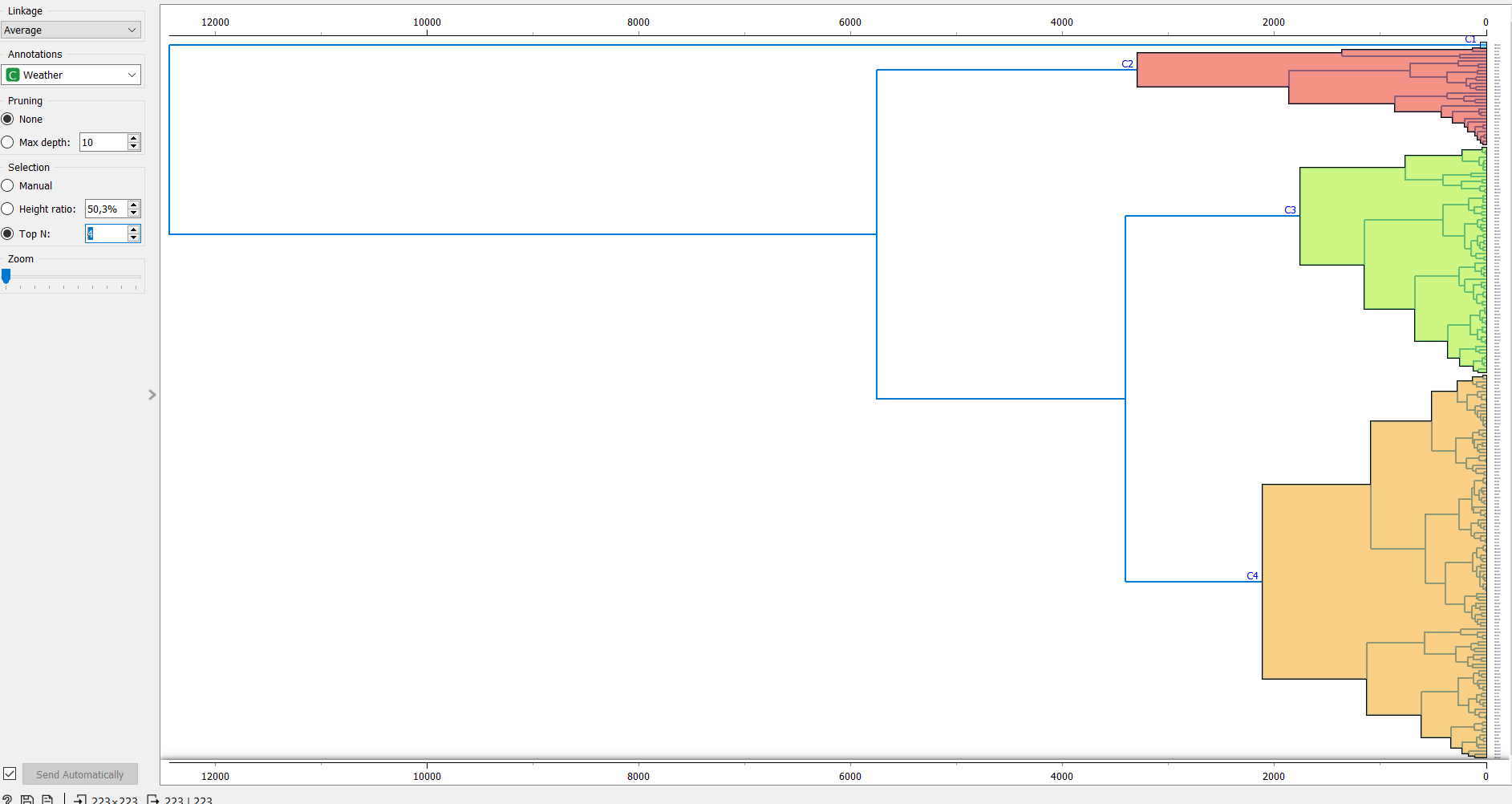
In the first experiment, having made one category, we see such a picture that all classes are under the same category and all have similarities.

****

In the second experiment, when I added the second category, there was a greater difference between the first category and the second, then the viewing angle decreased slightly and those who did not fall into the viewing angle simply switched to another category, thus the main category became 2.

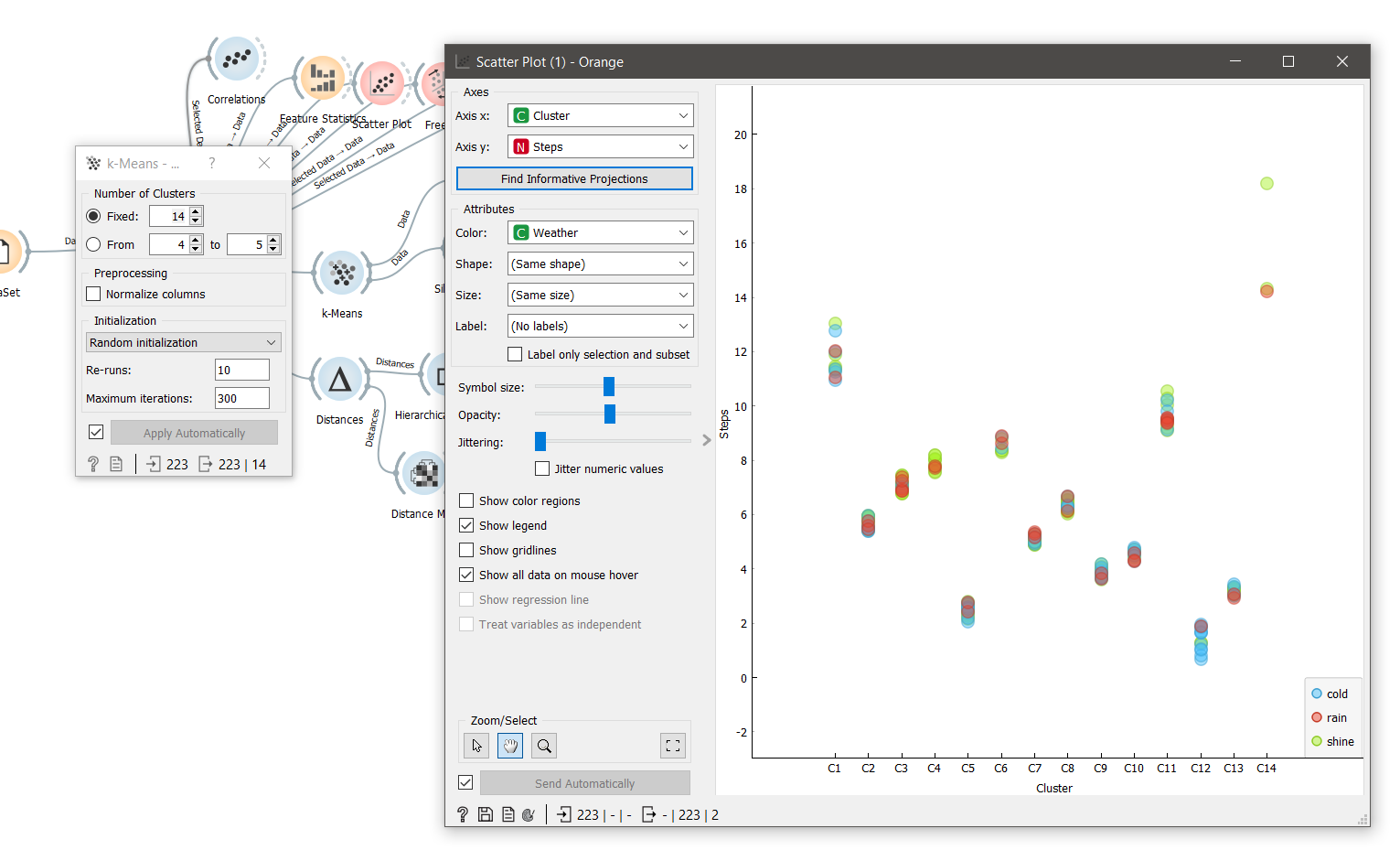
****

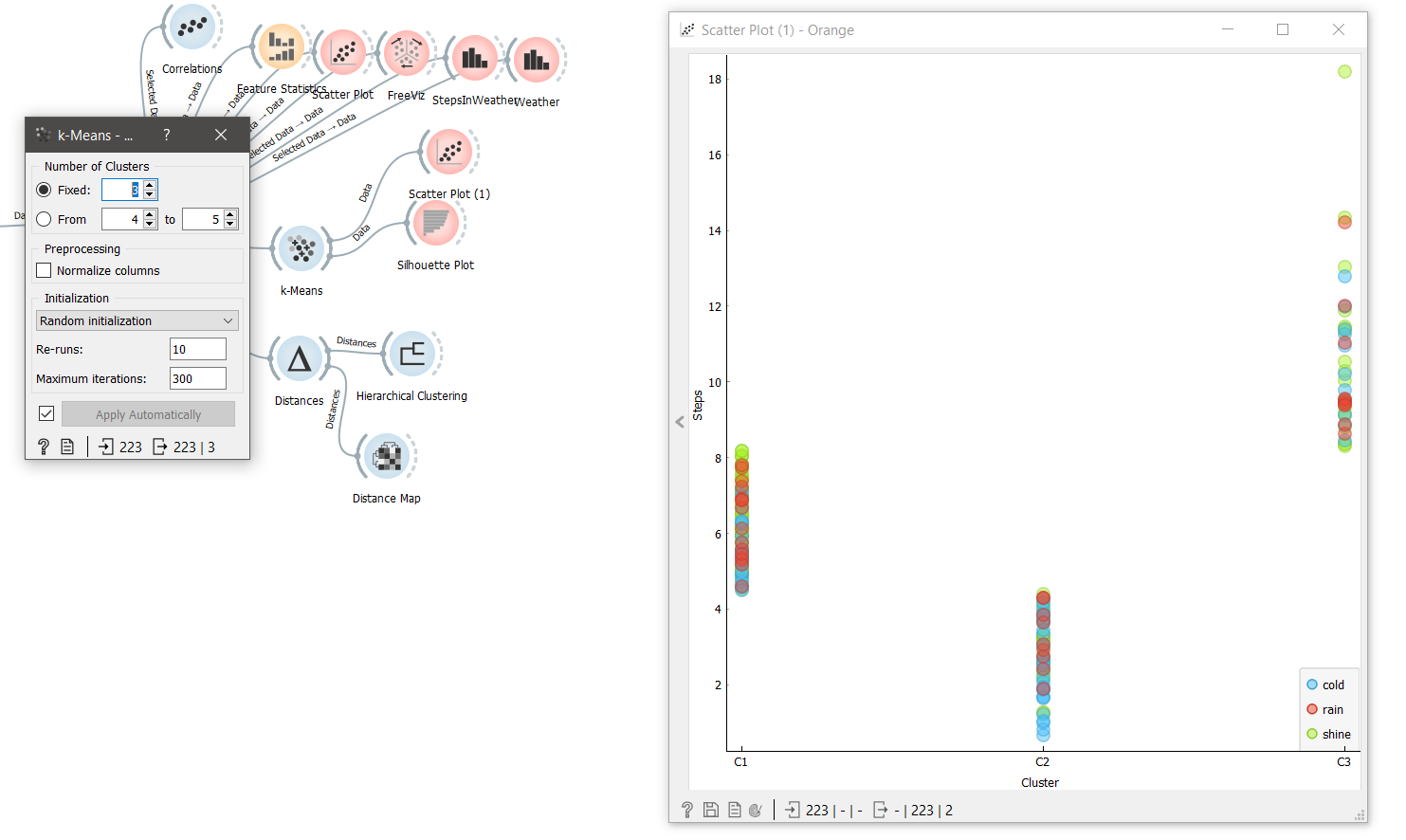
in the third experiment, when you have already made 3 categories, it is already visible, the new category pushes out the old one, thereby taking a big role, this shows the variance in my case is not large, due to this it turns out that each new category going deeper remains in the black, because there are similarities and in the deepening the visibility zone takes values concentrated on average.

****

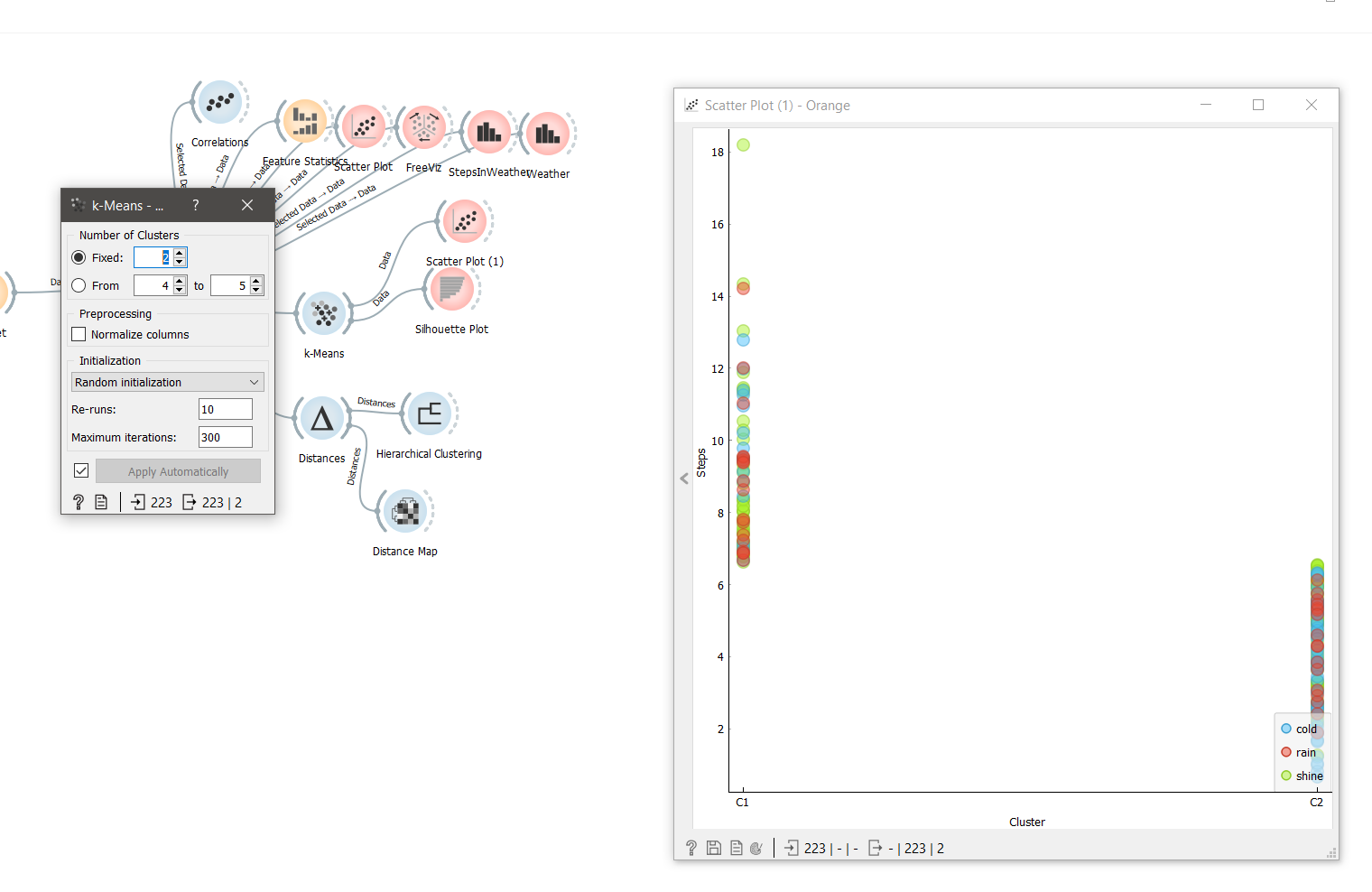
as I said above, it goes deeper into the values and more criteria for selection, those who do not fall into the selection zone simply go out and join categories in which the criteria are smaller.

**3) Perform experiments with the K-means algorithm using at least five different k values, calculate the Silhouette Score, and analyse the performance of the algorithm.**

**** Using the k-mean is not reasonable and simply spreading it over a large number of categories, the algorithm will distribute them so deeply that it will be difficult for a person to understand by what criteria it is divided. Below I will show in simple divisions into small categories.

****

Here I took 3 categories, more precisely clusters, and we can see that they are divided into minimum, average and maximum values in steps.

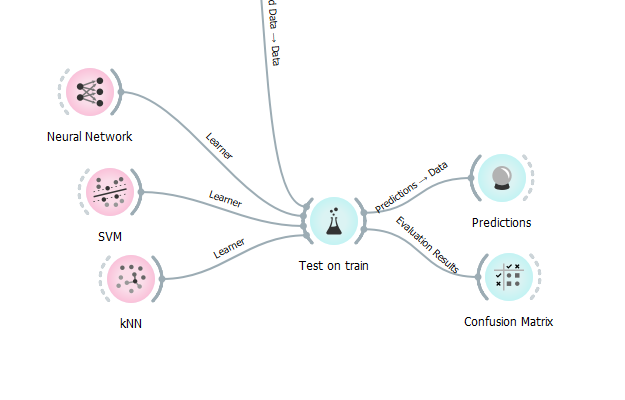
****

Further dividing into 2 categories, we can see that now the algorithm has divided them into maximum values and minimum ones, neglecting their classes. Steps have already played a role here, thereby distributing them into maximum and minimum.

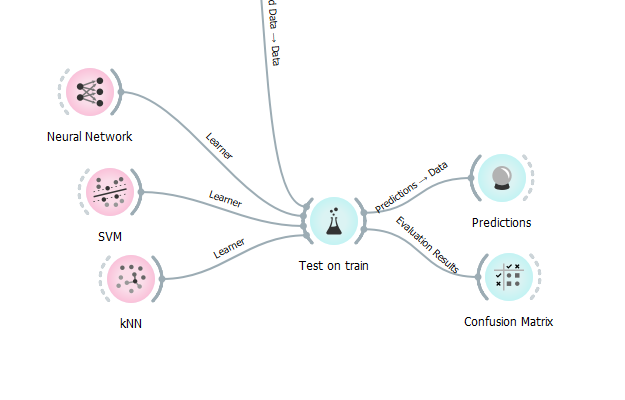
Based on the **analysis,** I can say that the best version, when they are divided by 3, it becomes no longer clear by what criteria it is divided, what plays a role in division and more and more, as I showed in the first experiment, when I immediately used 14 clusters, which is even an inveterate the person would be confused.

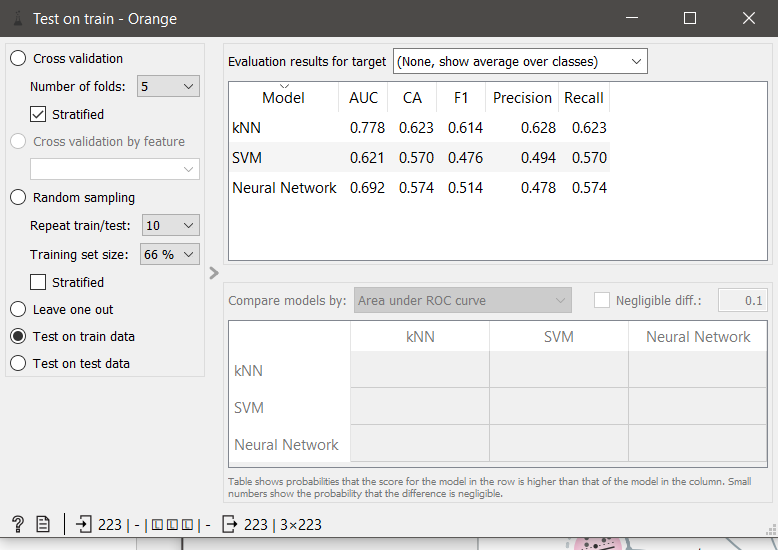
**III)**

1. **Choose at least two supervised learning methods that are suitable for classification task. You can use the methods considered in class and any other of the algorithms available in the Orange tool for classification task.**

****

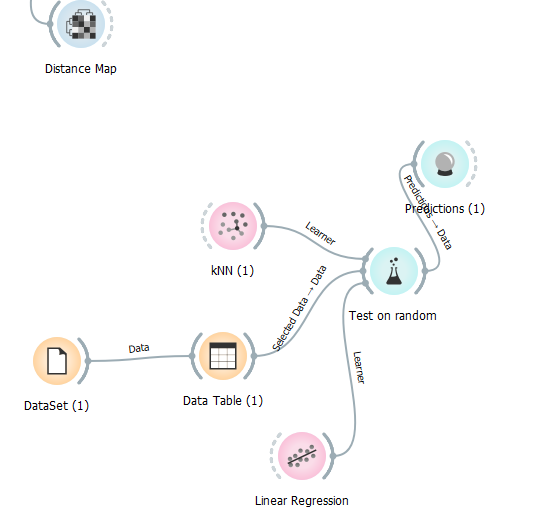
1. **Divide your dataset in training and test sets**

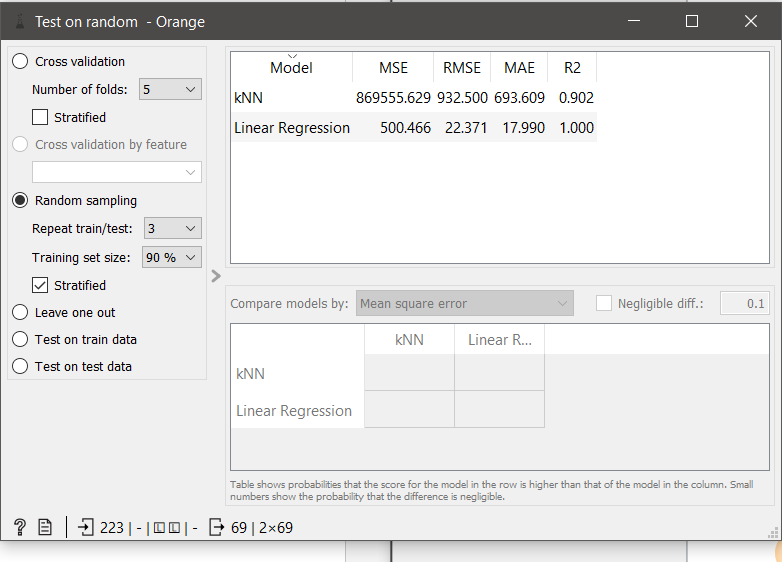


****

**Train test**

Train is a method for measuring the accuracy of a model. Training because I took 100% training dataset. Here I am training the model using the training set**.**

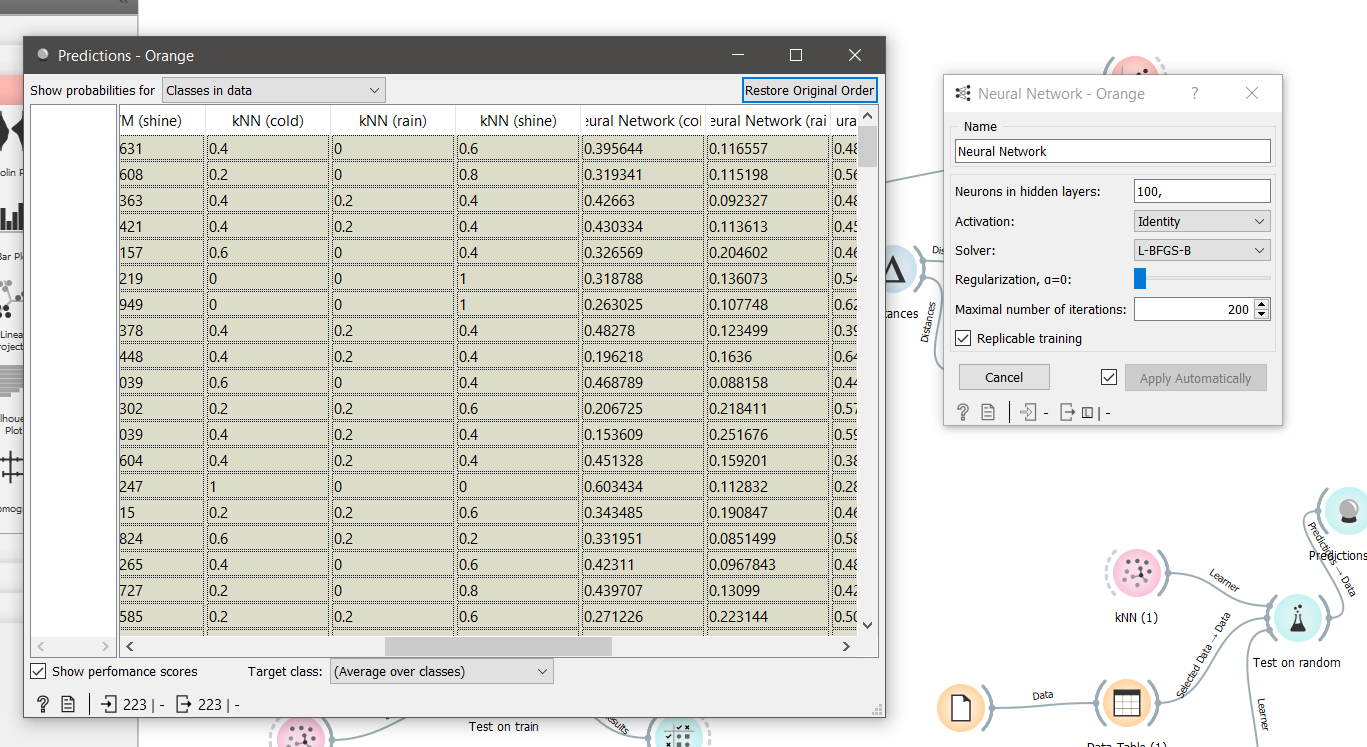
****

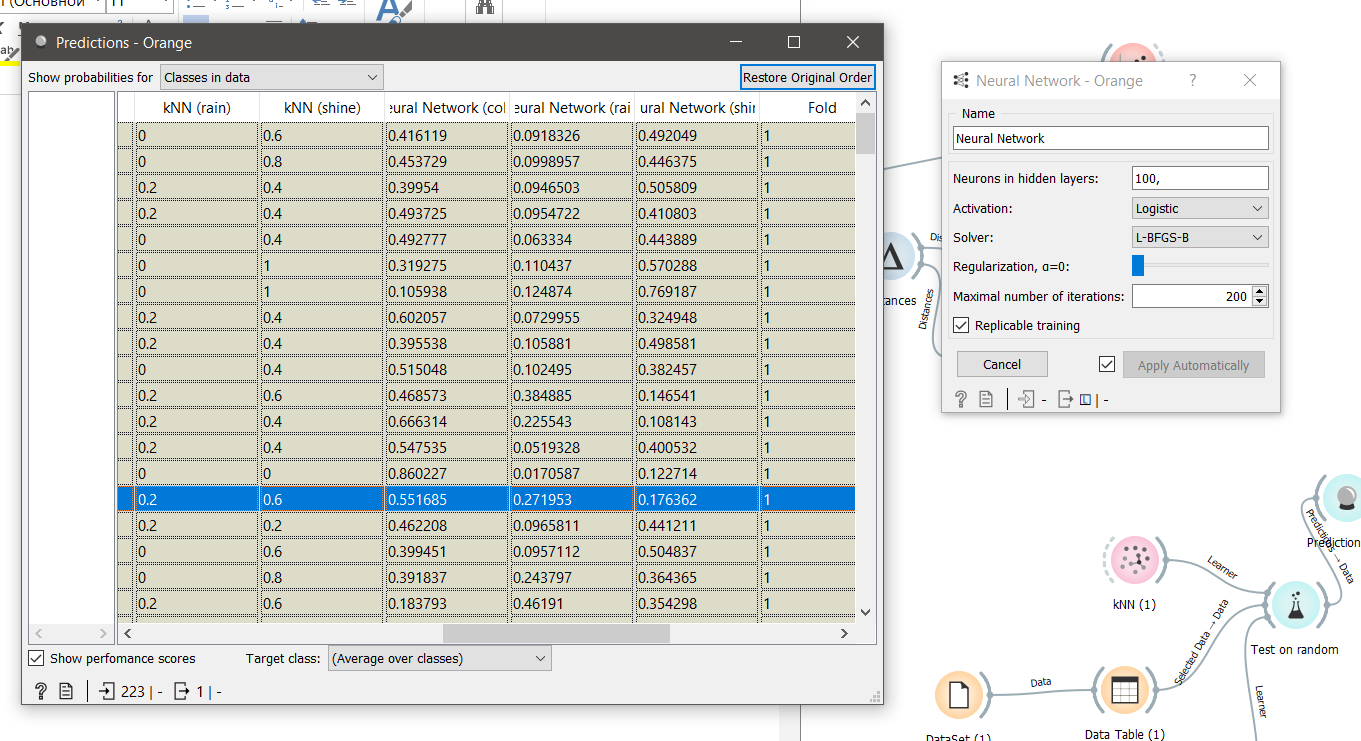
****

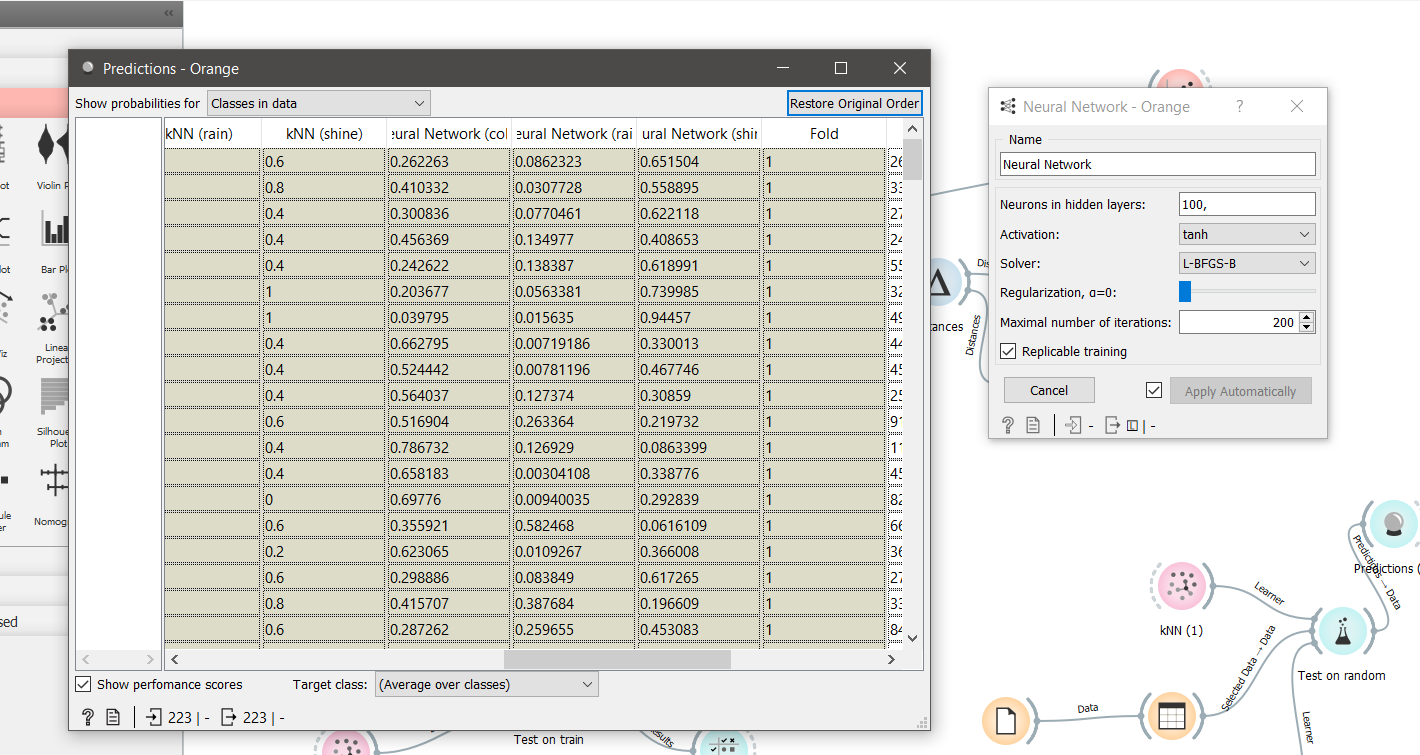
**Tests sets.**

A test is a method for measuring the accuracy of a model. The test because I took the training dataset for 90% and for the test 10%. And I use 10% of the whole set of data for the test. Ie 10% will be here in the lead role.

1. **For each algorithm, perform at least 3 experiments using the training dataset, changing the values of the algorithm hyperparameters and analysing the algorithm performance metrics.**



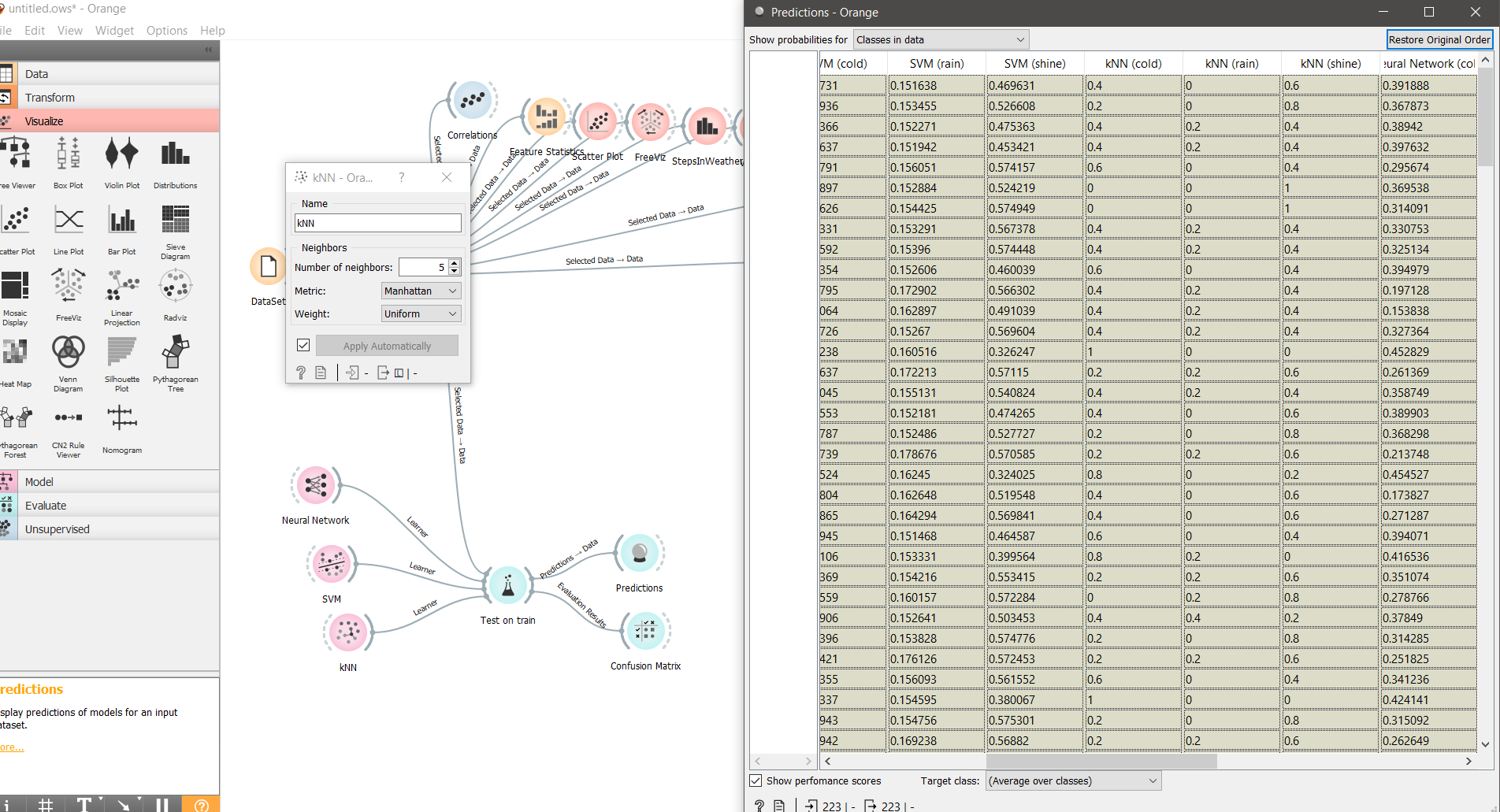


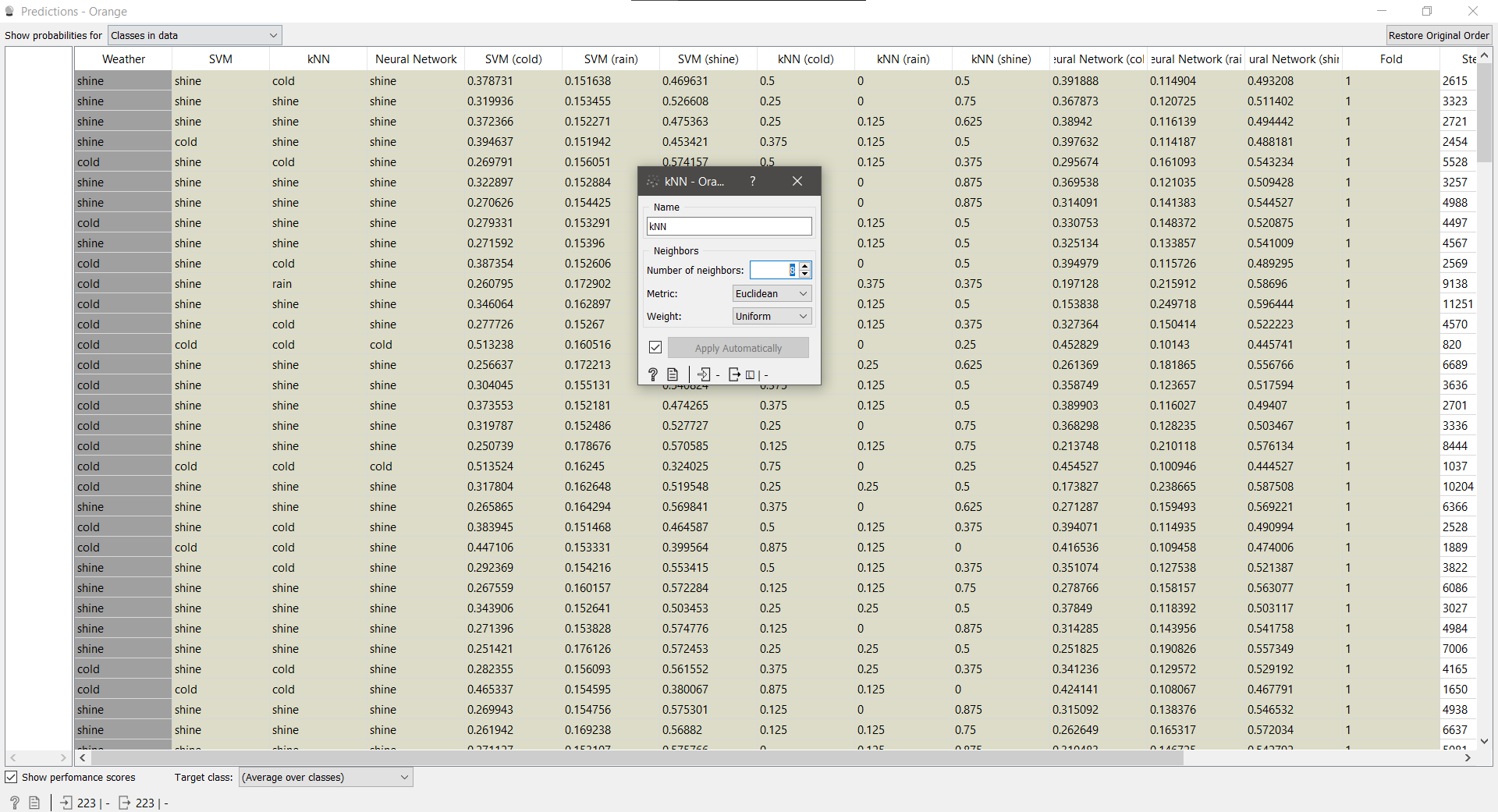


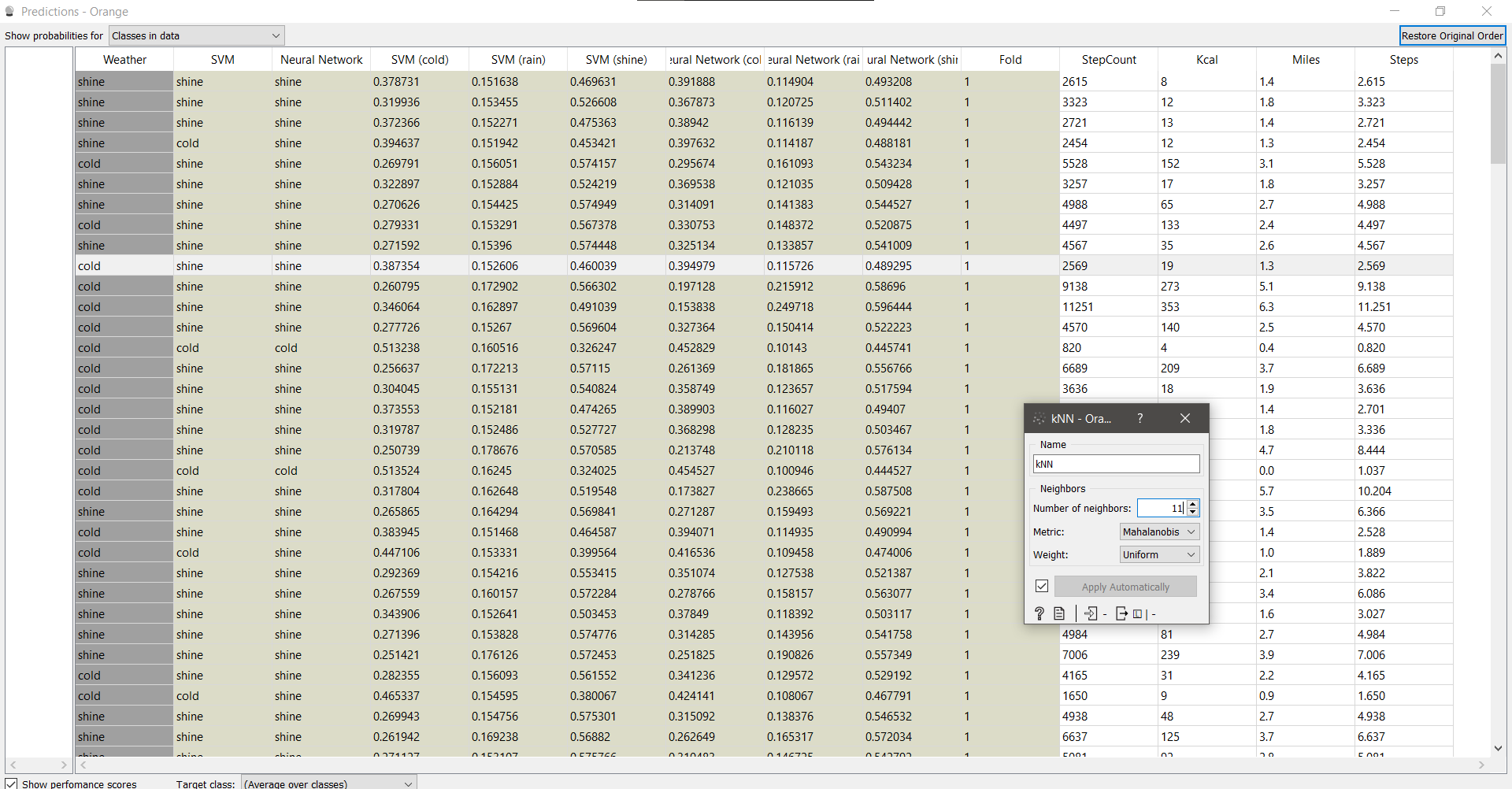
My first choice was neural networks. I changed both the Activation and the solver, but L-BFGS-B gave the best option for everyone and the smallest numbers, so I left it. But to test me, Activation, as a result, as you can see, each parameter will affect the prediction table. Identity activation showed that the non-ironic network gives 0.395 in the cold, I thought that was quite suitable, but after checking the rest, I was surprised.

I changed the activation to logistics, thereby increasing the non-ironic network by as much as 0.02, this already fit better, but differed from the main answer, this is also a mistake, but changing to TANH, it fell too low, almost 0.14 did not fit at all.

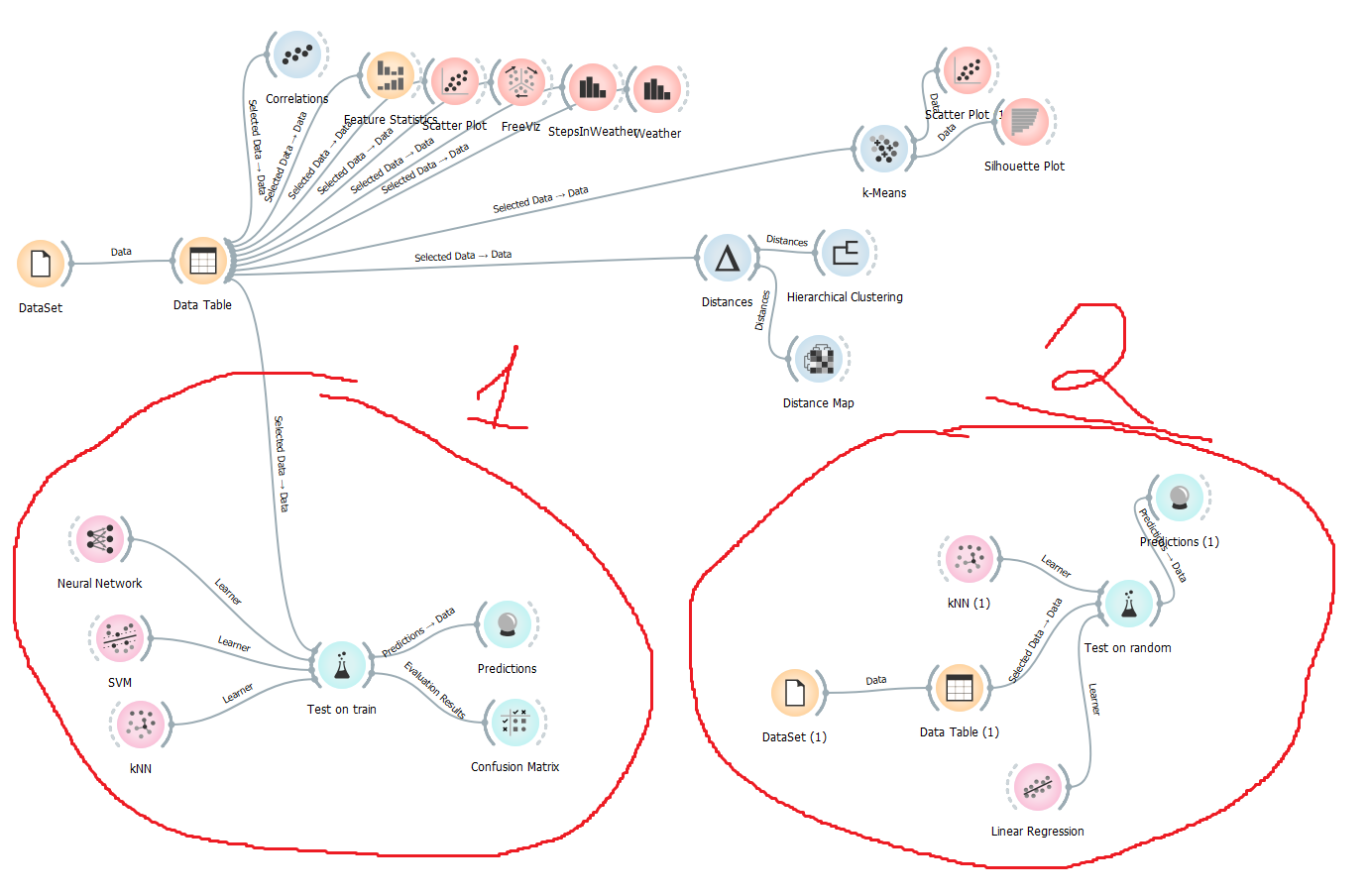
And the best option was Logistics,







the next one was KNN, it seems everything is simple, but he couldn’t predict exactly, but looking at other algorithms, KNN had more accuracy, in the first part I increased the neighbors to check whether the values will change and whether this will help the calculations. As it turned out - yes. I didn’t hit the first one, I did it at random, then when I found the best option for neighbors, I switched to metrics and I was surprised ... At mahalanobis, knn disappeared completely ... no neighbors helped bring it back, also with another , and I decided that it was better to leave Euclidian. And to choose the best neighbors for close calculation and guessing, so I stopped at 8 neighbors, the more neighbors, the less likely it is, and when it gets closer to 99, it completely disappears.



In the first case, I used the knn algorithm, this is the essence of the method on an intuitive level: look at the neighbors around, which of them prevail, you are the one. Next, SVM selects extreme points / vectors that help in creating a hyperplane. These edge cases are called support vector machines, and hence the algorithm is called a support vector machine. And then there was - Neural Network. Neural Network - is a series of algorithms that try to recognize the main relationships in a data set using a process that mimics the work of the human brain.

In the second case I used Linear Regression. Linear Regression- an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

Then I used the test and score for testing. The widget tests learning algorithms.

I also used the predictor widget to see the result of testing the data in the prediction.

Also at the end, I compared the two parts, the first in training and the second in test outputs. And maybe this is due to the fact that I gave 10% to the test, but basically the predictions were hit, and it’s not worth comparing steps and weather. But I couldn’t use letters in the second part, due to the fact that there is a linear regression, so skipping gave the target to steps. But in this case, the test one won.

**REF:**

[**https://orangedatamining.com/widget-catalog/evaluate/testandscore/**](https://orangedatamining.com/widget-catalog/evaluate/testandscore/)

[**https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/**](https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/)

[**https://www.investopedia.com/terms/**](https://www.investopedia.com/terms/)

[**https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm**](https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm)

[**https://proglib.io**](https://proglib.io)

**By Khusan Khujakeldiev 191ADB073**