

## **COMP 1942 Project Phase 2 - Design Report**

**Wijaya, Andrew Sebastian - 20994984**

**Kusnadi, Vicko Nicholas - 21022506**

### **I. Introduction**

In this report, we are going to design a report for a project that utilizes Classification with XLMiner. The project is designed to develop five distinct models, each offering unique parameters and prediction results that will be used in Phase 3 of the Project. The data used in this project comes from two sources. The first, referred to as “Data 1”, is raw data extracted from the training sheet of Phase 1. The second, “Data 2”, is a transformation of Data 1 into a numerical format, with the 'native-country' data attribute removed to streamline the analysis.

The classification learning models used in this project include the Decision Tree, Naive Bayesian, and K-Nearest Neighbor models. Each of these models brings a different approach to classification, providing a comprehensive analysis of the data. Through this project, we aim to explore the capabilities of these models and to guide the reader through the steps of each model, using the power of XLMiner to drive our data analysis.

### **II. Data Preparation**

To prepare the data we first duplicate the page “Training” twice and rename one of the duplicates into “Data 1” as shown on Fig 1.0. Then we split data, selecting the first 8000 entries as the training data and next 2000 entries as the testing data. This step is shown in Fig 1.1 and Fig 1.2.

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
1	24	Private	HS-grad	9	Married-civ-spouse	Tech support	Husband	White	Male	0	0	50	United States
2	41	Private	HS-grad	9	Married-civ-spouse	Transport moving	Husband	White	Male	0	0	50	United States
3	42	Private	9th	5	Married-civ-spouse	Craft repair	Husband	White	Male	0	0	40	Mexico
4	44	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Prof specialty	Husband	White	Male	0	0	80	United States
5	27	Private	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	United States
6	29	Local-gov	HS-grad	9	Married-civ-spouse	Prof specialty	Not-in-family	White	Female	0	0	50	United States
7	30	Private	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United States
8	44	Private	Bachelors	13	Married-civ-spouse	Tech support	Husband	White	Male	0	0	40	United States
9	44	Local-gov	Masters	16	Married-civ-spouse	Prof specialty	Husband	White	Male	0	0	40	United States
10	27	Private	Some-college	10	Never-married	Other service	Wife	White	Female	0	0	30	Germany
11	29	Private	Bachelors	13	Married-civ-spouse	Sales	Wife	White	Female	0	0	50	United States
12	18	Private	HS-grad	9	Married-spouse-absent	Other service	Down-child	Black	Male	0	0	40	Jamaica
13	42	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	40	United States
14	35	Self-emp-not-inc	Bachelors	13	Divorced	Adm-clerical	Not-in-family	White	Male	0	2862	40	United States
15	25	Self-emp-inc	HS-grad	9	Married-civ-spouse	Exec managerial	Husband	White	Male	0	0	40	United States
16	40	Local-gov	HS-grad	9	Married-civ-spouse	Protective serv	Wife	Black	Female	0	0	25	United States
17	33	Local-gov	Masters	16	Married-civ-spouse	Exec managerial	Husband	Black	Male	0	0	40	United States
18	29	Private	HS-grad	9	Never-married	Adm-clerical	Unmarried	Black	Female	0	0	40	Jamaica
19	29	Private	Bachelors	13	Married-civ-spouse	Other service	Husband	White	Male	7298	0	40	United States
20	31	State-gov	Some-college	10	Married-civ-spouse	Machine-op impct	Husband	White	Male	0	0	50	United States
21	32	Self-emp-inc	Bachelors	13	Never-married	Exec managerial	Not-in-family	White	Female	0	0	25	Canada
22	17	Private	HS-grad	9	Divorced	Sales	Unmarried	White	Female	0	0	40	Germany
23	24	Private	25th	7	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	Germany

Fig 1.0

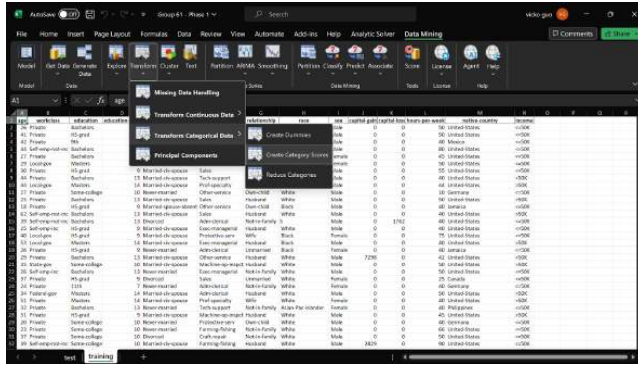
	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
24	40	Private	Bachelors	13	Never-married	Prof specialty	Not-in-family	White	Male	4050	0	40	United States
25	7069	Private	HS-grad	9	Married-civ-spouse	Other service	Husband	White	Male	0	0	40	Honduras
26	8023	Private	HS-grad	9	Married-civ-spouse	Sales	Not-in-family	White	Male	0	0	55	United States
27	8023	Private	Bachelors	13	Divorced	Sales	Not-in-family	White	Male	0	0	45	United States
28	8023	Private	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	44	United States
29	8023	Local-gov	HS-grad	9	Never-married	Protective serv	Not-in-family	White	Male	0	0	42	United States
30	8023	Local-gov	Some-college	10	Married-civ-spouse	Protective serv	Husband	White	Male	0	0	50	United States
31	8023	Private	Bachelors	13	Never-married	Sales	Not-in-family	White	Male	0	0	35	United States
32	8023	Private	HS-grad	9	Married-civ-spouse	Health care	Husband	White	Male	0	0	40	United States
33	8023	Private	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	5103	0	40	United States
34	8023	Private	Masters	16	Married-civ-spouse	Sales	Husband	White	Male	5103	0	40	United States
35	8023	Private	Bachelors	13	Never-married	Prof specialty	Not-in-family	White	Female	0	1977	40	United States
36	8023	Private	Some-college	10	Never-married	Protective serv	Husband	White	Male	0	0	40	United States
37	8023	Private	HS-grad	9	Married-civ-spouse	Protective serv	Husband	White	Male	0	0	40	United States
38	8023	Private	HS-grad	9	Married-civ-spouse	Machine-op impct	Husband	White	Male	7298	0	40	United States
39	8023	Private	Some-college	10	Married-civ-spouse	Prof specialty	Unmarried	White	Female	0	0	40	Germany
40	8023	Private	HS-grad	9	Married-civ-spouse	Transport moving	Husband	White	Male	0	0	40	United States
41	8023	Private	HS-grad	9	Married-civ-spouse	Prof specialty	Not-in-family	White	Male	0	0	40	United States
42	8023	Private	25th	7	Never-married	Sales	Not-in-family	White	Male	4050	0	50	United States
43	8023	Private	HS-grad	9	Married-civ-spouse	Exec managerial	Husband	White	Male	0	0	80	United States
44	8023	Private	Some-college	10	Divorced	Machine-op impct	Unmarried	White	Male	0	0	40	United States

Fig 1.1

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
45	8023	Private	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	44	United States
46	8023	Local-gov	HS-grad	9	Never-married	Protective serv	Not-in-family	White	Male	0	0	42	United States
47	8023	Local-gov	Some-college	10	Married-civ-spouse	Protective serv	Husband	White	Male	0	0	50	United States
48	8023	Private	Bachelors	13	Never-married	Sales	Not-in-family	White	Male	0	0	55	United States
49	8023	Private	HS-grad	9	Married-civ-spouse	Health care	Husband	White	Male	0	0	40	Mexico
50	8023	Private	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	5103	0	40	United States
51	8023	Private	Masters	16	Married-civ-spouse	Sales	Husband	White	Male	5103	0	40	United States
52	8023	Private	Bachelors	13	Never-married	Prof specialty	Not-in-family	White	Female	0	1977	40	United States
53	8023	Private	Some-college	10	Never-married	Protective serv	Husband	White	Male	0	0	40	United States
54	8023	Private	HS-grad	9	Married-civ-spouse	Machine-op impct	Husband	White	Male	7298	0	40	United States
55	8023	Private	Some-college	10	Married-civ-spouse	Prof specialty	Unmarried	White	Female	0	0	40	Germany
56	8023	Private	HS-grad	9	Married-civ-spouse	Transport moving	Husband	White	Male	0	0	40	United States
57	8023	Private	HS-grad	9	Married-civ-spouse	Prof specialty	Not-in-family	White	Male	0	0	40	United States
58	8023	Private	25th	7	Never-married	Sales	Not-in-family	White	Male	4050	0	50	United States
59	8023	Private	HS-grad	9	Married-civ-spouse	Exec managerial	Husband	White	Male	0	0	80	United States
60	8023	Private	Some-college	10	Divorced	Machine-op impct	Unmarried	White	Male	0	0	40	United States

Fig 1.2

The next step is to take the second duplicate of the “Training” page from phase 1 and transform it into numerical data (as shown on Fig. 2.0.0). We select all data attributes, except “native-country” attributes and several other attributes that already have numerical value. We don’t use “native-country” attribute as it exceeds the limit of XLminer, and transform it into numerical data (as shown on Fig. 2.0.1). The resulting data is shown on Fig. 2.0.2.



*Fig. 2.0.0*

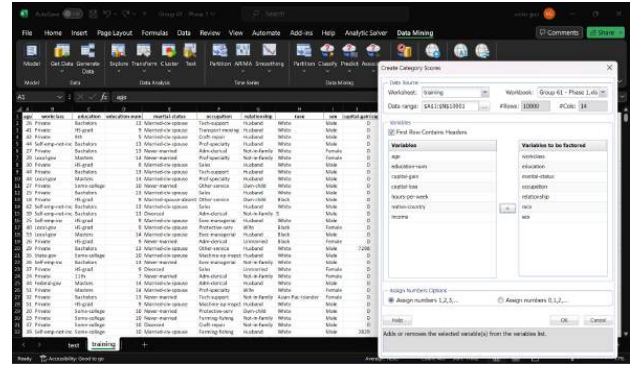
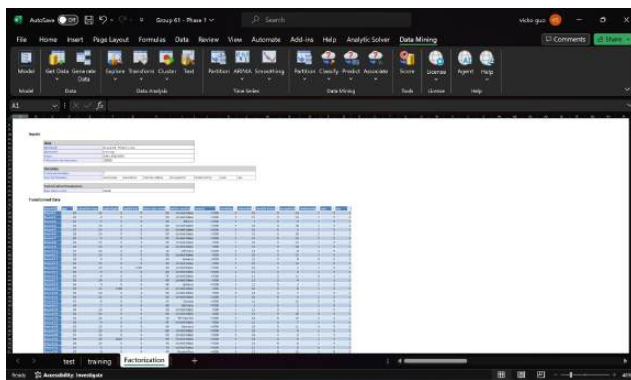


Fig. 2.0.1



*Fig. 2.0.2. Resulting numerical data from transformation*

The result of the numerical data page is renamed to “Data 2” (as shown on Fig. 2.1.0). We will now take this data and split the data for training and testing. We will use the first 8000 entries for training and the next 2000 entries for testing. We split the data by inserting 2 new rows below the 8000th entry and copying the corresponding attribute names there (as shown in Fig. 2.1.1). The final result of our preparation for “Data 2” is shown in Fig. 2.1.2 and Fig. 2.1.3. That concludes the data preparation step needed to start generating our model. Our model will then use “Data 1” and “Data 2”.

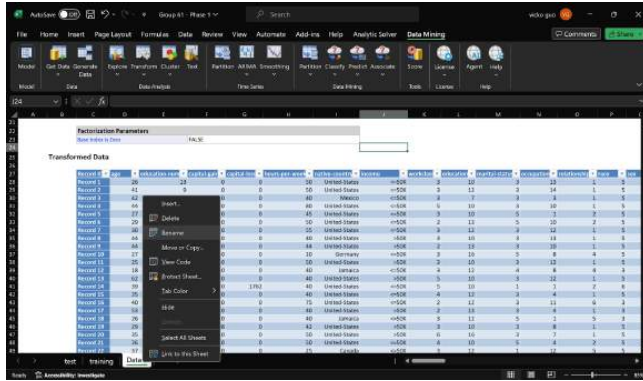


Fig. 2.1.0. Rename the data as “Data 2”

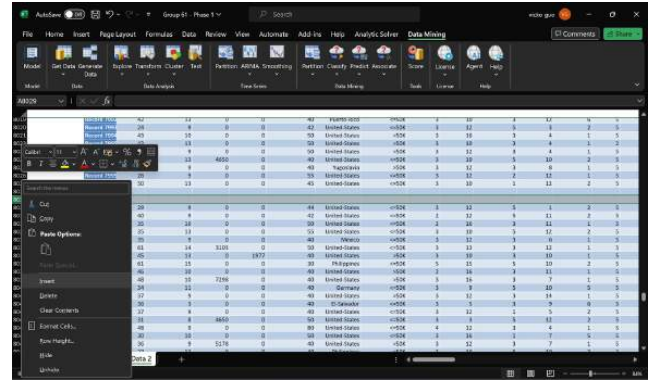


Fig. 2.1.1

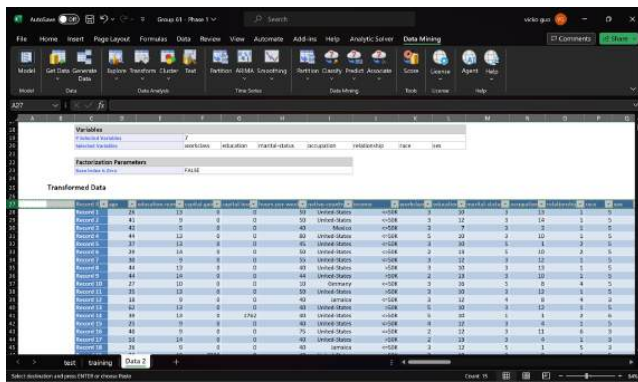


Fig. 2.1.2

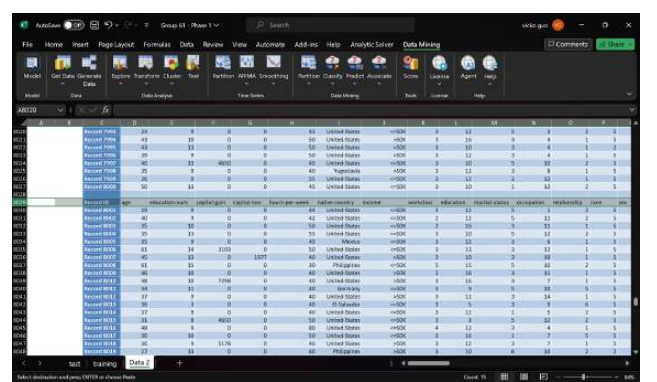


Fig 2.1.3

### III. Model 1

- ❖ Data: Data 2
- ❖ Type of Model: Decision Tree
- ❖ Parameter used:
  - Success Class: >50K
  - Number of Classes: 2
  - Success probability cut-off: 0.5
  - Tree to display: Fully Grown
  - Records in Terminal Nodes: 1
  - Maximum Number of Leaves: 7
  - Set the Test set and Training set variables to Match by Name

The first model we are generating using Classification Tree algorithm. The parameters used for this model are setting the number of Classes to 2 by default, success probability cut-off to 0.5, set Tree to display to Fully Grown, set Records in Terminal Nodes to 1, and Maximum Number of Leaves: 7. In addition, we also set the Success Class to “>50K”.

To generate, first select classify then classification tree in “Data 2” sheet (as shown in Fig. 3.0). Now we see the data tab, set the tab to “Data” tab, and set the data range to C27 until C8027, as we want to use the first 8000 entries as our training data (as shown on Fig 3.1). We set all the variables in “Variables in Input Data” to selected Variables except “Record ID”, “native-country”, and “Income”, set the “Output Variable” to “Income”. Finally, keep the success Probability Cutoff to 0.5, the Number of Classes to 2, and also the Success class to >50K.

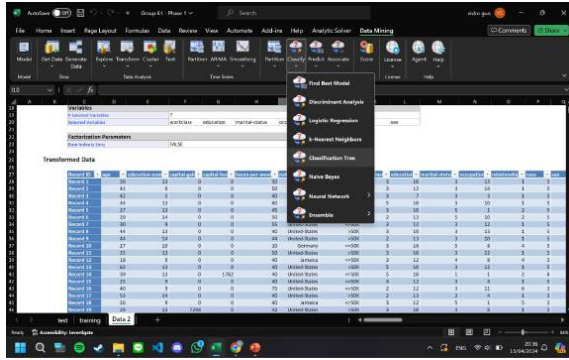


Fig. 3.0. Choose the classification Tree from XLMiner

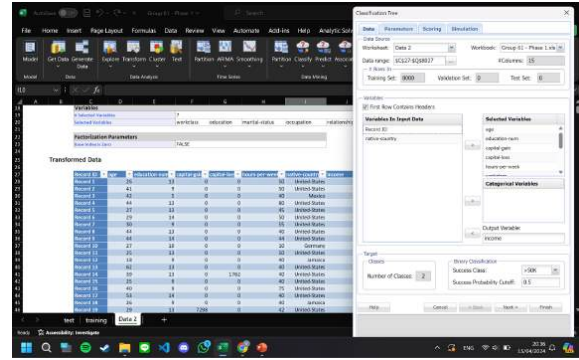


Fig. 3.1. Data tab, set the data range for the training set (C27:C8027)

In the Parameters Tab, we make changes to certain settings. We enable the inclusion of records in terminal nodes and set it to 1. We also limit the maximum number of levels to 7, which is the maximum allowed. Clicking on Trees to Display allows us to choose the Fully Grown option for displaying the complete classification tree. The settings can be seen on Fig.3.2 and Fig. 3.3.

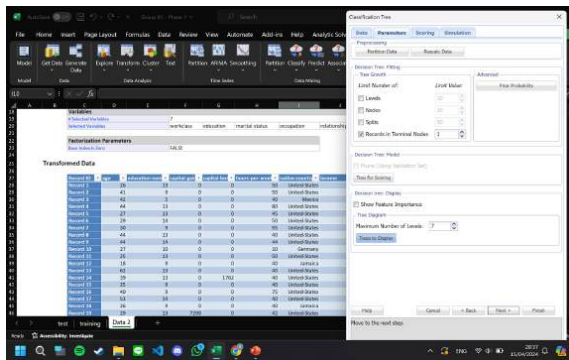


Fig 3.2. Parameters Tab

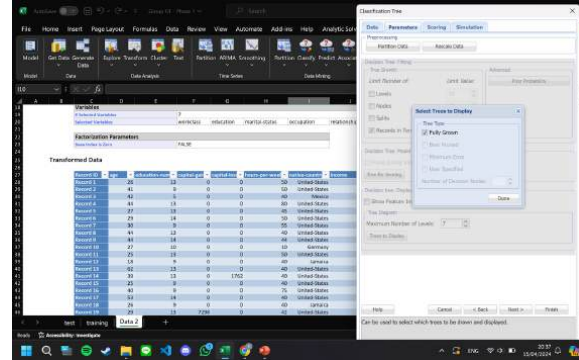


Fig 3.3. Select Tree to Display type to Fully Grown

Moving to the Scoring Tab, we evaluate the performance of the model on the training data by selecting all the relevant checkboxes which are “Detailed Report”, “Summary Report”, “Lift Charts”, and “Frequency Chart” (as shown on Fig. 3.4). In the Score New Data section, we choose the option to score data within the same worksheet as the training set. This leads to the creation of a new tab called the New Data (WS) Tab, where we set up the test data. The test data range is specified as C8029 to Q10029, with a total of 2000 data points (as shown on Fig. 3.5). We use the “Match By Name” option to match variables with the training set and then click Finish to complete the setup of the test data.



The result obtained from XLMiner may show some limitations or unexpected outcomes, which require careful analysis.

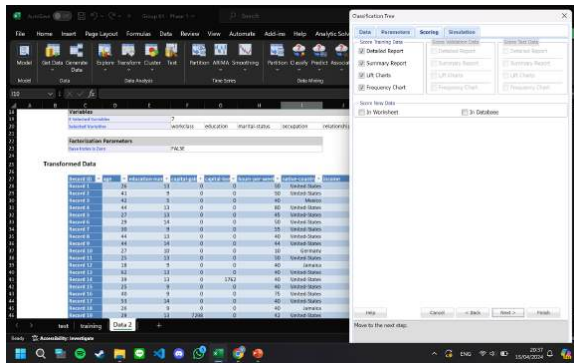


Fig. 3.4. Scoring tab

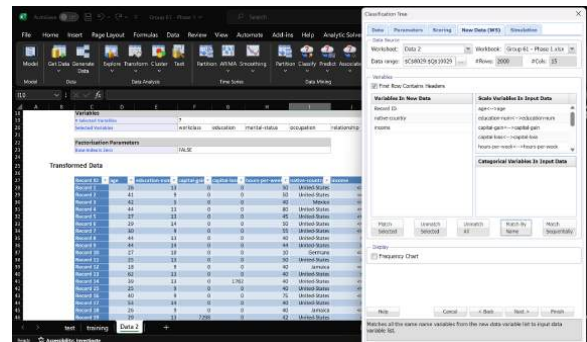


Fig. 3.5. New Data (WS) Tab. Set the data range for test set (C8029:C10029); Press “match by name” to match the training set attributes and test set attributes by name

## IV. Model 2

- ❖ Data: Data 1
- ❖ Type of Model: Naive Bayesian
- ❖ Parameter used:
  - Success Class: >50K
  - Number of Classes: 2
  - Success probability cut-off: 0.5
  - Laplace Smoothing: False
  - Prior Probability Method: Empirical
  - Show Prior Conditional Probability: True
  - Show Log Density: True
  - Set the Test set and Training set variables to Match by Name

In the second model, we will generate the training data using the Naive Bayesian algorithm. The parameter used for the second model is similar to the first model with a bit of modification. The parameters are set the Number of Classes to 2 and the success probability cut-off to 0.5. In addition, we also disable the Laplace smoothing, use the empirical method for the Prior Probability Method, Show Prior Conditional Probability, Show Log Density, and assign “>50K” as the Success Class.

The first step to generate the data is to select "Classify" and then "Naive Bayes" in the "Data 1" sheet (as shown in Fig. 4.0). In the data tab, set the data range to A1 until A8001, as we want to use the first 8000 entries as our training data (as shown in Fig 4.1). Then, set the success probability cutoff to 0.5, the number of classes to 2, and the success class to ">50K". Next, press the "Next" button.



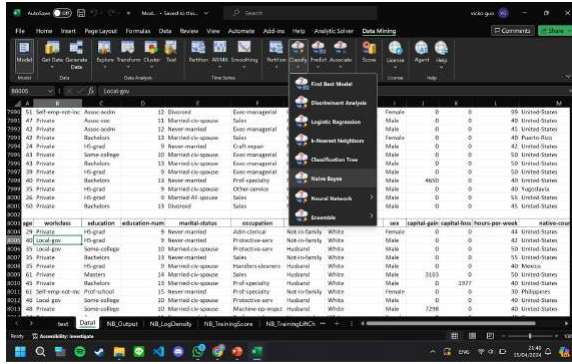


Fig.4.0. Naive Bayesian Classifier Selection

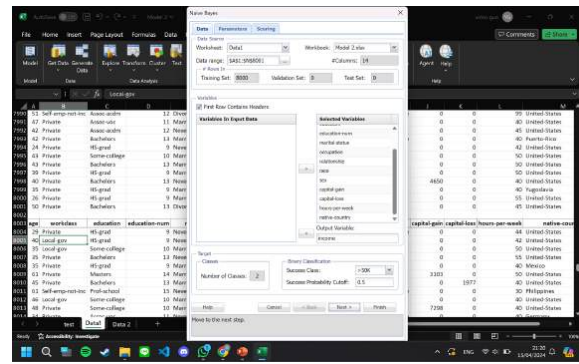


Fig.4.1. Data. TabSelected Variable and the Output Variable Selection and Data range selection from A1 to A8001

In the parameters tab, disable Laplace Smoothing and select "Prior Probability" (as shown in Fig 4.2). In the pop-up, select the "Empirical" option for the Prior Probability Method (as shown in Fig 4.3). In the "Display Options", check all the boxes in the "Naive Bayes: Display" section, and then press the "Next" button again to move to the "Scoring Tab" (as shown in Fig 4.4).

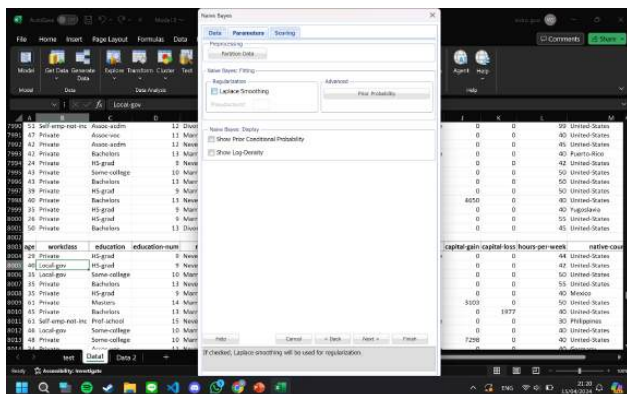


Fig.4.2. Parameter Tabs

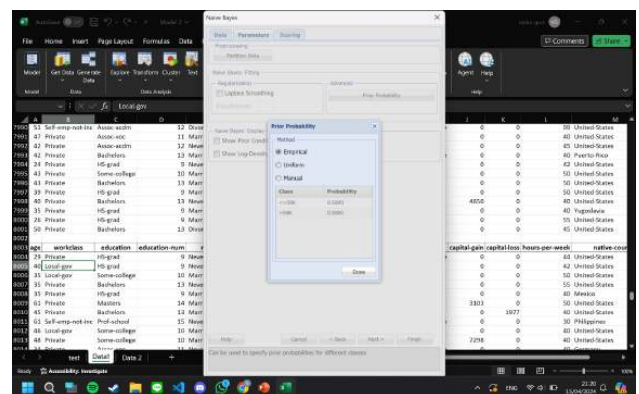


Fig.4.2.Prior Probability Tab

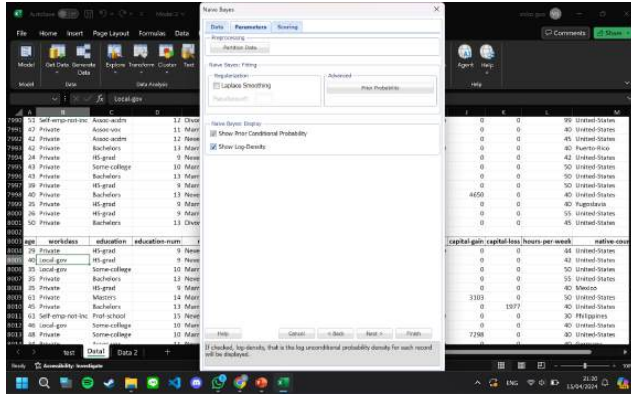


Fig.4.4. Parameter Tabs

In the "Scoring Tab", check all the boxes in the "Score training data" section, namely "Detailed Report", "Summary Report", "Lift Charts", and "Frequency Chart" (as shown in Fig 4.5). In addition, in the "Score New Data" area, select the "In Worksheet" option. After selecting the "In Worksheet" option, the "New Data (WS)" tab will appear (as shown in Fig 4.6).

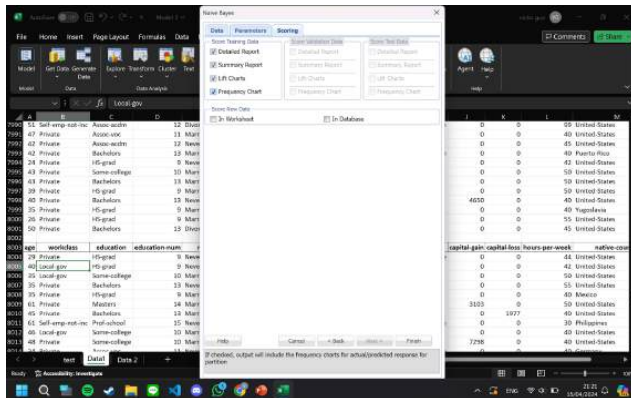


Fig.4.5. Scoring Tab

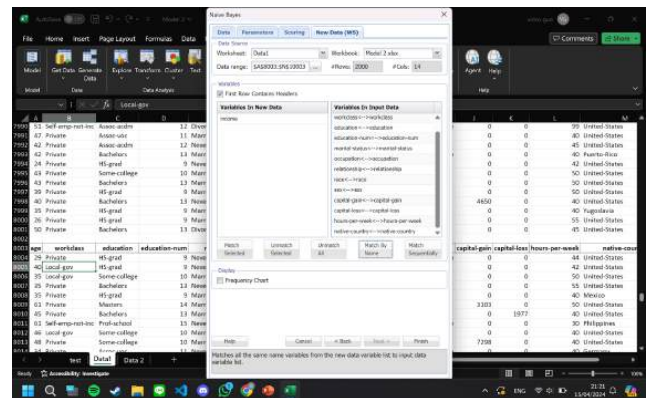


Fig.4.6. New Data(WS) Tab. Set the data range for the test set (A8003:N10003); Press "match by name" to match the training set attributes and test set attributes by name

The "New Data (WS)" tab is used to set the test set. Set the data range from A8003 to N10003, as we are using 2000 data points for testing. Finally, click the 'Match By Name' button to match the attributes of the training data with the test data set, and click 'Finish' to generate the results by XLMiner (as shown in Fig 4.6).

## V. Model 3

- ❖ Data: Data 2
- ❖ Type of Model: K Nearest Neighbor
- ❖ Parameter used:
  - Success Class: >50K
  - Number of Classes: 2
  - Success probability cut-off: 0.5
  - Fixed K with K = 3
  - Prior Probability to Empirical
  - Set the Test set and Training set variables to Match by Name

Model 3, based on K Nearest Neighbor, was applied to the Data 2 dataset with specific parameters. The analysis involved classifying the data into two classes based on the "Income" variable, with the success class defined as ">50K." A success probability cutoff of 0.5 was used to determine the class assignment. The K Nearest Neighbor algorithm was configured with a fixed value of K=3. Prior probability estimation was performed using the empirical method.

In the Data 2 Sheet, select Data mining tab on Excel, and then select Classify, and choose K Nearest Neighbor as the Algorithm type for this model (as shown in Fig 5.0). In the Data Tab, set the data range from C27 to C8029, since 8000 data points are for training sets. All variables except "Record ID," "native-country," and "Income" were selected as the Selected Variables. The "Income" variable was designated as the Output Variable (as shown in Fig 5.1). After that, press the Next button to go to the Parameter Tab

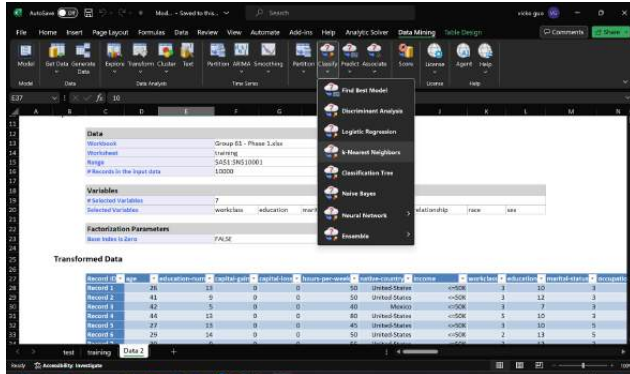


Fig.5.0. K Nearest Neighbor Classifier Selection

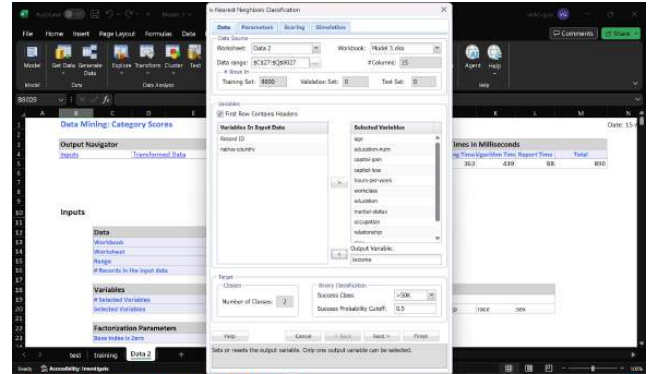


Fig.5.1. Data Tab. Selected Variable and the Output Variable Selection and Data range selection from C27 to C8027

Moving to the Parameters Tab, the K Nearest Neighbor algorithm was configured with a value of K=3, indicating that the three nearest neighbors were considered for classification. The nearest neighbor search method was set to "Fixed K.". After that, press the prior probability button, and the Prior Probability Tab will appear, set the Prior Probability Empirical. Then, press the next button again to move to the Scoring tab

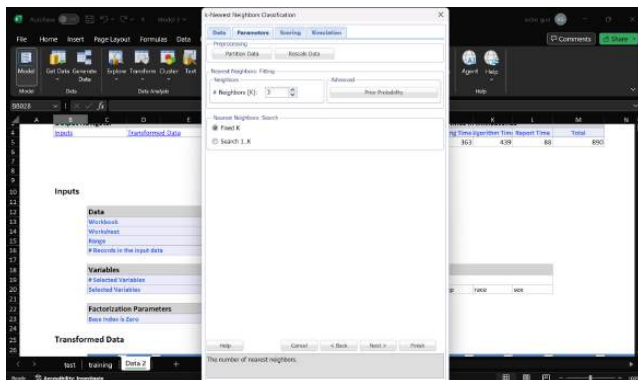


Fig.5.2. Parameter Tabs

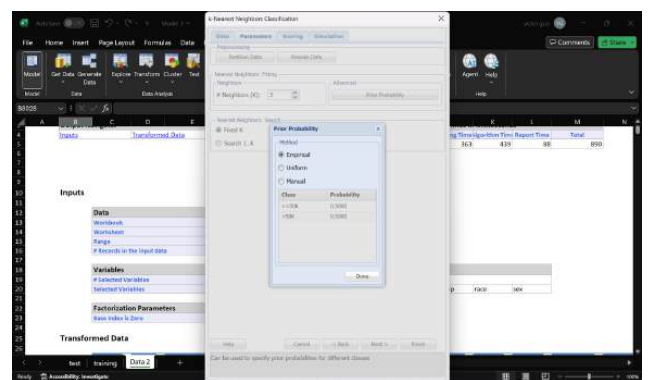


Fig.5.3. Prior Probability Tab

In the Scoring Tab, all the checkboxes in the Score Training data section were selected to evaluate the model's performance on the training set (as shown in Fig.5.4). Next, in the scoring new data section, enable the In Worksheet option, to indicate the test data was present in the same worksheet as the training set.

After enabling the In Worksheet option, a new tab w, the New Data (WS) Tab, will appear to set up the test data (as shown in Fig.5.5). The test data range was specified as C8029 to Q10029,

comprising 2000 data points. The variables in the test data were matched with the training set using the Match By Name function. Upon completion of the test data setup, the analysis was conducted. Finally, press the Finish button, and the result will be generated by XLMiner.

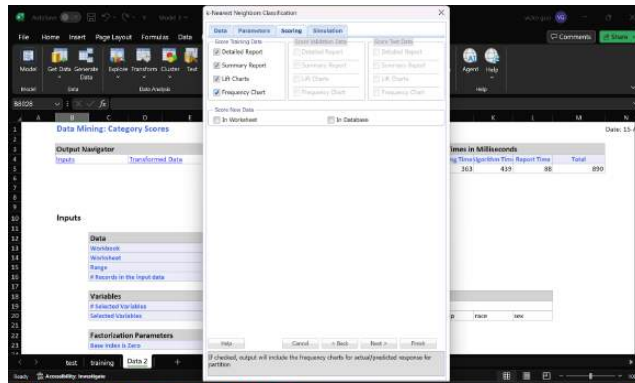


Fig.5.4. Scoring Tabs

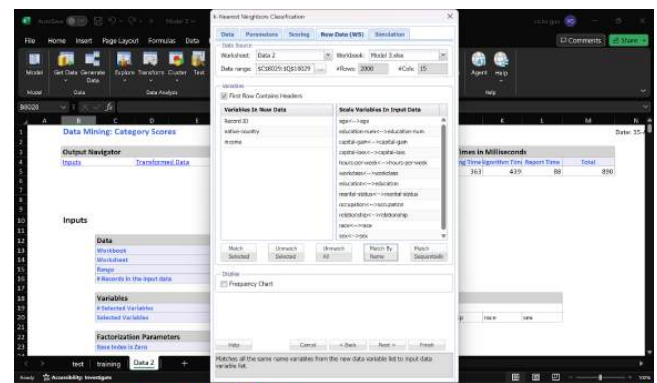


Fig.5.5. New Data (WS) Tab. Set the data range for the test set (C8029:C10029); Press “match by name” to match the training set attributes and test set attributes by name

## VI. Model 4

- ❖ Data: Data 2
- ❖ Type of Model: K Nearest Neighbor
- ❖ Parameter used:
  - Success Class: >50K
  - Number of Classes: 2
  - Success probability cut-off: 0.5
  - Fixed K with K = 5
  - Prior Probability to Empirical
  - Set the Test set and Training set variables to Match by Name

Similar with Model 3, in Model 4, we use the K Nearest Neighbor and Data 2 dataset with different parameters from Model 3. The parameter used for this Model is classifying the data into two classes based on the "Income" variable, with the success class defined as ">50K". and also a success probability cutoff of 0.5 to determine the class assignment. In addition, The K Nearest Neighbor algorithm was configured with a fixed value of K=5 and the prior probability estimation was performed using the empirical method.

In the Data 2 Sheet, select the Data mining tab on Excel, and then select Classify, and choose K Nearest Neighbor as the Algorithm type for this model (as shown in Fig 5.0). In the Data Tab, set the data range from C27 to C8029, since 8000 data points are for training sets. All variables except "Record ID," "native-country," and "Income" were selected as the Selected Variables. The "Income" variable was designated as the Output Variable (as shown in Fig 5.1). After that, press the Next button to go to the Parameter Tab



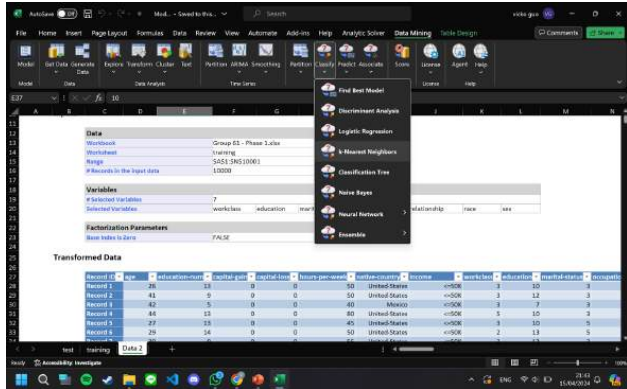


Fig.6.0. K Nearest Neighbor Classifier Selection

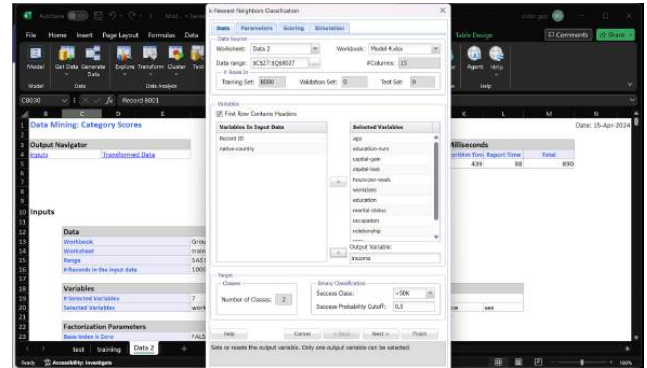


Fig.6.1. Data Tab. Selected Variable and the Output Variable  
Selection and Data range selection from C27 to C8027

Moving to the Parameters Tab, the K Nearest Neighbor algorithm was configured with a value of K=5, indicating that the three nearest neighbors were considered for classification. The nearest neighbor search method was set to "Fixed K.". After that, press the prior probability button, and the Prior Probability Tab will appear, set the Prior Probability Empirical. Then, press the next button again to move to the Scoring tab

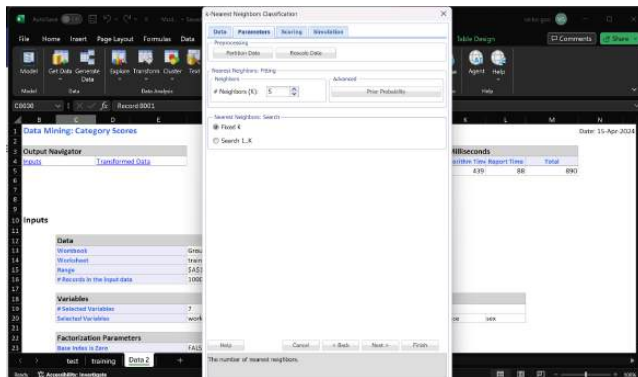


Fig.6.2. Parameter Tabs

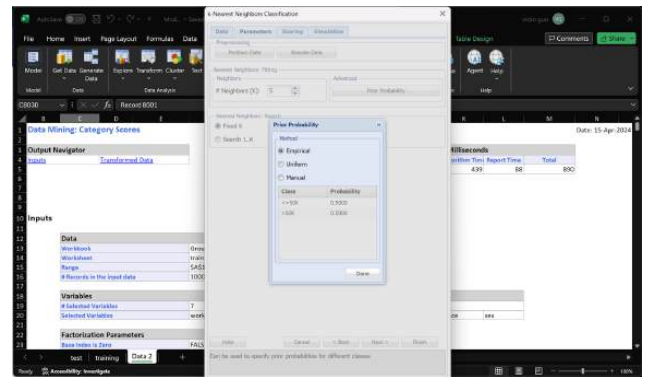


Fig.6.3. Prior Probability Tab

In the Scoring Tab, all the checkboxes in the Score Training data section were selected to evaluate the model's performance on the training set (as shown in Fig.5.4). Next, in the scoring new data section, enable the In Worksheet option, to indicate the test data was present in the same worksheet as the training set.



After enabling the In Worksheet option, a new tab w, the New Data (WS) Tab, will appear to set up the test data (as shown in Fig.5.5). The test data range was specified as C8029 to Q10029, comprising 2000 data points. The variables in the test data were matched with the training set using the Match By Name function. Upon completion of the test data setup, the analysis was conducted. Finally, press the Finish button, and the result will be generated by XLMiner.

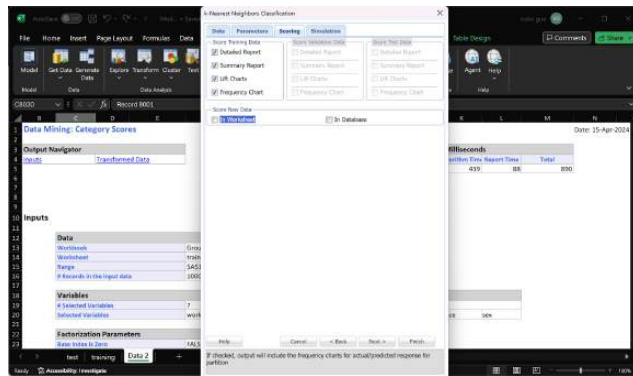


Fig.6.4. Scoring Tabs

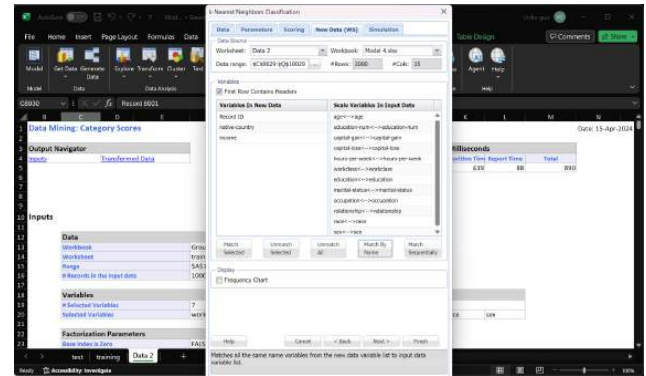


Fig.6.5. New Data (WS) Tab. Set the data range for the test set (C8029:C10029); Press “match by name” to match the training set attributes and test set attributes by name

## VII. Model 5

- ❖ Data: Data 2
- ❖ Type of Model: K Nearest Neighbor
- ❖ Parameter used:
  - Success Class: >50K
  - Number of Classes: 2
  - Success probability cut-off: 0.5
  - Fixed K with K = 10
  - Prior Probability to Empirical
  - Set the Test set and Training set variables to Match by Name

For Model 5, we are using “Data 2” and the model is K nearest neighbor, similar to models 3 and 4. We will be changing the parameters to hopefully achieve more accurate results. The parameter used for this Model is classifying the data into two classes based on the "Income" variable, with the success class defined as ">50K". We open the sheet and select “Predict” then select “K Nearest Neighbour” (as shown on Fig. 7.0). The success probability cutoff is set to 0.5 to determine the class assignment. We now set the K Nearest Neighbor algorithm which is configured with a fixed value of K=10 and the prior probability estimation was performed using the empirical method.

In the Data 2 Sheet, select Data mining tab on Excel, and then select Classify, and choose K Nearest Neighbor as the Algorithm type for this model (as shown in Fig 7.1). In the Data Tab, set the data range from C27 to C8029, since 8000 data points are for training sets. All variables except "Record ID," "native-country," and "Income" were selected as the Selected Variables. The "Income" variable was designated as the Output Variable (as shown in Fig 7.1). After that, press the Next button to go to the Parameter Tab

Moving to the Parameters Tab, the K Nearest Neighbor algorithm was configured with a value of K=5, indicating that the three nearest neighbors were considered for classification. The nearest neighbor search method was set to "Fixed K."(as shown in Fig 7.2). After that, press the prior probability button, and the Prior Probability Tab will appear, set the Prior Probability Empirical(as shown in Fig 7.3). Then, press the next button again to move to the Scoring tab

In the Scoring Tab, all the checkboxes in the Score Training data section were selected to evaluate the model's performance on the training set (as shown in Fig.7.4). Next, in the scoring new

data section, enable the In Worksheet option, to indicate the test data was present in the same worksheet as the training set.

After enabling the In Worksheet option, a new tab w, the New Data (WS) Tab, will appear to set up the test data (as shown in Fig.7.5). The test data range was specified as C8029 to Q10029, comprising 2000 data points. The variables in the test data were matched with the training set using the Match By Name function. Upon completion of the test data setup, the analysis was conducted. Finally, press the Finish button, and the result will be generated by XLMiner.

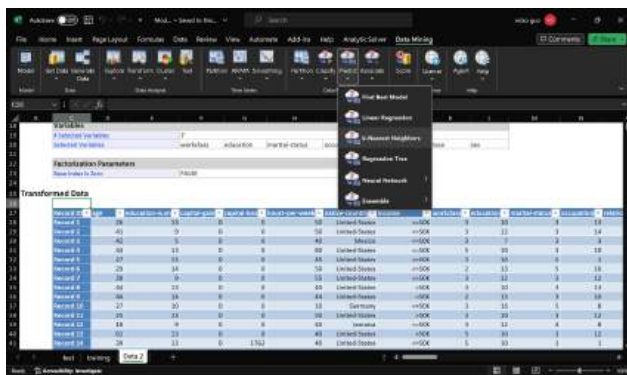


Fig.7.0. K Nearest Neighbor Classifier Selection

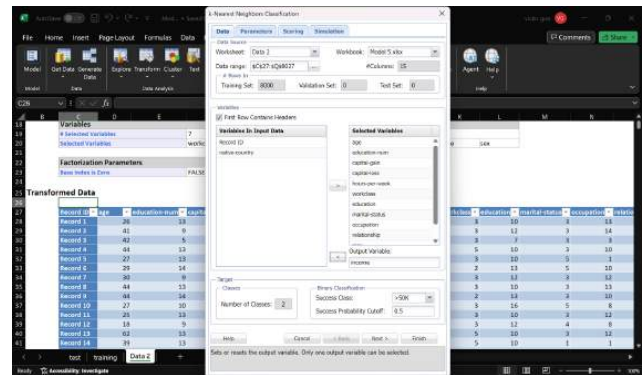


Fig.7.1. Data Tab. Selected Variable and the Output Variable Selection and Data range selection from C27 to C8027

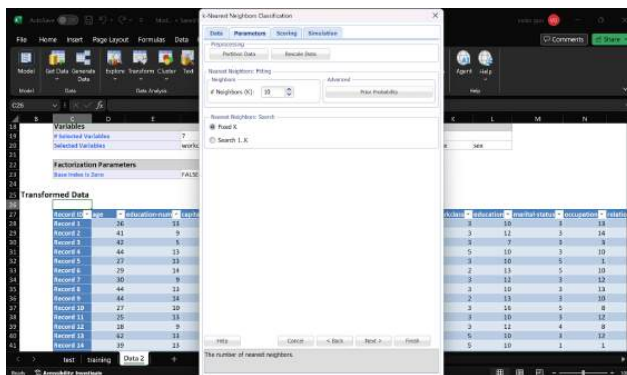


Fig.7.2. Parameter Tabs

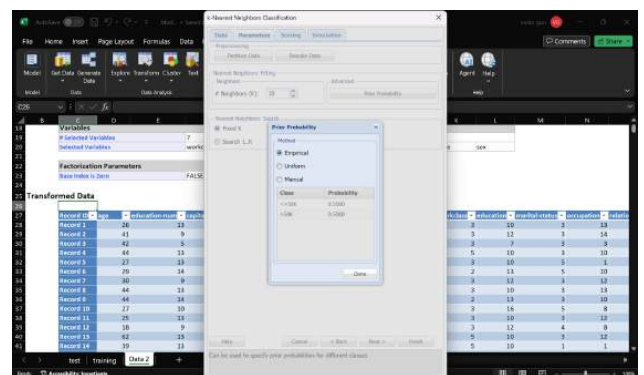


Fig.7.3. Prior Probability Tab

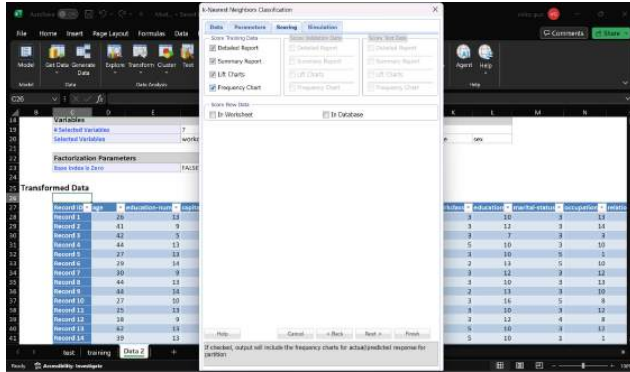


Fig.7.4. Scoring Tabs

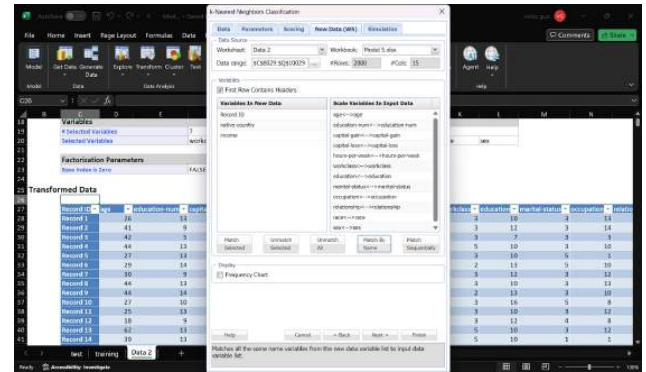


Fig.7.5. New Data (WS) Tab. Set the data range for the test set (C8029:C10029); Press “match by name” to match the training set attributes and test set attributes by name

## VIII. Images

The screenshot shows the Orange3 data mining software interface. The top menu bar includes File, Home, Insert, Page Layout, Formulas, Data, Review, View, Automate, Add-ins, Help, Analytic Solver, and Data Mining. Below the menu is a toolbar with icons for various data processing steps. The main workspace displays a data table with 14 columns: age, workclass, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country. The table contains 28 rows of data. The bottom status bar shows the current project is 'test1' and the active data table is 'Data1'.

	A	B	C	D	E	F	G	H	I	J	K	L	M
	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
1	26	Private	Bachelors	13	Married-civ-spouse	Tech support	Husband	White	Male	0	0	50	United-States
2	41	Private	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	50	United-States
3	42	Private	9th	5	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	Mexico
4	44	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	80	United-States
5	27	Private	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	45	United-States
6	29	Local-gov	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	0	0	50	United-States
7	30	Private	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	55	United-States
8	44	Private	Bachelors	13	Married-civ-spouse	Tech support	Husband	White	Male	0	0	40	United-States
9	44	Local-gov	Masters	14	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	44	United-States
10	27	Private	Some-college	10	Never-married	Other-service	Own-child	White	Male	0	0	30	Germany
11	25	Private	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States
12	38	Private	HS-grad	9	Married-spouse-absent	Other-service	Own-child	Black	Male	0	0	40	Jamaica
13	62	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	40	United-States
14	34	Self-emp-not-inc	Bachelors	13	Divorced	Adm-clerical	Not-in-family	White	Male	0	1762	40	United-States
15	27	Self-emp-inc	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States
16	40	Local-gov	HS-grad	9	Married-civ-spouse	Wrtite	Black	Female	0	0	0	75	United-States
17	35	Local-gov	Masters	14	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	40	United-States
18	26	Private	HS-grad	9	Never-married	Adm-clerical	Unmarried	Black	Female	0	0	40	Jamaica
19	29	Private	Bachelors	13	Married-civ-spouse	Other-service	Husband	White	Male	7298	0	42	United-States
20	35	State-gov	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	50	United-States
21	26	Self-emp-inc	Bachelors	13	Never-married	Exec-managerial	Not-in-family	White	Male	0	0	50	United-States
22	37	Private	HS-grad	9	Divorced	Sales	Unmarried	White	Female	0	0	21	Canada
23	24	Private	11th	7	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	30	Germany

*Fig 1.1*

The screenshot displays the Microsoft Excel interface. At the top, the ribbon shows the 'Formulas' tab. The main workspace contains a data table with columns A through M. A right-click context menu is open over cell A803, showing options like 'Cut', 'Copy', 'Paste Options...', 'Paste Special...', 'Insert', 'Delete', 'Clear Contents', 'Format Cells...', 'Wrap Text...', and 'Hide'. The data table includes columns for marital status, occupation, spouse status, gender, age, and country.

	A	B	C	D	E	F	G	H	I	J	K	L	M
7997	39	Private	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	50	United-States
7998	40	Private	Bachelors	13	Never-married	Prof-specialty	Not-in-family	White	Male	4650	0	40	United-States
8000	Calculated	11	HS-grad	9	Married-civ-spouse	Other-service	Husband	White	Male	0	0	40	Turkey
8001	Calculated	11	HS-grad	9	Married-AF-spouse	Sales	Husband	White	Male	0	0	55	United-States
8002	Calculated	11	HS-grad	13	Divorced	Sales	Not-in-family	White	Male	0	0	45	United-States
8003	Calculated	11	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	44	United-States
8004	Calculated	11	HS-grad	9	Never-married	Protective-serv	Not-in-family	White	Male	0	0	42	United-States
8005	Calculated	11	HS-grad	10	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	50	United-States
8006	Calculated	11	HS-grad	13	Never-married	Sales	Not-in-family	White	Male	0	0	55	United-States
8007	Calculated	11	HS-grad	9	Married-civ-spouse	Handlers-cleaners	Husband	White	Male	0	0	40	Mexico
8008	Calculated	11	HS-grad	14	Married-civ-spouse	Sales	Husband	White	Male	3104	0	50	United-States
8009	Calculated	11	HS-grad	9	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	1977	40	United-States
8010	Calculated	11	HS-grad	13	Never-married	Prof-specialty	Not-in-family	White	Female	0	0	30	Philippines
8011	Calculated	11	HS-grad	10	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States
8012	Calculated	11	HS-grad	10	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	7298	0	40	United-States
8013	Calculated	11	HS-grad	11	Never-married	Prof-specialty	Unmarried	White	Female	0	0	40	Germany
8014	Calculated	11	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	40	United-States
8015	Calculated	11	HS-grad	3	Married-civ-spouse	Wife	Wife	White	Female	0	0	40	El-Salvador
8016	Calculated	11	HS-grad	9	Divorced	Farming-fishing	Not-in-family	White	Male	0	0	40	United-States
8017	Calculated	11	HS-grad	8	Never-married	Sales	Not-in-family	White	Male	4650	0	50	United-States
8018	Calculated	11	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States
8019	Calculated	11	HS-grad	10	Divorced	Machine-op-inspct	Unmarried	White	Male	0	0	50	United-States
8020	Calculated	11	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	4179	0	40	United-States

*Fig 1.0*



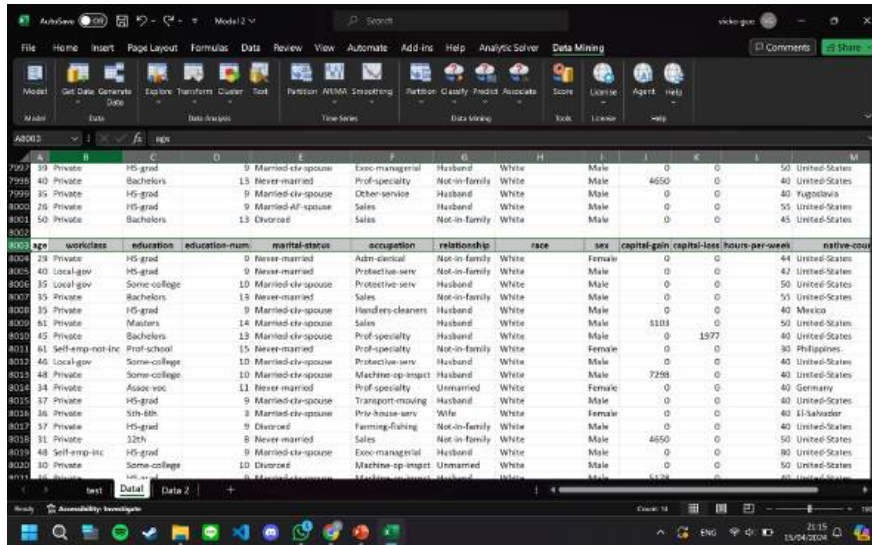


Fig. 1.2

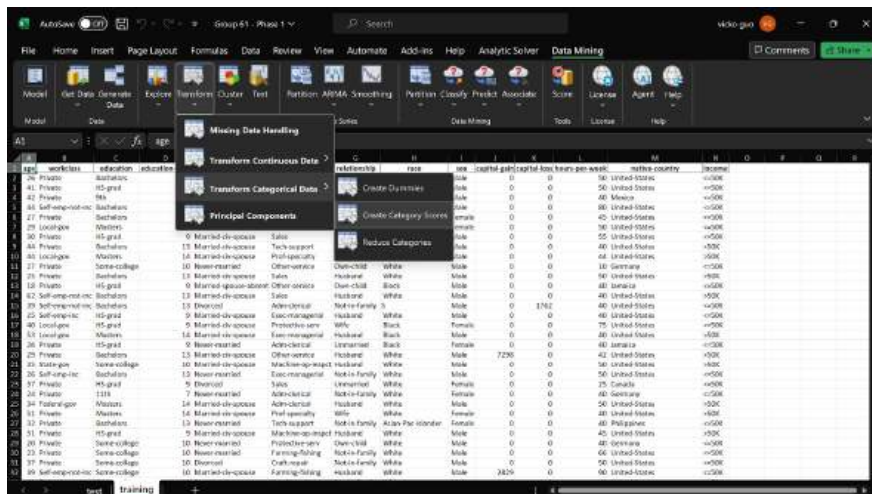


Fig. 2.0.0

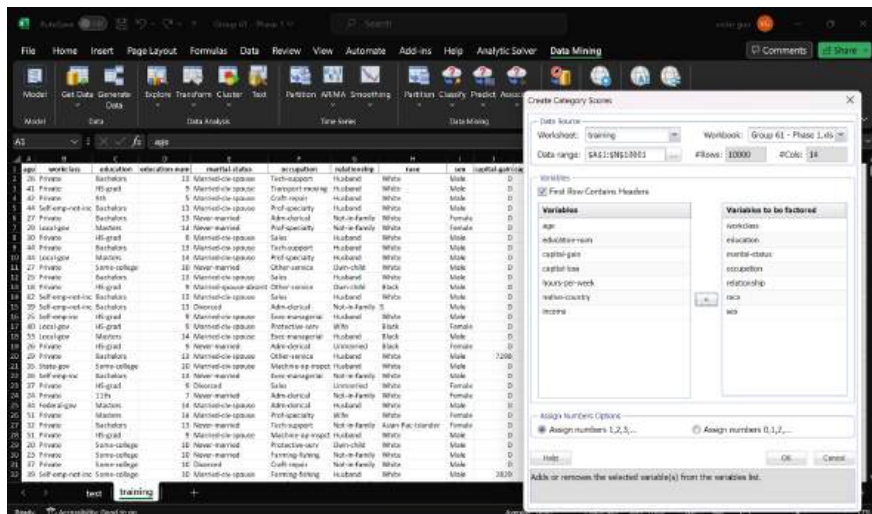
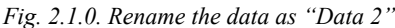


Fig. 2.0.1





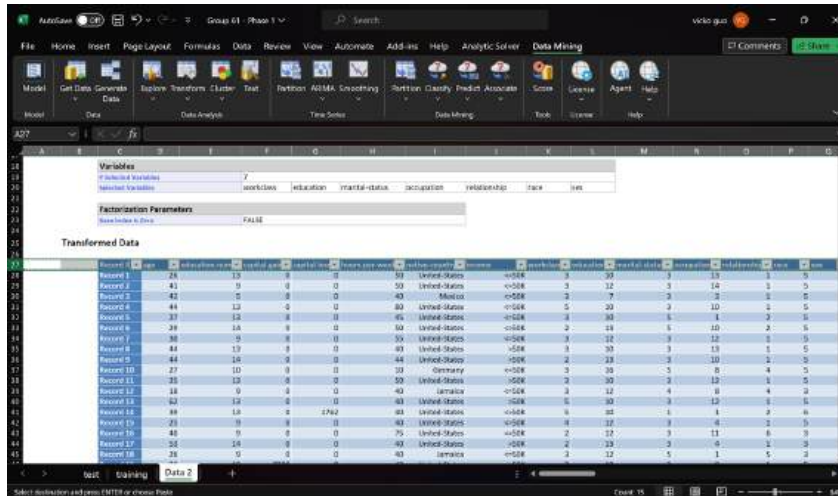


Fig. 2.1.2

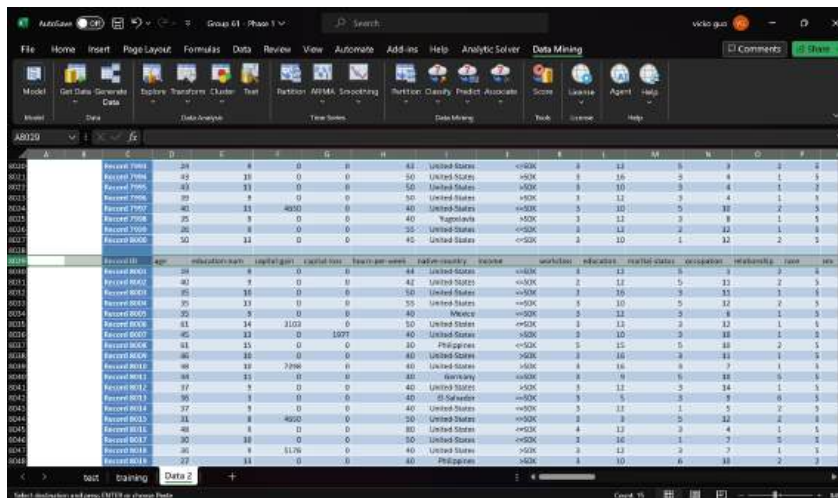


Fig 2.1.3

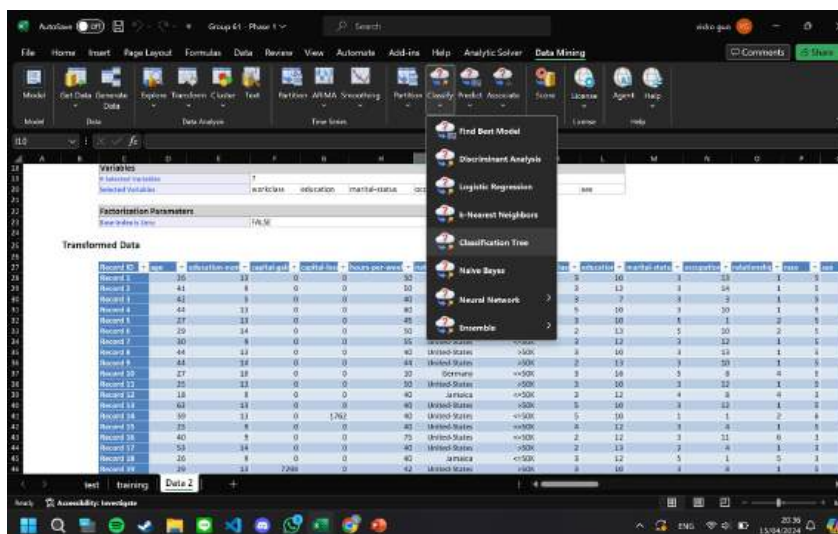


Fig. 3.0

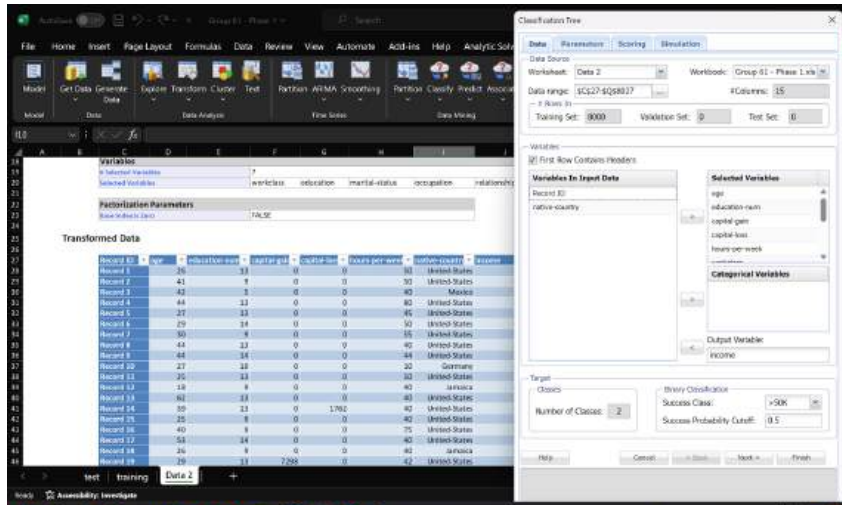


Fig. 3.1

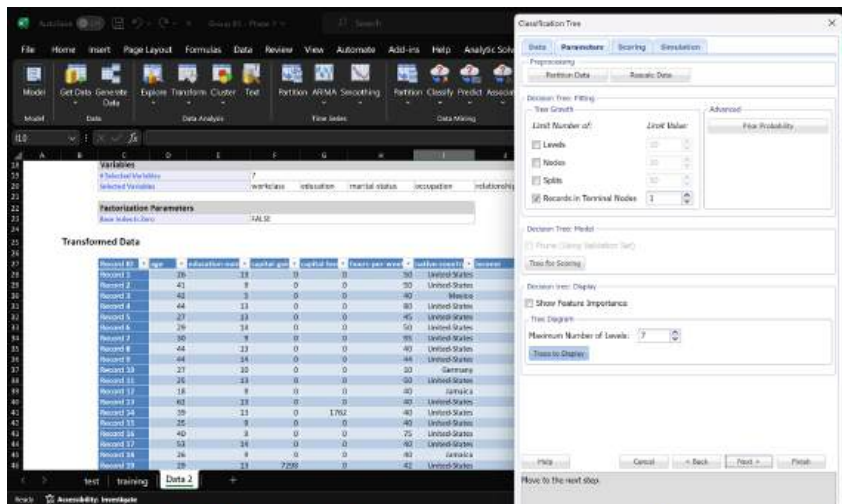


Fig. 3.2. Parameters Tab

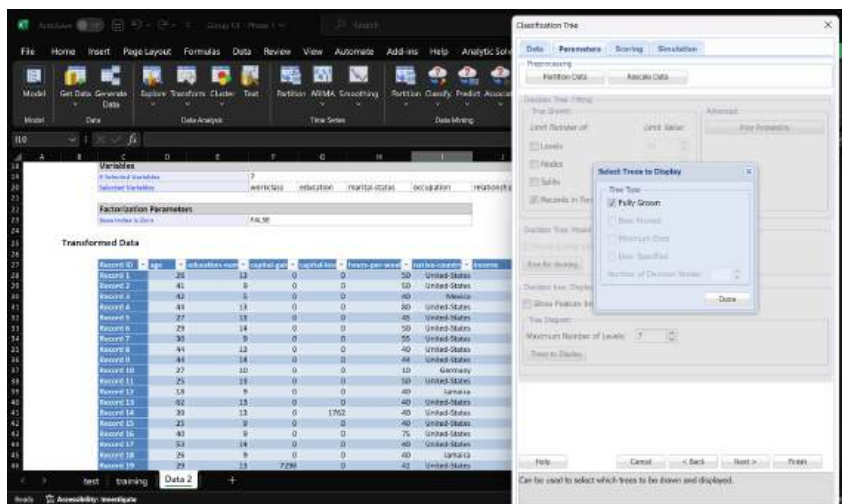


Fig. 3.3. Select Tree to Display type to Fully Grown

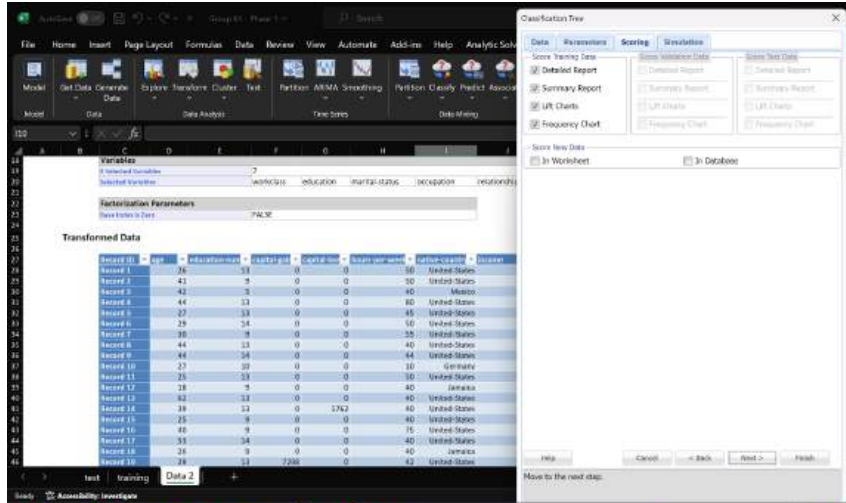


Fig. 3.4. Scoring tab

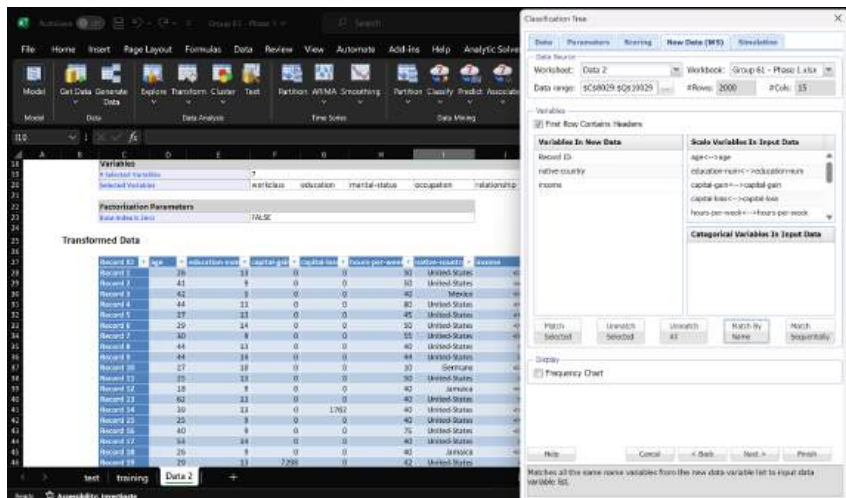


Fig. 3.5

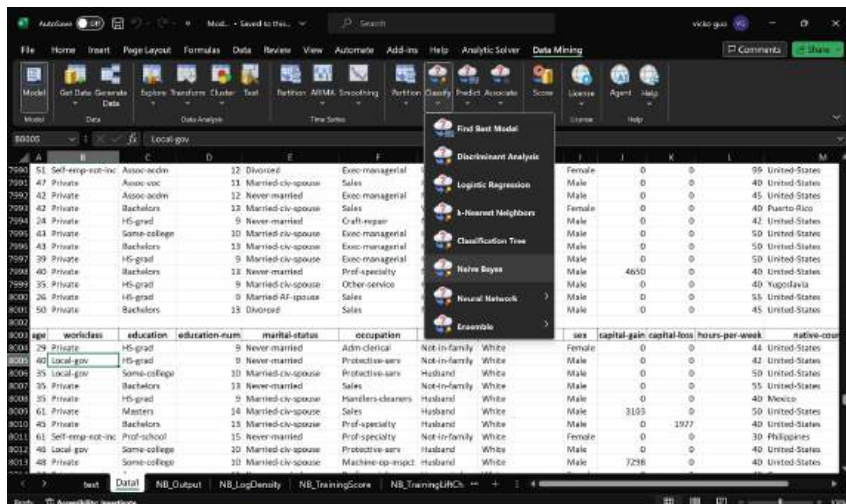


Fig. 4.0. Naive Bayesian Classifier Selection



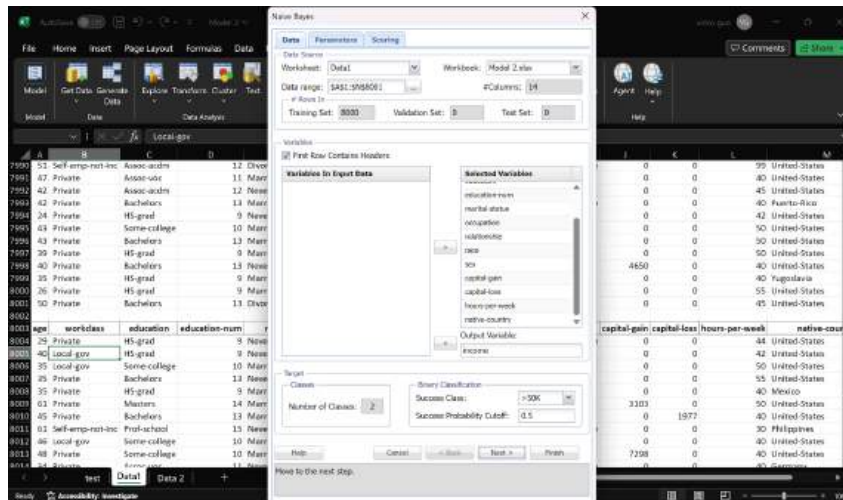


Fig. 4.1. Data Tab Selected Variable and the Output Variable

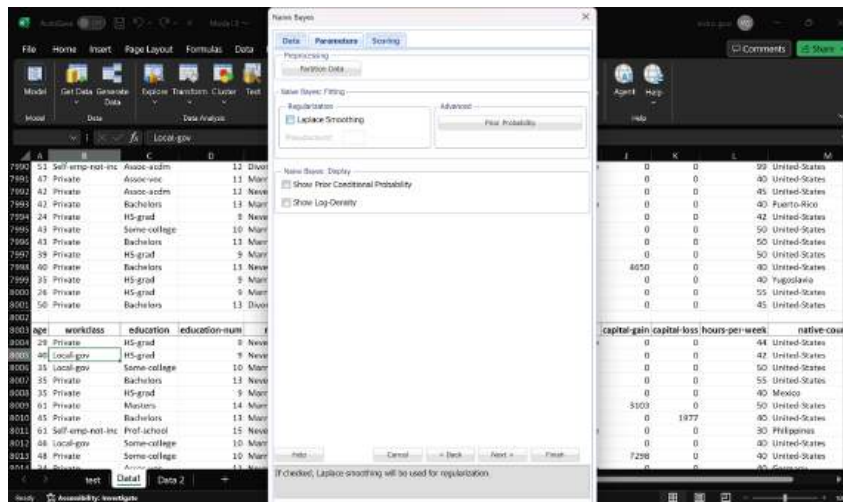


Fig. 4.2. Parameter Tabs

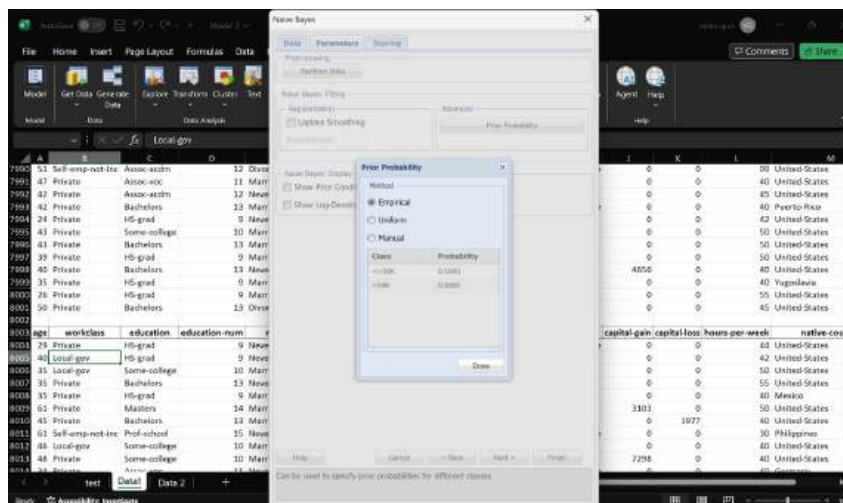


Fig. 4.2. Prior Probability Tab

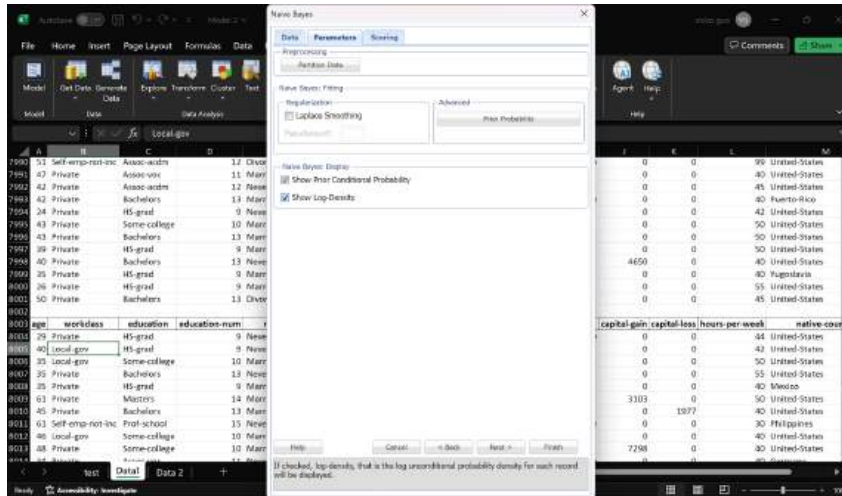


Fig. 4.4. Parameter Tabs

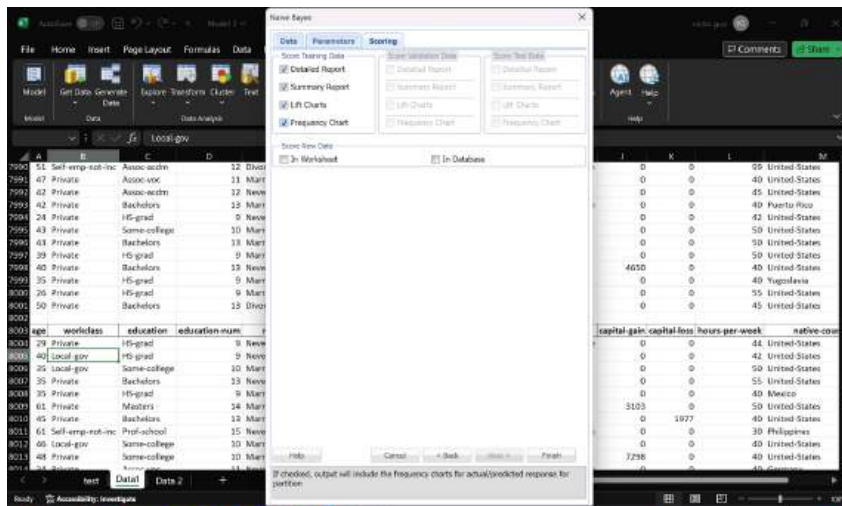


Fig. 4.5. Scoring Tab

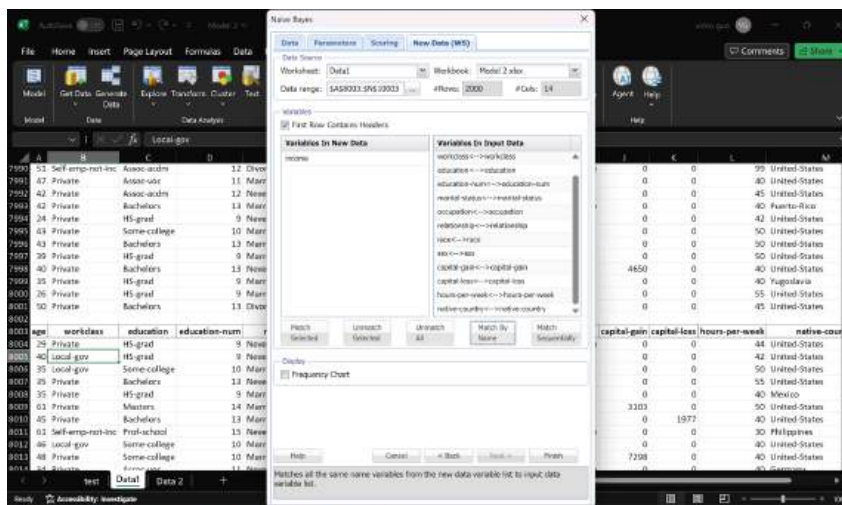


Fig. 4.6. New Data(WS) Tab. Set the data range for the test set

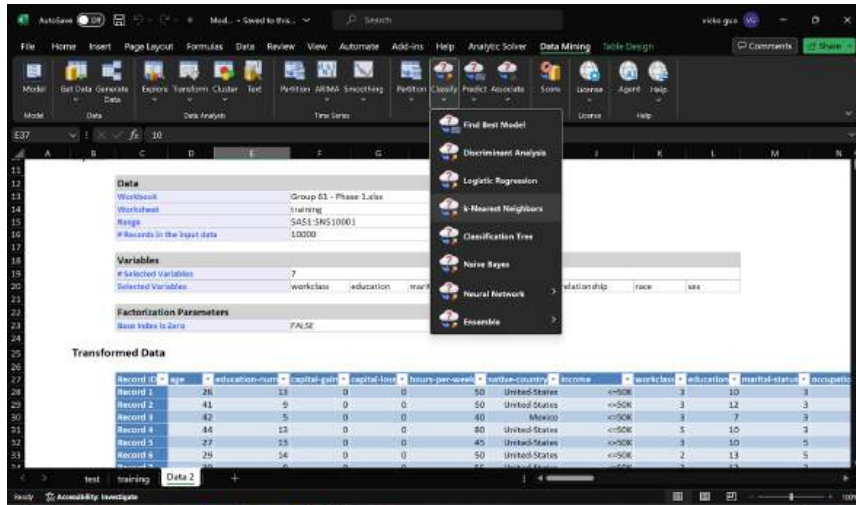


Fig. 5.0. K Nearest Neighbor Classifier Selection

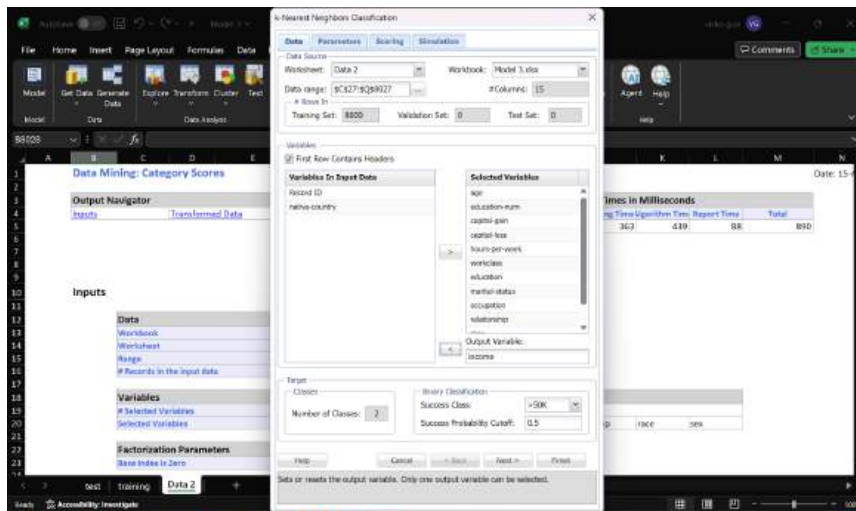


Fig. 5.1. Data Tab. Selected Variable and the Output Variable

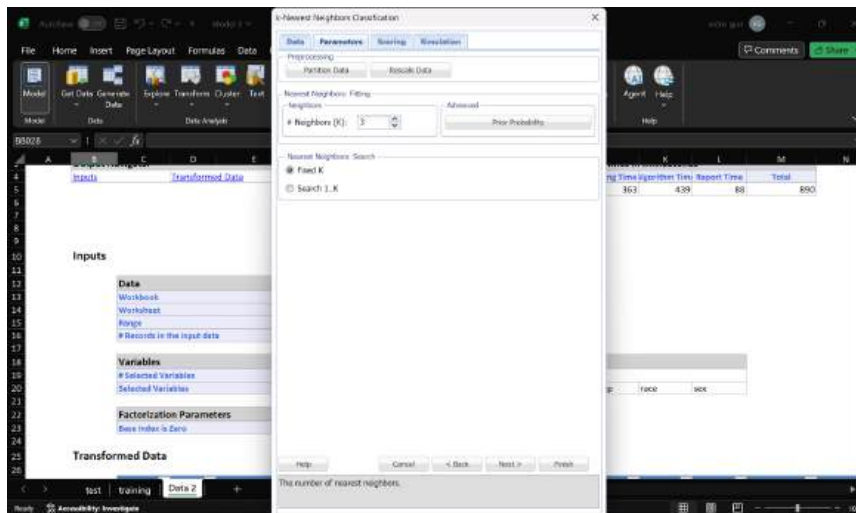


Fig. 5.2. Parameter Tabs

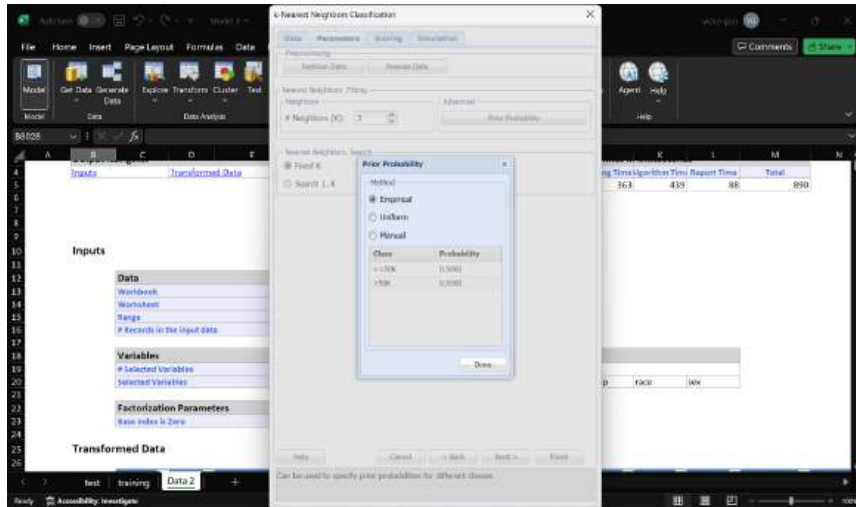


Fig. 5.3. Prior Probability Tab

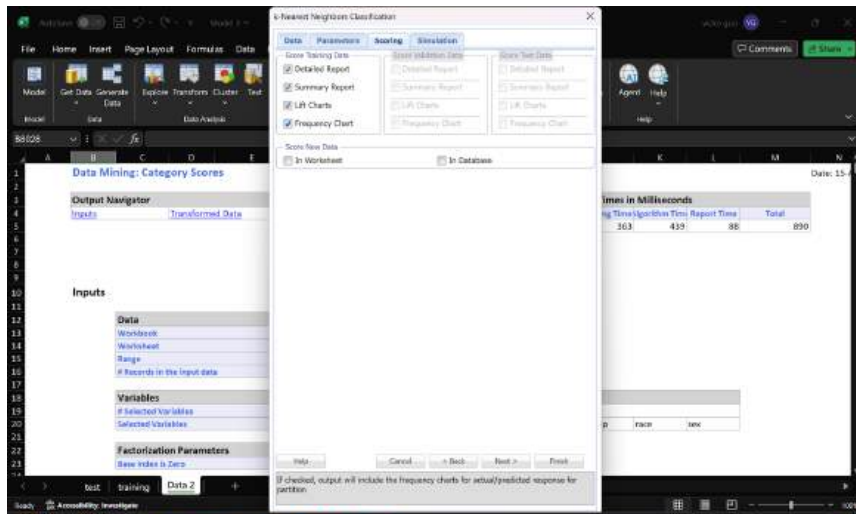


Fig. 5.4. Scoring Tabs

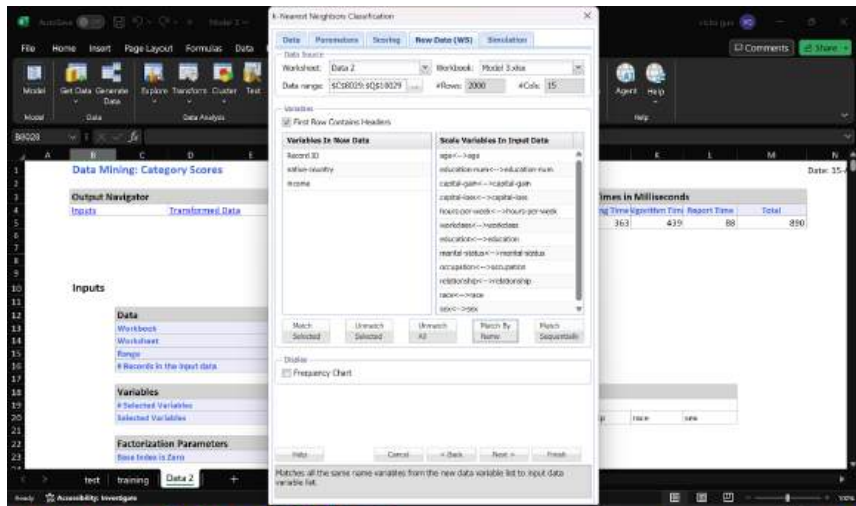


Fig. 5.5. New Data (WS) Tab.



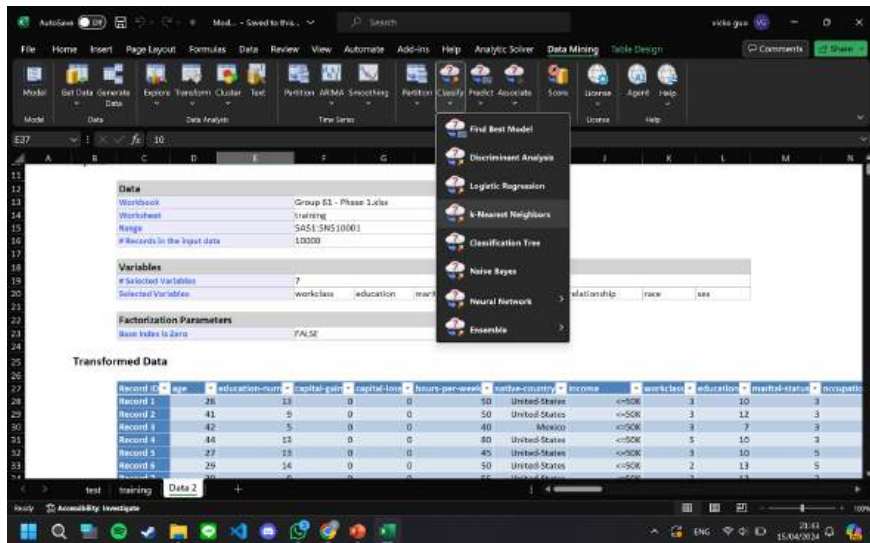


Fig. 6.0. K Nearest Neighbor Classifier Selection

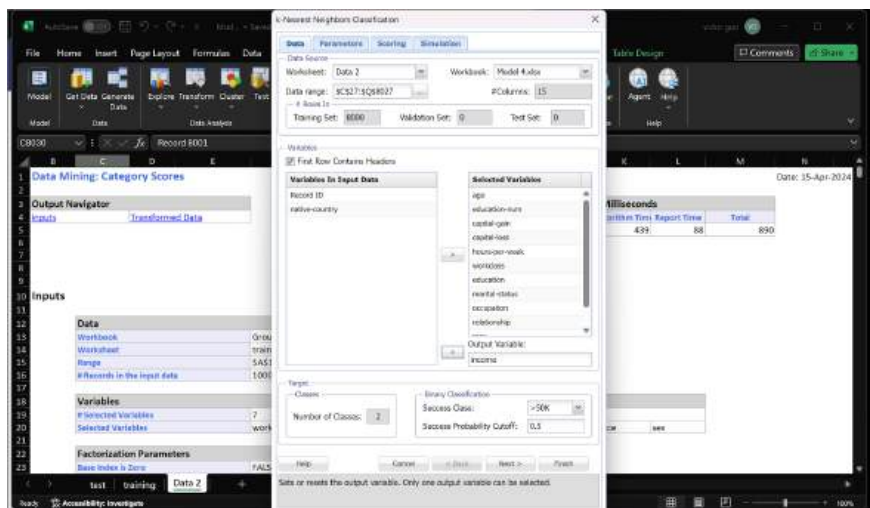


Fig. 6.1. Data Tab

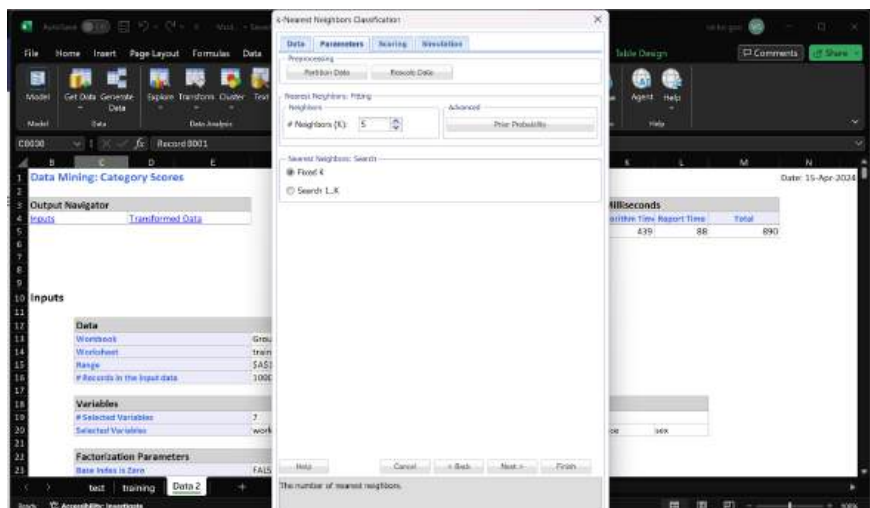


Fig. 6.2. Parameter Tabs

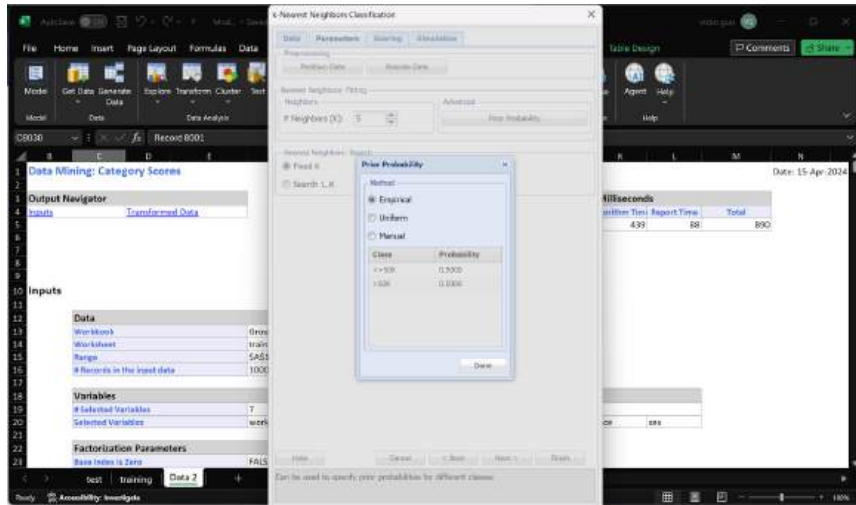


Fig. 6.3. Prior Probability Tab



Fig. 6.4. Scoring Tabs

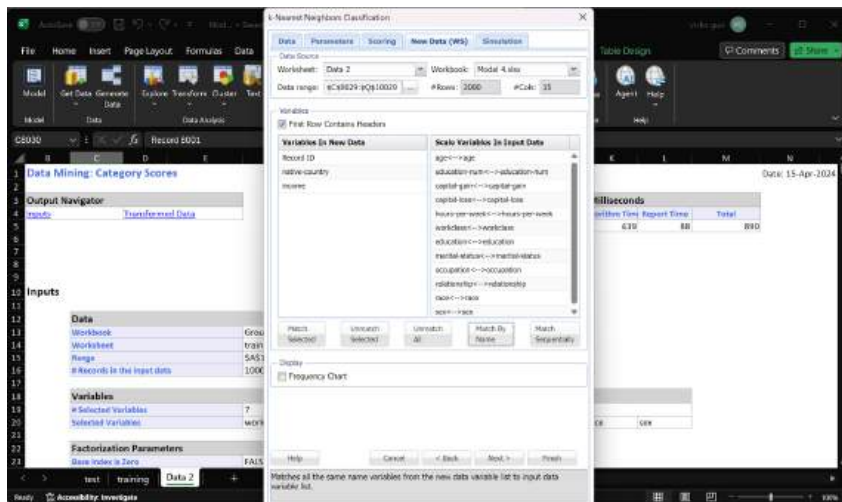


Fig. 6.5. New Data (WS) Tab

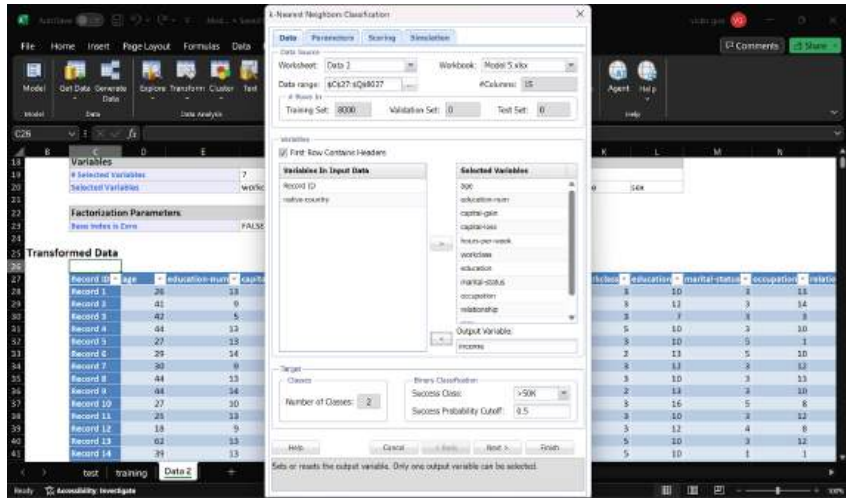


Fig. 7.1

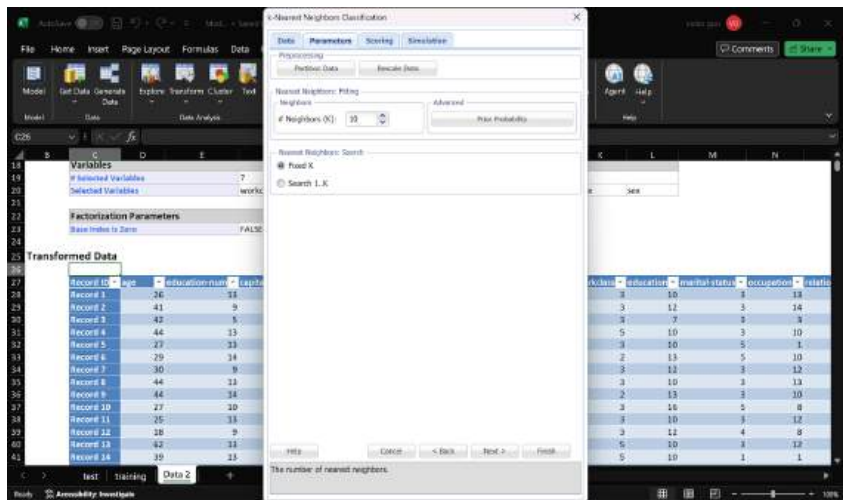


Fig. 7.2

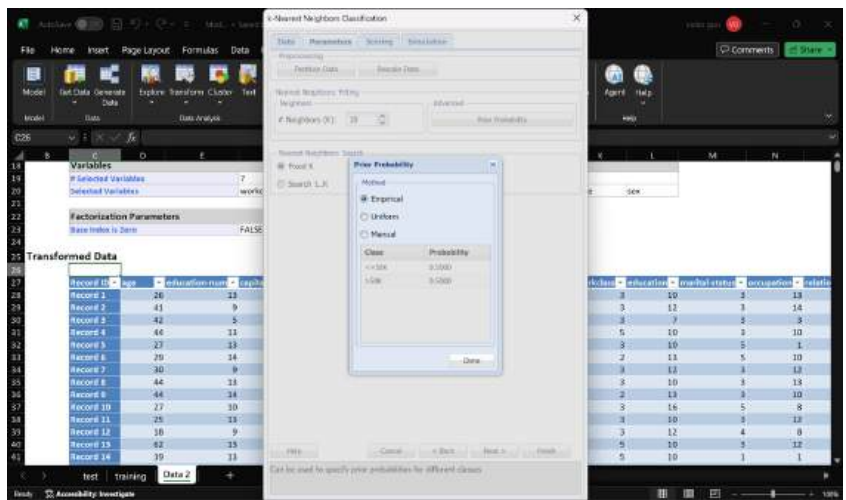


Fig. 7.3

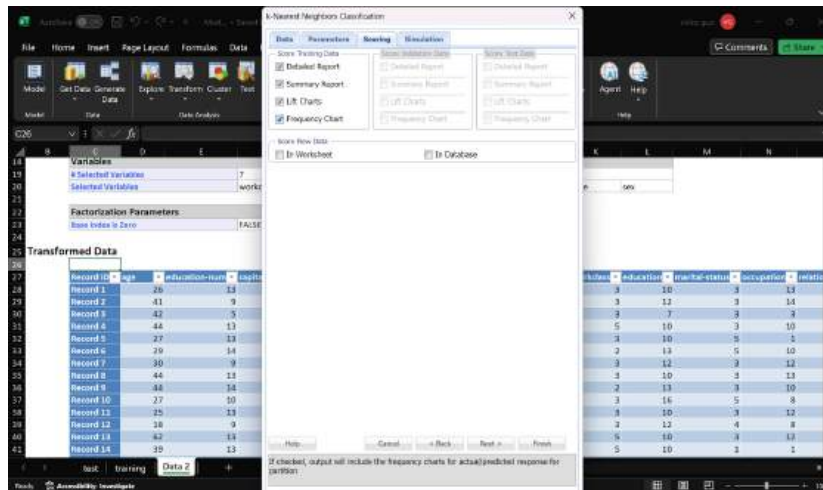


Fig. 7.4

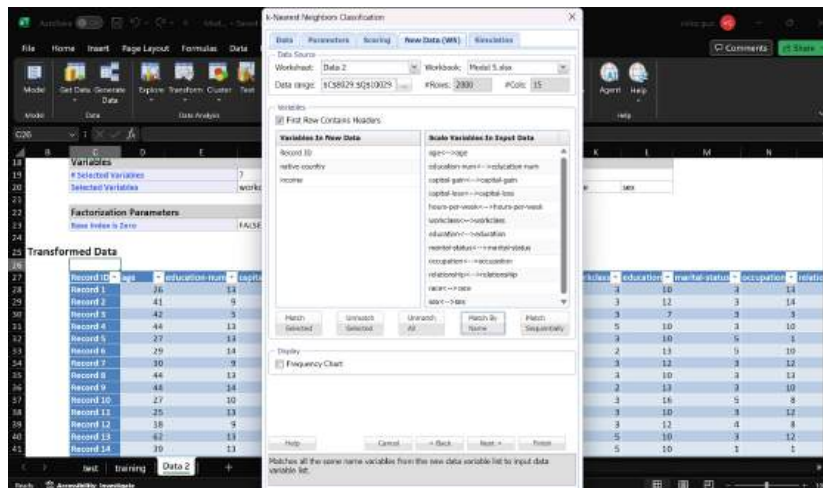


Fig. 7.5