DV User Manual

Table of Contents

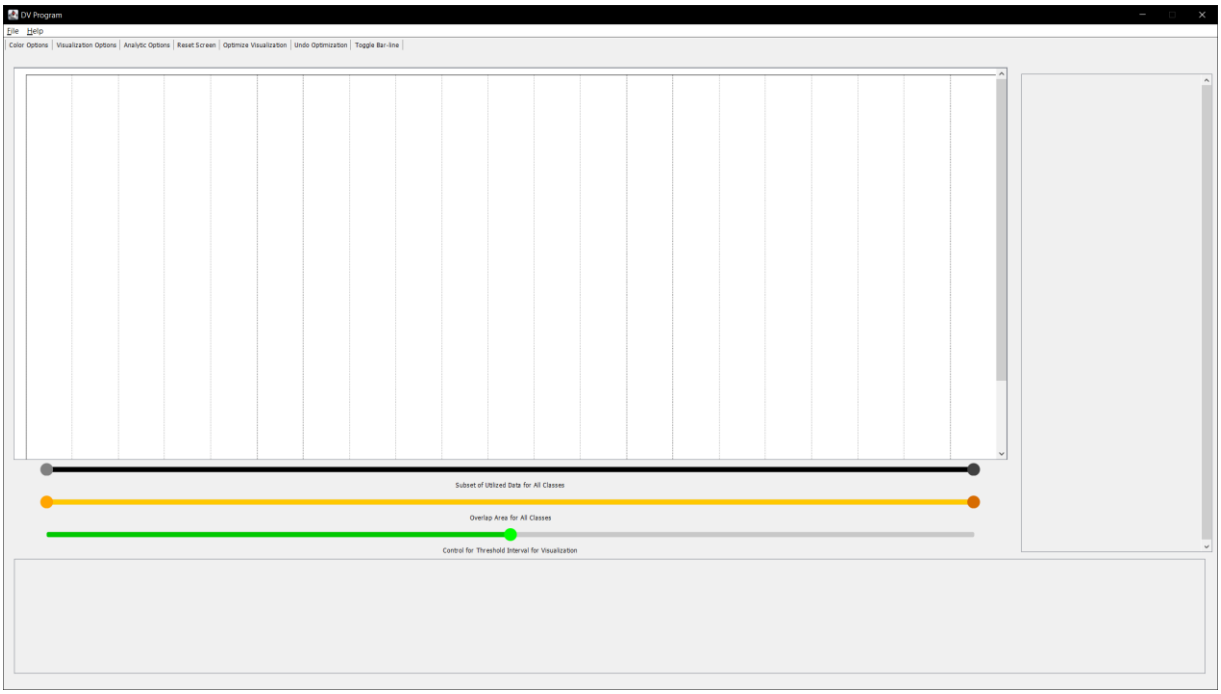
By: Daniel Van Houten, Morgan Leblanc, Tyler Swan, Fawziah Alkharnda, Stephan Adams, and Lincoln Huber

Introduction

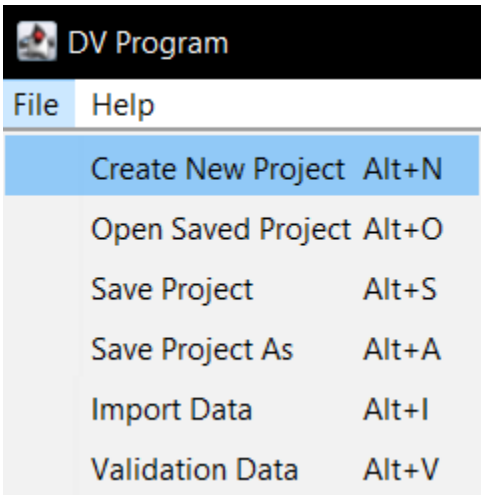
The purpose of the DV program is to render a visualization of data contained in one or more .csv files to improve data analysis. The graph is composed of many lines which represent a single data object or row of the dataset. Each line is composed of multiple vectors associated with the field's value or the columns of the data set. This allows a user to visualize data with multiple dimensions through the methodology of the program. In addition, the user will also be able to interact with the visualization through various program functions, such as domain constraints and field angle manipulation. Using the tools of the DV program a user can view data with many different dimensions, make classifications of data objects, or draw other conclusions from the data.

Getting Started

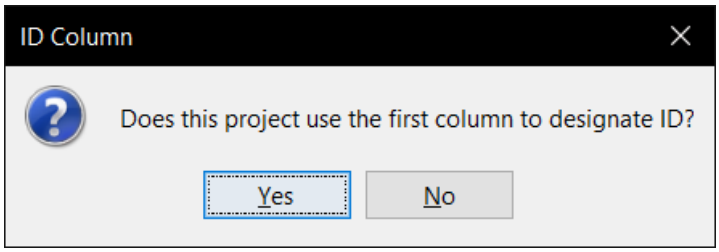
After the DV program has finished loading, it will display the following:



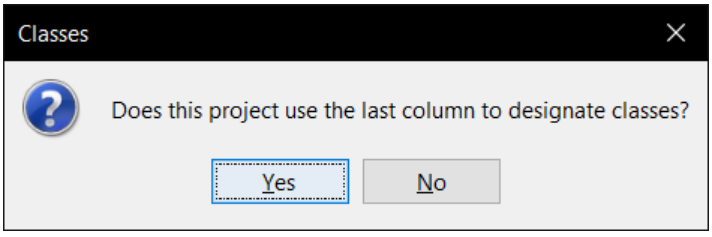
The initial screen is where a user will begin the process to create a new project or pick up from another project. To start a new project, click on the “File” drop-down menu or press Alt + N for a keyboard shortcut. Underneath the “File” drop-down menu, choosing “Create New Project” opens a new window, beginning the process to create a new project from comma delimited data (e.g. .csv file). Select the desired dataset using the file browser displayed.



Commonly, the first column of a dataset may represent the index of the row, thus the program will ask the user if the first column is used as the index upon the initial click of “Create New Project”.

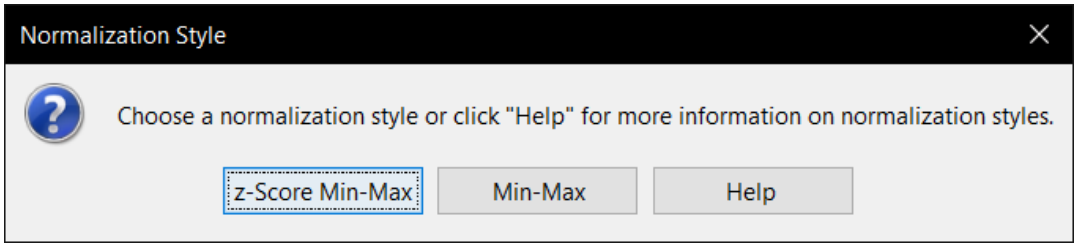


Similarly, a dataset may user the last column as the class indicator. Select the option yes if this is true, and no if not.

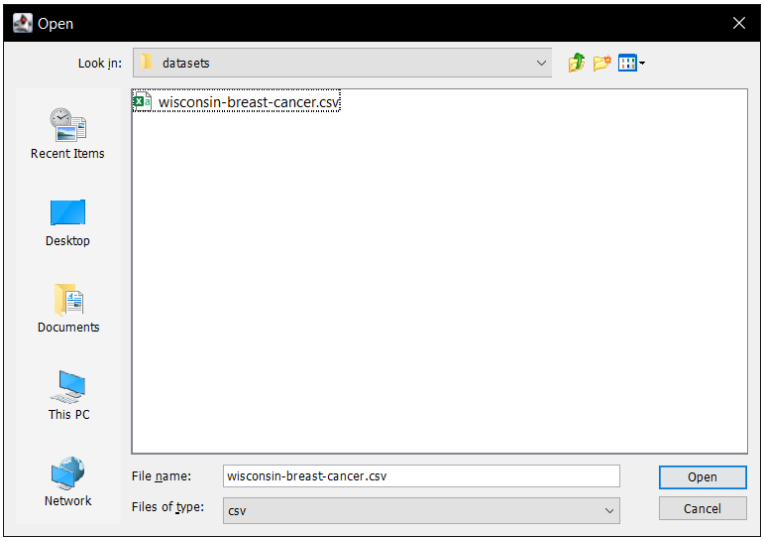


Then, the program will ask the user how to normalize the data. A z-Score Min-Max normalization will perform a z-Score standardization before normalizing the data from [0, 1]

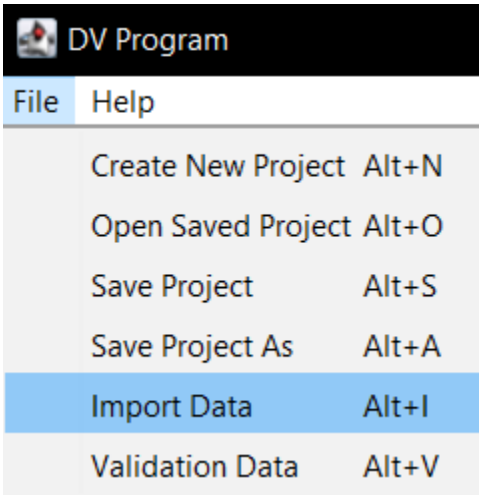
and a Min-Max normalization will directly normalize the data from [0, 1]. A user can also click "Help" for information of the exact mathematical formulas used for each option.



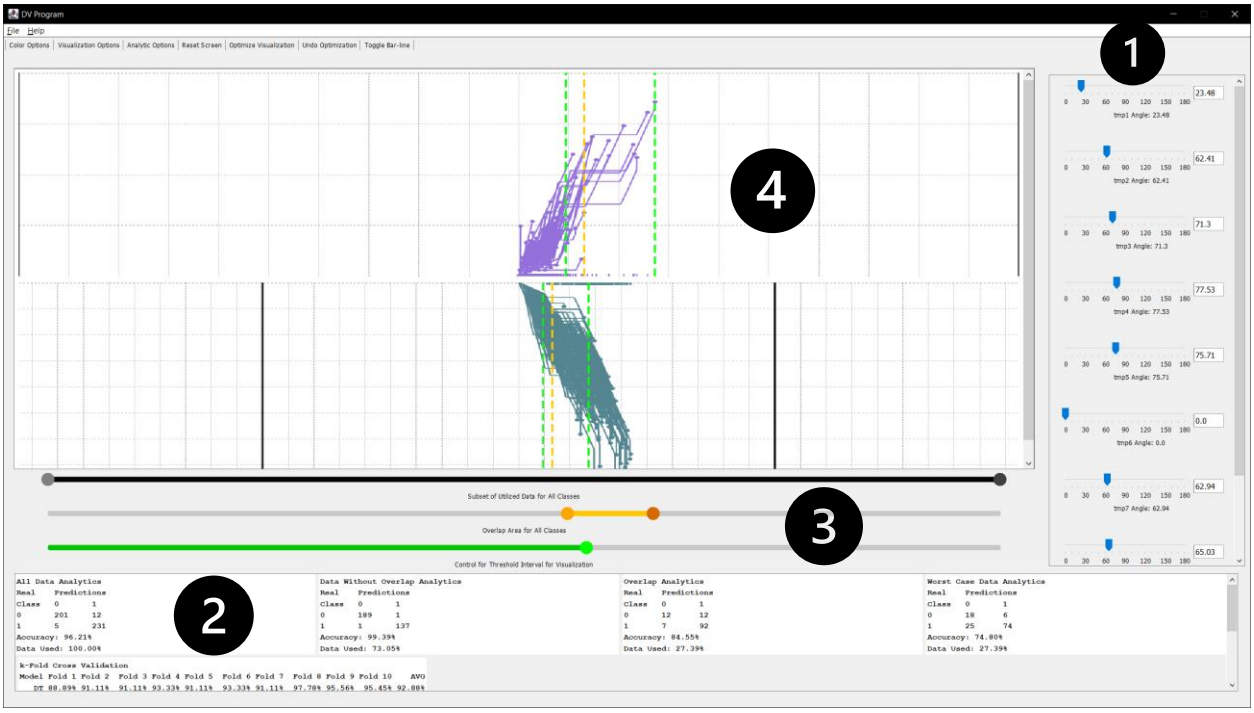
Last, a file browser will open, and a dataset can be chosen.



If a dataset's classes are separated by files, DV can still visualize all classes simultaneously. Again, the user will create a project in the same process as above, however they must then return to the "File" drop-down menu and click the option "Import Data." This option opens the file browser where the user can select another dataset.



The Below image is an example of what the program shall render once the user has finished the creation process. In this diagram are several important labels. The label (1) shows the location of the **field angles** for user interaction. Label (2) is where the **analytics** of the dataset will be displayed. Label (3) shows slider locations for **range**, **overlap**, and **threshold control**. Lastly, (4) indicates where the data will be visualized.



Manipulating Field Angles

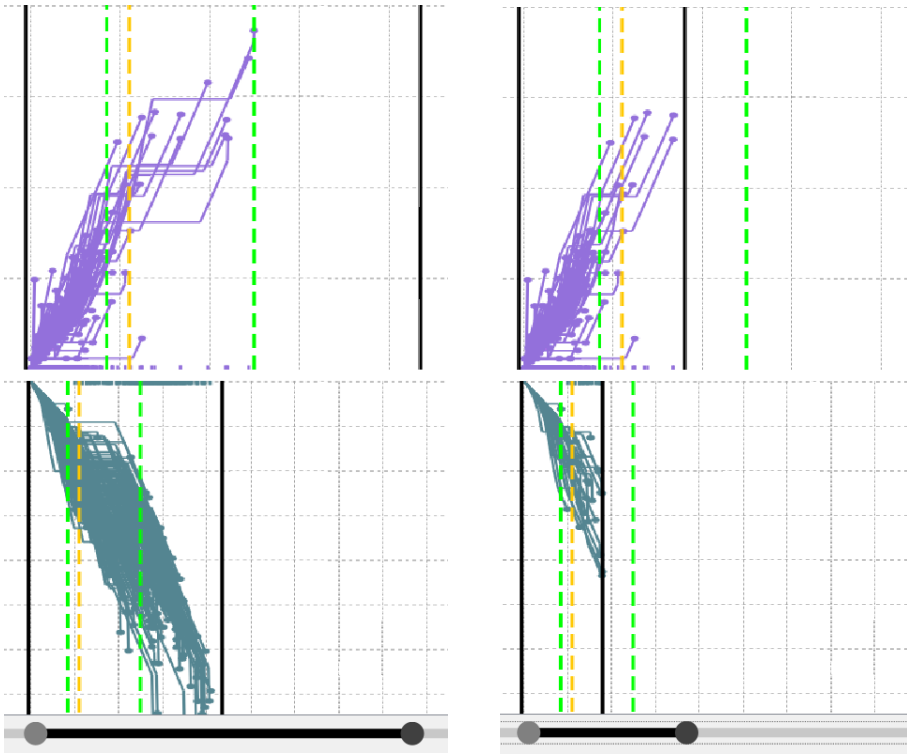
When rendering a visualization for a new project, the DV program will use linear discriminant analysis to find the optimal angles for the visualization, as shown in the figure above. The user can change these using components found in the field angles panel (denoted by (1)).

With these sliders, a user can manipulate the angle of the dimension which will be

rendered in real time as the user interacts. This can also be done by inputting a number to the textbox. It should also be noted that the title of the dimension is displayed in the field angle's label. For example, the name of this dimension would be "feature 1."

Applying Domain Constraints

The domain constraints can be utilized through the **Subset of Utilized Data for all Classes** slider shown by label (3). By dragging this slider, the user can apply domain constraints to the visualization. This will exclude data from the visualization which resides outside of the specified domain. The closer together the sliders are, the more data that will be constrained.



Generate Analytics

Below the domain slider, pointed to by label (3) the **Overlap area for all classes** and **Control for Threshold Interval for Visualization** sliders can be found. By default, the DV program will user linear discriminant analysis to find the optimal placement of the threshold and will take the leftmost and rightmost misclassified points to use as the range for the overlap area. At this point, five different analytics will appear: four confusion matrices and one k-fold cross validation. The confusion matrices will be for "All Data," "Data Without Overlap," "Overlap," and "Worst Case Data." The "All Data" confusion matrix is the confusion matrix for all datapoints shown in the visualization. The "Data Without Overlap" confusion matrix is for all datapoints shown in the visualization that are not within the overlap area. The "Overlap" confusion matrix is for all datapoints shown in the visualization that are within the overlap area. Both the "Data Without Overlap" and "Overlap" confusion matrix are generated after running linear discriminant analysis on the data to get the optimal separation. Then, the "Worst Case Data" confusion matrix is created by used the datapoints from the "Overlap" confusion matrix, but with the function used to get the "Data Without Overlap." As for the k-fold cross validation, by default there will be 10 folds. To change these analytics, the user can adjust the sliders to designate different predicted classifications for different intervals. The confusion matrices will reflect these changes by displaying updated accuracies for their classifications. Analytics for the visualization above can be seen on the next page.

```
All Data Analytics
Real     Predictions
Class    0        1
0        429      15
1        3        236
Accuracy: 97.36%
Data Used: 100.00%
```

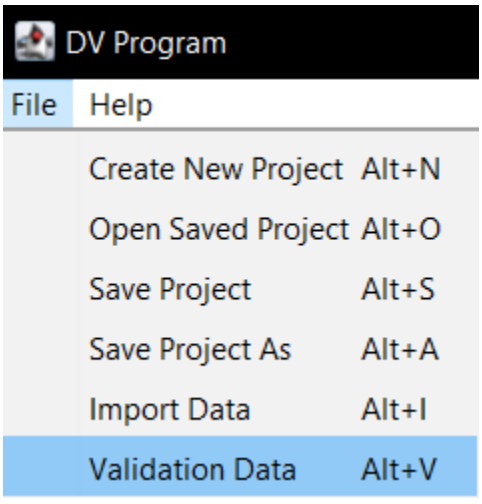
```
Data Without Overlap Analytics
Real     Predictions
Class    0        1
0        422      1
1        1        146
Accuracy: 99.65%
Data Used: 83.46%
```

```
Overlap Analytics
Real     Predictions
Class    0        1
0        9        13
1        6        87
Accuracy: 83.48%
Data Used: 16.84%
```

```
Worst Case Data Analytics
Real     Predictions
Class    0        1
0        15       7
1        27       66
Accuracy: 70.43%
Data Used: 16.84%
```

```
k-Fold Cross Validation
Model Fold 1 Fold 2  Fold 3 Fold 4 Fold 5  Fold 6 Fold 7  Fold 8 Fold 9 Fold 10    AVG
   DT 88.89% 82.22%  97.78% 91.11% 88.89%  88.89% 91.11%  95.56% 97.78%  93.18% 91.54%
  SGD 95.56% 95.56% 100.00% 91.11% 91.11%  93.33% 97.78%  95.56% 95.56% 100.00% 95.56%
   NB 91.11% 97.78%  95.56% 88.89% 91.11%  97.78% 95.56%  95.56% 97.78%  95.45% 94.66%
  SVM 88.89% 95.56%  97.78% 91.11% 91.11%  97.78% 95.56%  97.78% 97.78%  97.73% 95.11%
  KNN 86.67% 95.56% 100.00% 91.11% 91.11%  97.78% 95.56% 100.00% 97.78%  97.73% 95.33%
   LR 88.89% 91.11% 100.00% 91.11% 91.11%  97.78% 93.33%  97.78% 97.78% 100.00% 94.89%
  LDA 86.67% 86.67%  97.78% 93.33% 93.33%  97.78% 93.33%  97.78% 97.78% 100.00% 94.44%
  MLP 91.11% 93.33% 100.00% 91.11% 91.11%  97.78% 93.33%  97.78% 97.78%  97.73% 95.11%
   RF 91.11% 95.56% 100.00% 91.11% 88.89% 100.00% 95.56%  97.78% 97.78%  97.73% 95.55%
   SD  2.75%  5.09%   1.61%  1.11%  1.34%   3.35%  1.96%   1.48%  0.74%   2.27%  2.17%
  AVG 89.88% 92.59%  98.77% 91.11% 90.86%  96.54% 94.57%  97.28% 97.53%  97.73% 94.69%
```

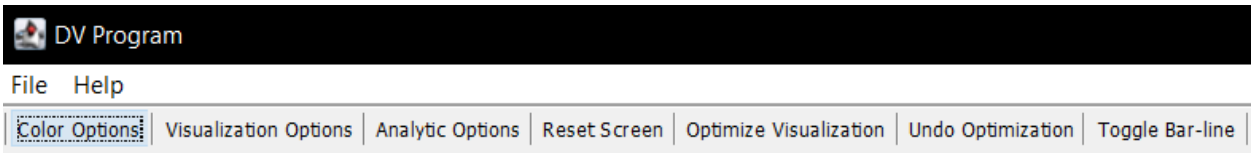
User Validation Data



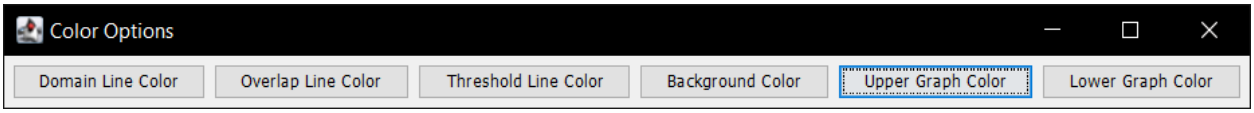
Along with the analytics mentioned in the previous section, users can also import their own validation data by going to the “Validation Data” option under the “File” drop-down menu. Clicking on this option will open the file browser where a user can search for the validation data. The validation data must be in a csv file and formatted similarly to previously entered data. Once the validation data has been successfully input, a user validation confusion matrix will be generated.

Color Options

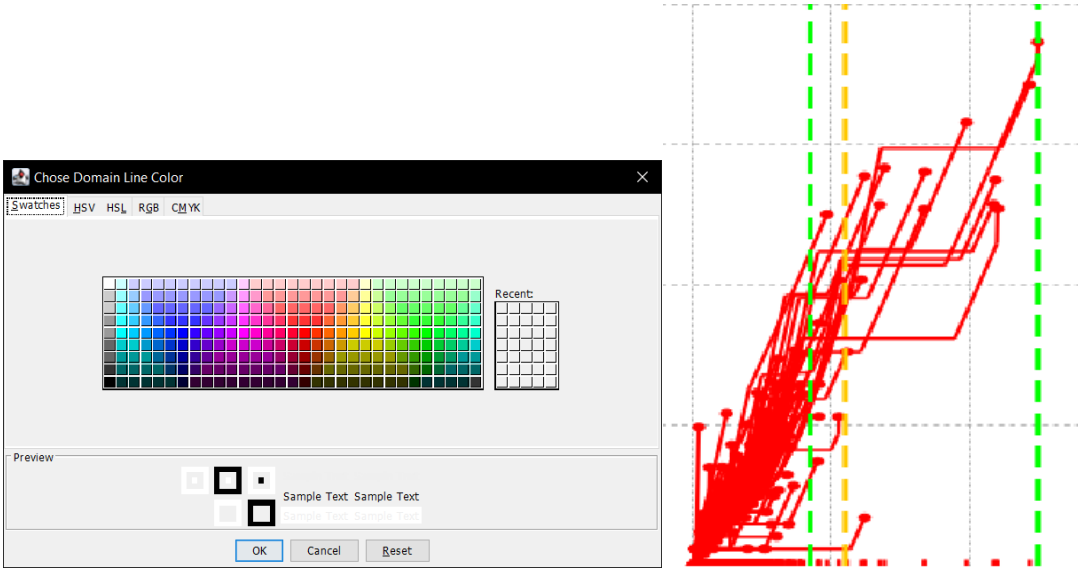
The user can find color options by navigating to “Color Options” on the toolbar.



These options allow the user to change the color of the domain lines, overlap lines, threshold line, graph background, upper graph, and lower graph.

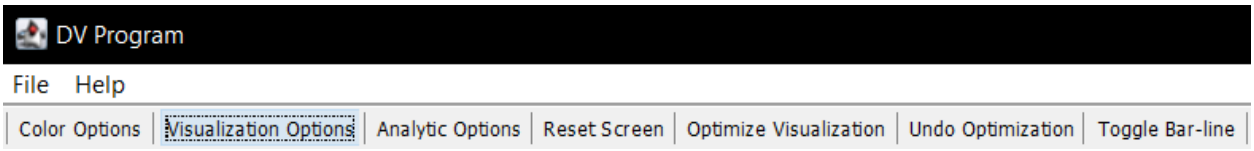


After choosing an option, a color choosing menu will appear. Choosing a color will change the color of the chosen option.



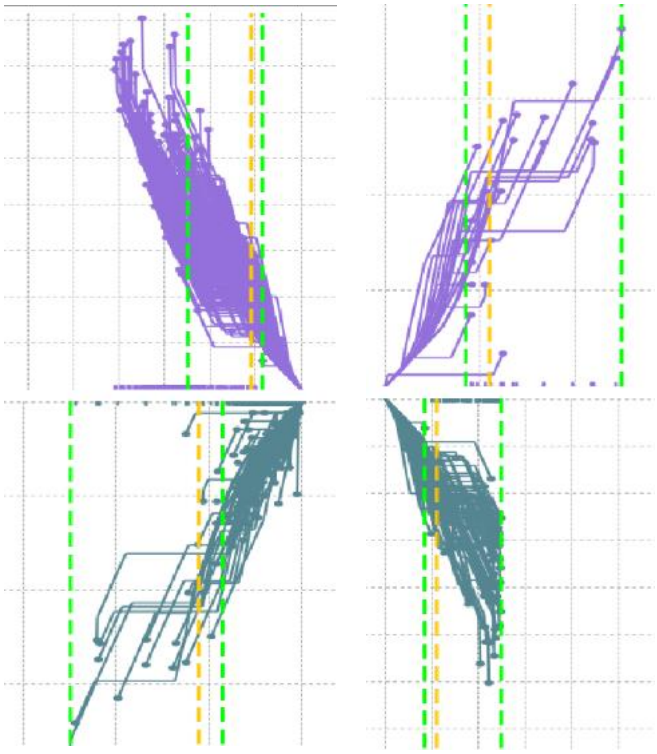
Visualization Options

The user can find visualization options by navigating to the “Visualization Options” on the toolbar.



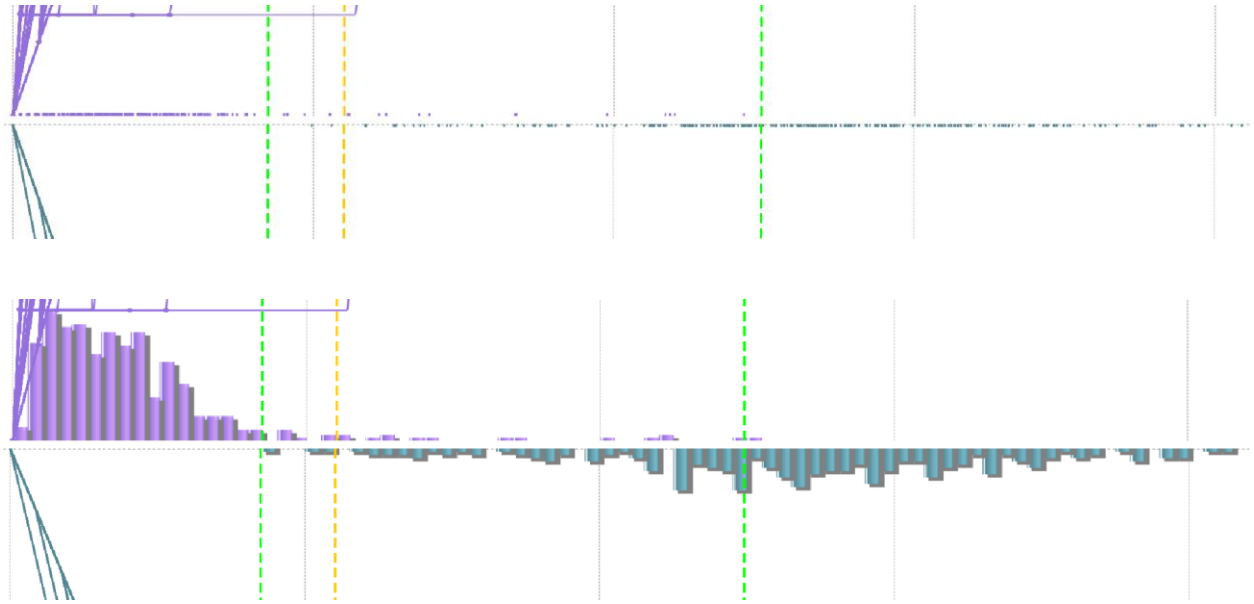
These options allow the user to change the visualization by changing which class is visualized on the upper graph, specifying the visualization in 3+ class visualization, and visualizing only the overlap.

On the right, the image on the left shows the upper class changed and the image on the right shows only the overlap visualized. Information on “Specify Visualization” can be found in the Section 3+ Class Visualizations.



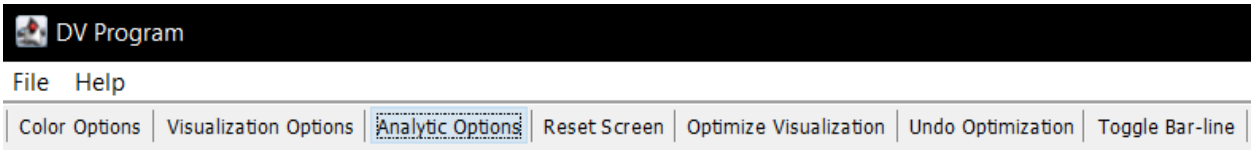
Bar-line

In addition to the visualization options mentioned in the previous section, users can also toggle on/off a bar-line of grouped frequency bars instead of individual values when data are packed into a small area. Below, the first image shows individual values, while the second image shows the bar-line of grouped frequency bars.

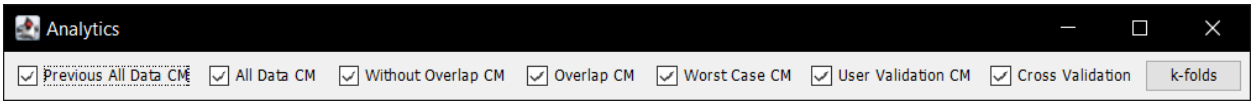


Analytic Options

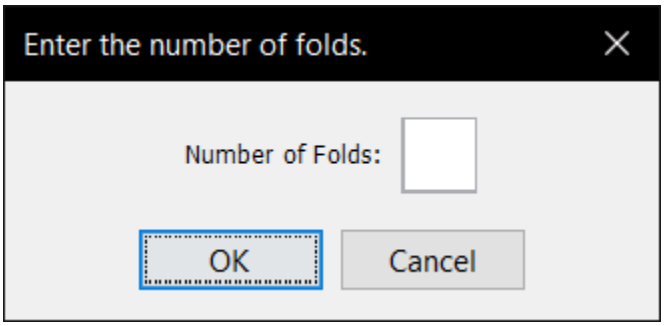
The user can find analytic options by navigating to the “Analytic Options on the toolbar.



These options allow the user to toggle on/off all analytic options as well as change the number of folds in k-fold validation.



After choosing on option, the selected analytic will be removed from the analytic panel, pointed to by label (2), or the following menu will appear if changing the number of folds in k-fold cross validation.

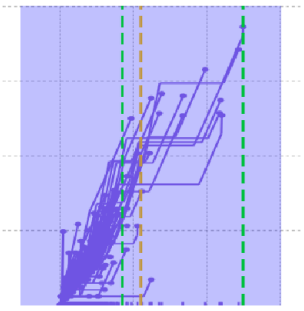


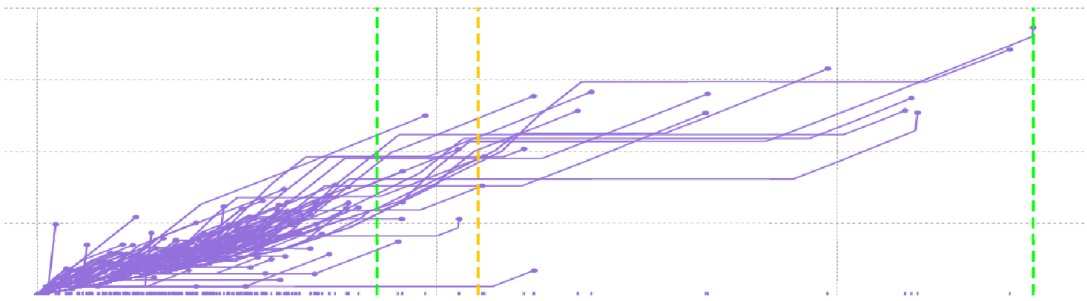
Visualization Optimization

The DV program can generate a possible optimization of angles and thresholds to find the highest accuracy. DV continuously looks for a new set of angles and threshold until a new combination resulting in higher accuracy is found. This can be initiated by clicking “Optimize Visualization” on the toolbar. If the user wishes to revert to the configuration prior to optimization, an “Undo Optimization” option exists to the right on the same toolbar.

Panning and Zooming Functions

The DV program allows users to pan across graphs by holding ctrl and dragging left click. For zooming, the DV program provides two options: one that preserves the aspect ratio of the graphs and one that scales the graphs. For the former option, by using the scroll wheel users can zoom in and out of the graphs at will. For the latter option, the DV program provides a drag and drop box method to select an area to zoom, starting from the top left corner, moving down and right. An example of this is shown below; the first image is the selection, and the second image is the selected area zoomed in and scaled.





Similarly, the user can zoom back out by using the drag and drop box method in reverse, i.e., starting from the bottom-right and moving to the top-left. To return to the original view, just click the “Reset Screen” button on the tool bar at the top of the screen.

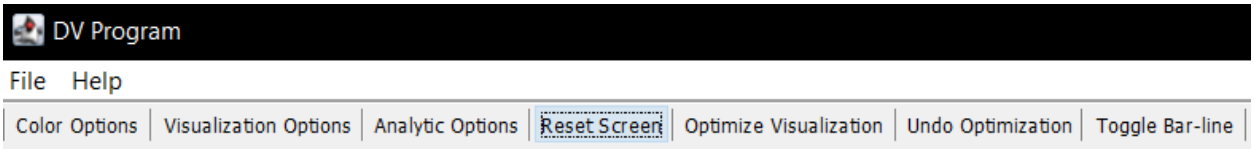


Image Options

The DV program also can export the visualization to a few different image options. By right clicking on a specific graph, the user may choose one of three options to export the image. Should the user choose “print…” a new window will option prompting the user to confirm setting to print the image. If the user chooses “Save as,” DV will open a file browser prompting the user to select where to save the PNG image. The last possible image option is to simply copy the PNG image to the user’s clipboard.

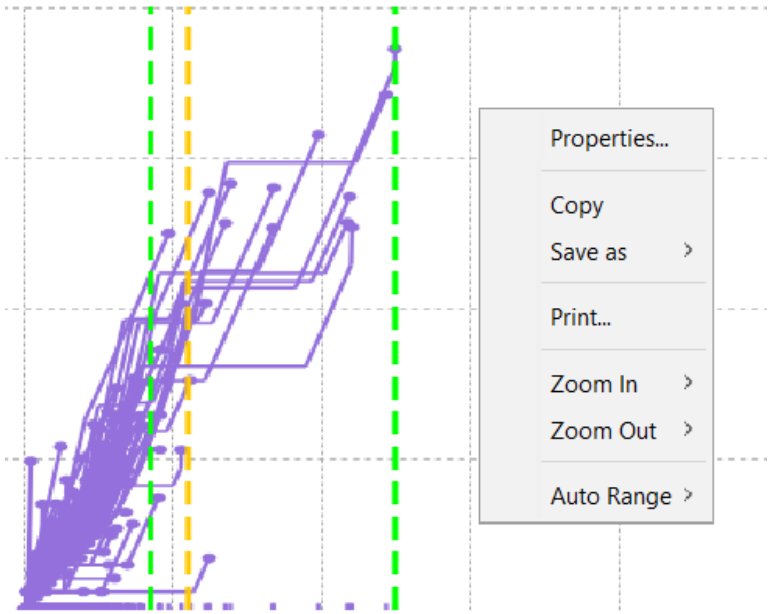
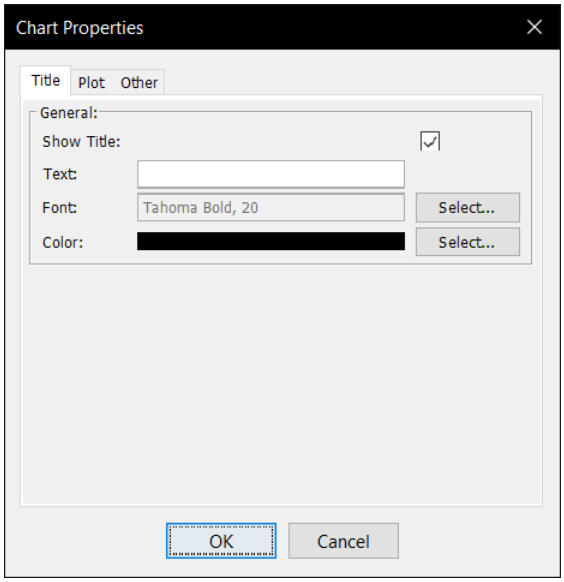


Chart Properties

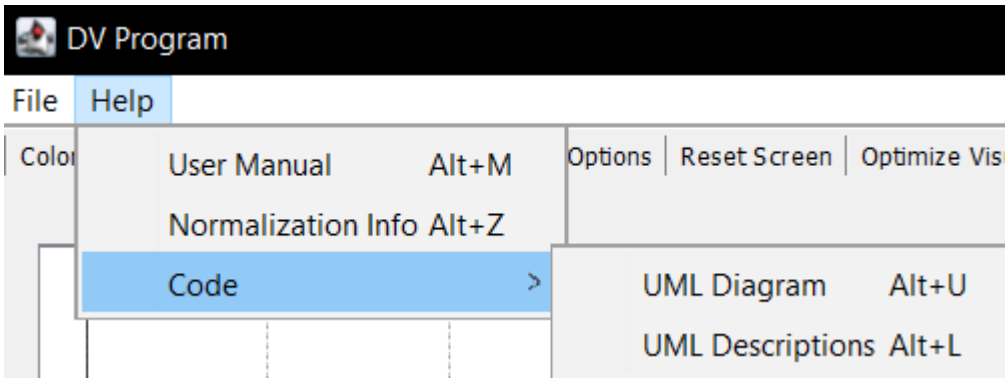
Additional customization options can be set using the Chart Properties window, which can be found in the “Properties” option when the user right clicks on the graph (shown in the previous

section). This will display the Chart Properties window where a user can add titles, label axes, and change colors of the graph elements. Options to add custom labels to the axes are found under the “Plot” tab.



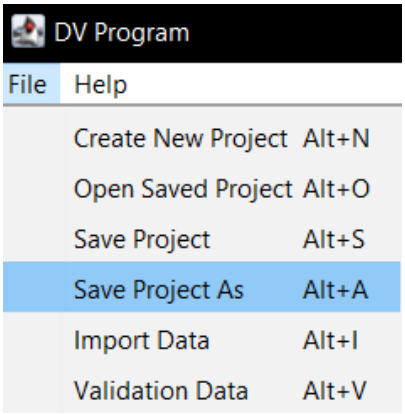
Help

If in need of help, finding the “Help” drop-down menu provides the following options: the user manual, normalization info, and the UML diagram and description of the code.



The user manual provides information on how to use the DV program, normalization info provides information on the different normalization methodologies used in the program, and both the UML diagram and descriptions are for the computer scientist looking to understand the programs code.

Saving a Project



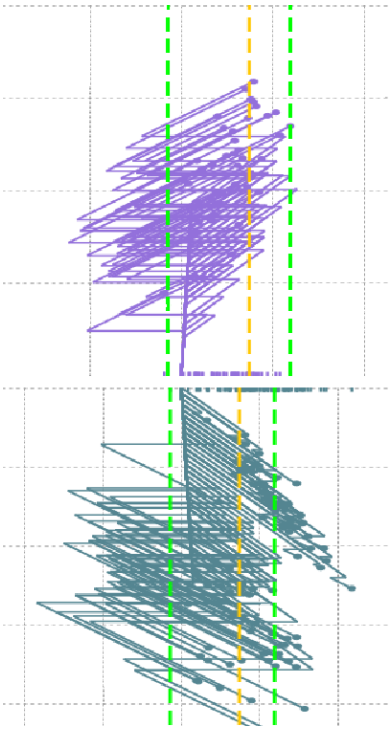
If a user wants to save a project for later use, the DV program has this option located under the “File” drop-down menu. For the initial save of a project, the user should click “Save Project As” bringing up the file browser. The VD program specifically needs .csv files for later use, thus the user should save a file with the “csv” extension. Once the project file has been created, the user may use the “Save Project” option from then on.

Opening a Saved Project

When a user wants to open a saved project, the option “Open Saved Project” under the “File” drop-down menu will open past projects. This button will bring up the file browser, where the user can search for their desired project. The DV program requires that all files are of the type “.csv” to open past projects.

3+ Class Visualizations

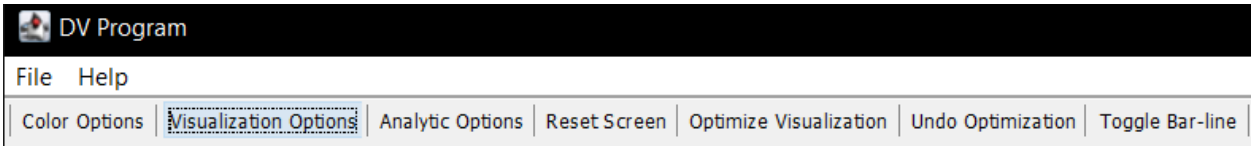
Since the DV program separates classes linearly on a single dimension, directly visualizing 3+ classes at the same time is nearly impossible with good results. Because of this, in 3+ class visualizations one class is visualization by itself on the upper graph, while all other classes are visualized together on the lower graph. This is shown below with the Iris dataset. The Iris-versicolor class is visualized on the upper graph, while the Iris-Setosa and Iris-Virginica classes are visualized on the lower graph.

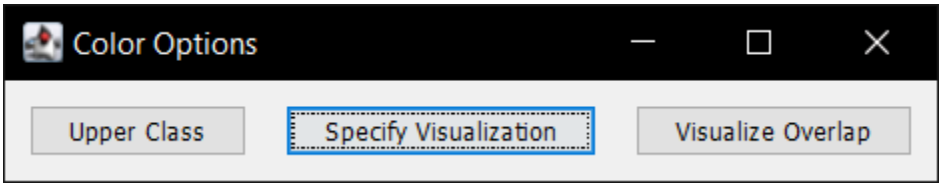


The confusion matrices also reflect these combined classes. The "All Data" confusion matrix on the right shows classes 0 and 2 (Iris-setosa and Iris-Virginica) combined.

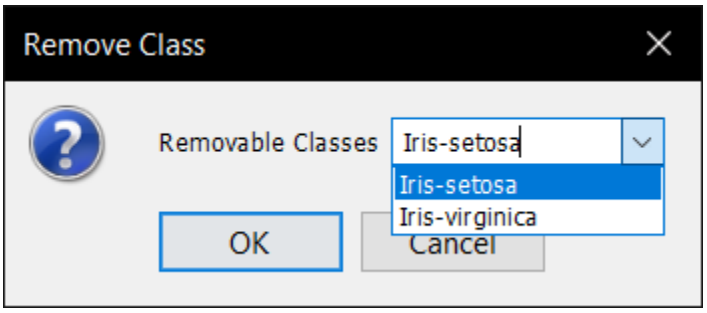
```
All Data Analytics
Real     Predictions
Class    1        0,2
1        40       10
0,2      26       74
Accuracy: 76.00%
Data Used: 100.00%
```

Then, users can use the “Specify Visualization” option in “Visualization Options” menu located on the toolbar.

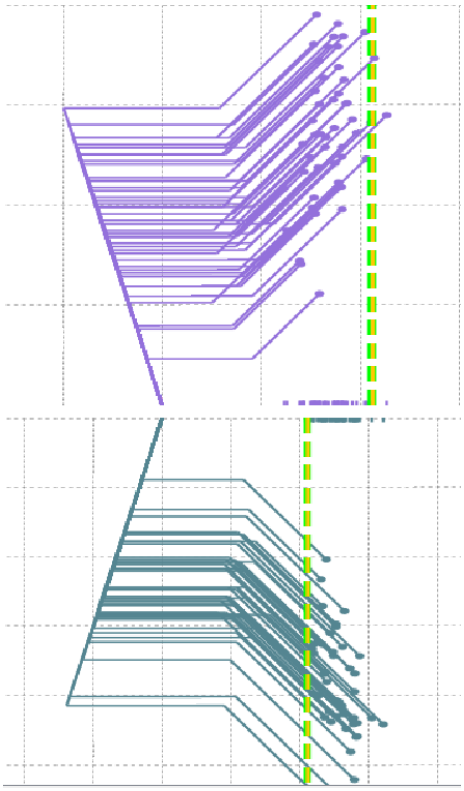




After selecting “Specify Visualization” a menu will appear giving a list of classes capable of being filtered out of the combined classes on the lower graph.



Selecting the Iris-setosa class to be removed from the visualization will create a new visualization with only the Iris-versicolor and Iris-virginica classes. This new visualization is shown below.



Additionally, the DV program will save the previous “All Data” confusion matrix for comparison to new the “All Data” confusion matrix. Moreover, the new “All Data” confusion matrix will include an additional analytic of “Overall Accuracy” this is the overall accuracy of both visualizations. These analytics can be seen below.

All Data Analytics

Real Class	Predictions 1	0,2
1	40	10
0,2	26	74

Accuracy: 76.00%

Data Used: 100.00%

All Data Analytics

Real Class	Predictions 1	2
1	48	2
2	1	49

Accuracy: 97.00%

Overall Accuracy: 84.40%

Data Used: 66.67%

All steps mentioned previously can be repeated with any number of classes until only two remain.