

## MODUL 14

# SAS® VDMML: BUILD & COMPARE SEVERAL TYPES OF MODELS

### THEME DESCRIPTION

Students understand and are able to implement the data science project using the SAS Visual Data Mining and Machine Learning to provides procedures for data exploration, model building, model comparison and selection.

### WEEKLY LEARNING OUTCOMES (SUB-LESSONS)

CLO-1-Sub-CLO-14:

Able to implement machine learning and its applications, and provide honest assessment using SAS VDMML with Forest and SVM Algorithm – C5.

Through the following learning steps:

1. Data Preparation
2. Model Building
3. Model Selection and Scoring

### PRACTICUM SUPPORT

- a. Windows Operating System
- b. (any) Browser Application

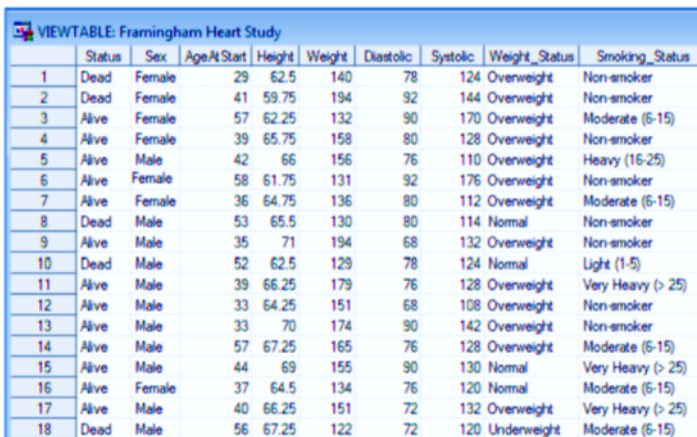
### PRACTICUM STEPS

Today practicum is an outline of the implementation of a data science project, including data preparation, building models, up to MODEL SELECTION & SCORING to determine the best and optimal model.

#### 1. DATA PREPARATION

##### Selecting and Exploring Data

- ▶ Selecting the Explore and Visualize Data tile takes you to the data screen:

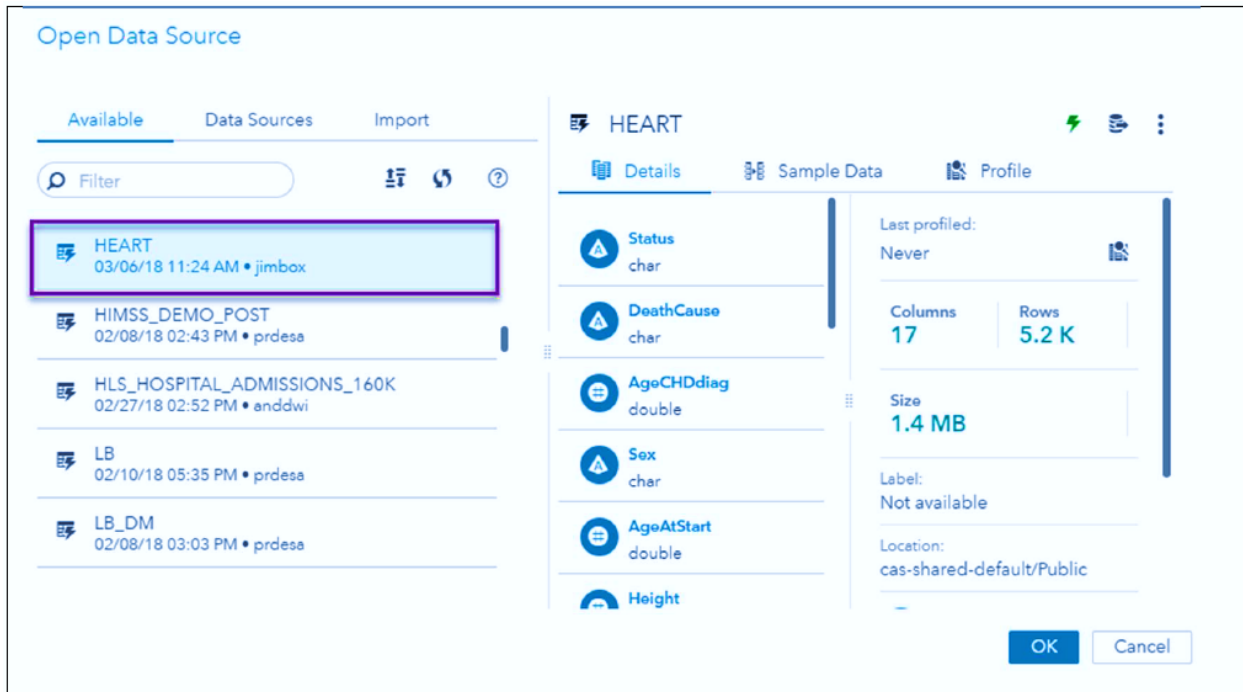


|    | Status | Sex    | AgeAtStart | Height | Weight | Diastolic | Systolic | Weight_Status | Smoking_Status    |
|----|--------|--------|------------|--------|--------|-----------|----------|---------------|-------------------|
| 1  | Dead   | Female | 29         | 62.5   | 140    | 78        | 124      | Overweight    | Non-smoker        |
| 2  | Dead   | Female | 41         | 59.75  | 194    | 92        | 144      | Overweight    | Non-smoker        |
| 3  | Alive  | Female | 57         | 62.25  | 132    | 90        | 170      | Overweight    | Moderate (6-15)   |
| 4  | Alive  | Female | 39         | 65.75  | 158    | 80        | 128      | Overweight    | Non-smoker        |
| 5  | Alive  | Male   | 42         | 66     | 156    | 76        | 110      | Overweight    | Heavy (16-25)     |
| 6  | Alive  | Female | 58         | 61.75  | 131    | 92        | 176      | Overweight    | Non-smoker        |
| 7  | Alive  | Female | 36         | 64.75  | 136    | 80        | 112      | Overweight    | Moderate (6-15)   |
| 8  | Dead   | Male   | 53         | 65.5   | 130    | 80        | 114      | Normal        | Non-smoker        |
| 9  | Alive  | Male   | 35         | 71     | 194    | 68        | 132      | Overweight    | Non-smoker        |
| 10 | Dead   | Male   | 52         | 62.5   | 129    | 78        | 124      | Normal        | Light (1-5)       |
| 11 | Alive  | Male   | 39         | 66.25  | 179    | 76        | 128      | Overweight    | Very Heavy (> 25) |
| 12 | Alive  | Male   | 33         | 64.25  | 151    | 68        | 108      | Overweight    | Non-smoker        |
| 13 | Alive  | Male   | 33         | 70     | 174    | 90        | 142      | Overweight    | Non-smoker        |
| 14 | Alive  | Male   | 57         | 67.25  | 165    | 76        | 128      | Overweight    | Moderate (6-15)   |
| 15 | Alive  | Male   | 44         | 69     | 155    | 90        | 130      | Normal        | Very Heavy (> 25) |
| 16 | Alive  | Female | 37         | 64.5   | 134    | 76        | 120      | Normal        | Moderate (6-15)   |
| 17 | Alive  | Male   | 40         | 66.25  | 151    | 72        | 132      | Overweight    | Very Heavy (> 25) |
| 18 | Dead   | Male   | 56         | 67.25  | 122    | 72        | 120      | Underweight   | Moderate (6-15)   |

Figure 1. Framingham Heart Study Viewtable

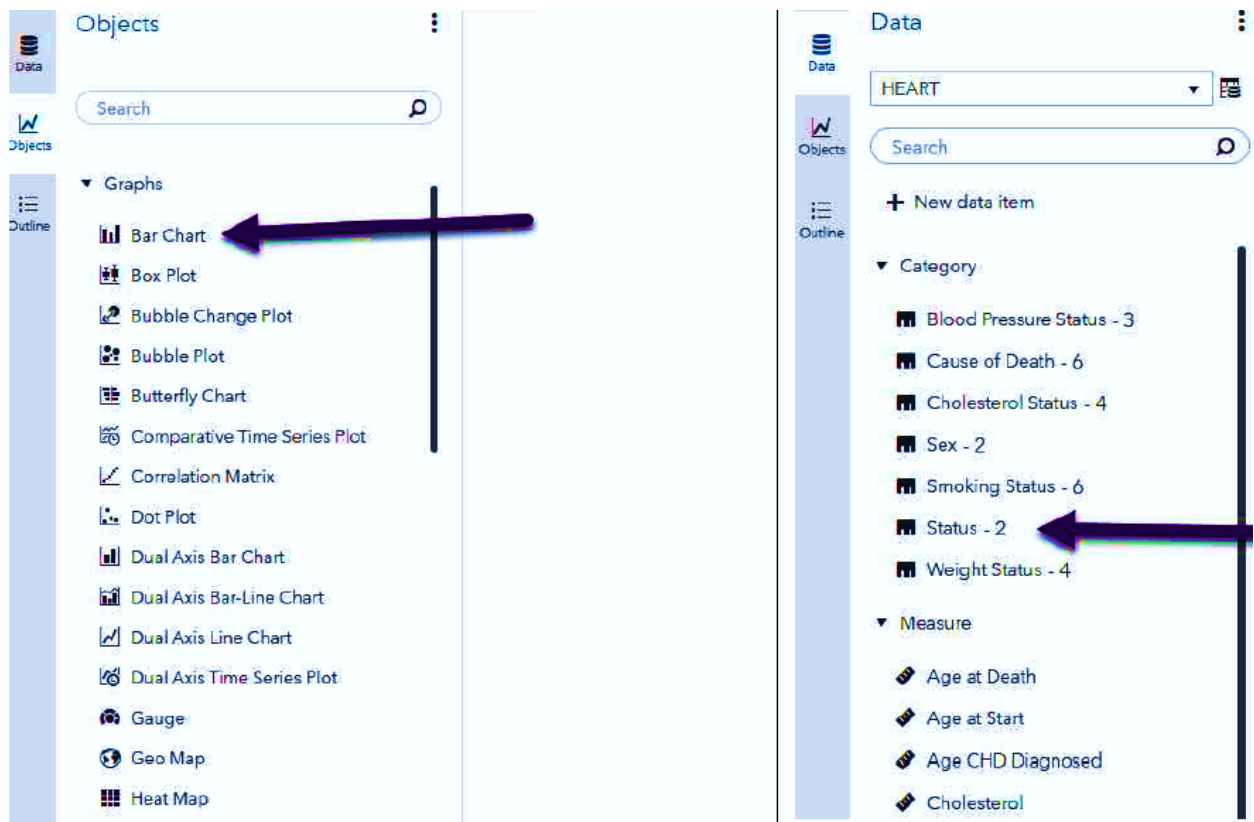
- a. For this example, we will use the SASHELP.HEART data, which has results from the Framingham Heart Study. Figure 1 and Figure 2 shows the data selection screen, and it gives us a count of the

number of rows and columns. We can also get a snapshot of the data in a table to further look at the values.



**Figure 2: Data Selection**

- b. The response variable we are interested in is the Status variable – it is an indicator of survival. We'll use the exploration tools to get an understanding of this response. Figure 3 shows that we select the objects button from the left menu and select a bar chart, and then select the Status variable to get a distribution of the responses



**Figure 3: Create Bar Chart of Status**

- c. We can set up the bar chart a variety of ways by choosing different Options and Roles. Figure 4 shows the results of using vertical bars and showing Data Labels in the Options menu and selecting both Frequency and Frequency Percent as Measures in the Roles menu. As a result, we see that just over 38% of the records show a status of “Dead.” We would like to explore some different model options and fits a good model that predicts death as a status for future patients.

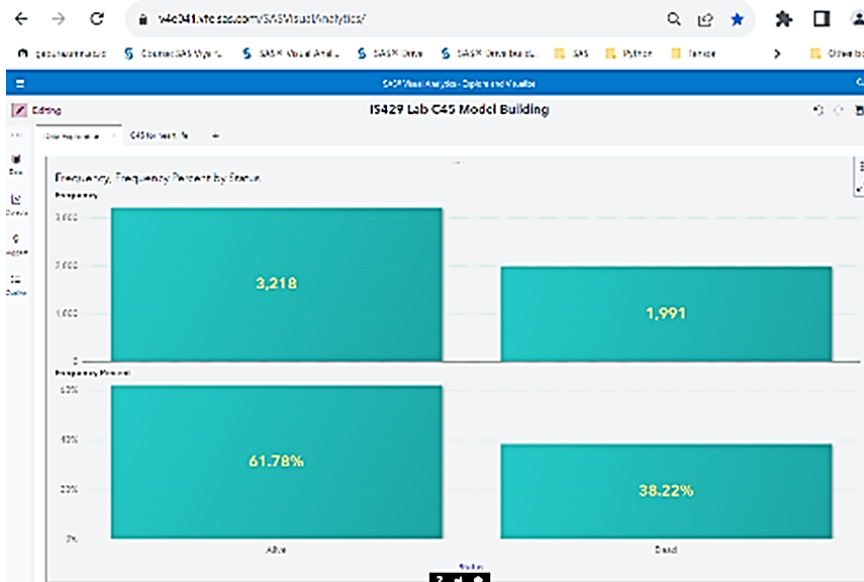


Figure 4: Endpoint Breakdown

## Partitioning Data

- ▶ When developing models that will be used for prediction, it's important to evaluate their predictive power.
  - ▶ It's also important to not overfit the data – we want the model to be useful at predicting outcomes for data that was not part of the construction of the model.
- d. If a partition variable exists in the dataset (usually coded as a 0 or 1, randomly assigned), we can identify the variable as a partition data type, but generally we'll need to make a partition variable. Figure 5 shows how we accomplish this. First, we click on the data table and select the “Add partition data item...” option. Then we give the partition column a name, tell it that we want to use two partitions, and set the training partition size to 70%.

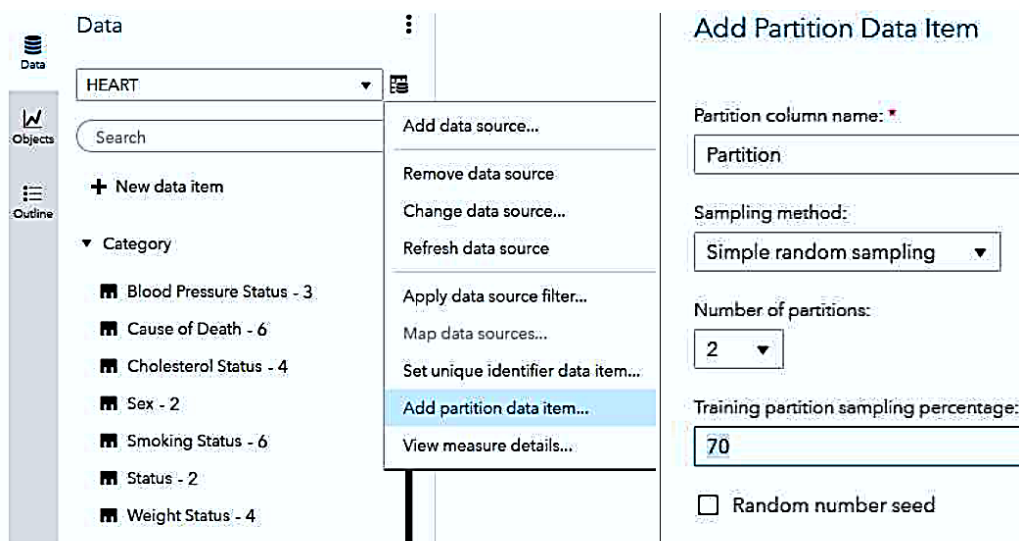


Figure 5: Partition Creation

## 2. MODEL BUILDING

- ▶ There are several types of models we could build to predict the status variable. For this demonstration, we will make four different models with the training data, then have the system compare the four models and tell us which one did the best job correctly predicting the status of patients in the validation partition.
- ▶ There are two general types of models we might try out: statistical models and machine learning models.

### STATISTICAL MODELS

- The first approach is to look at some of the traditional statistical models that could be used. Figure 6
- shows the different model types we could use. Since our response variable has a binary outcome, we will use the two main approaches for this type of data: a Decision Tree and a Logistic Regression.

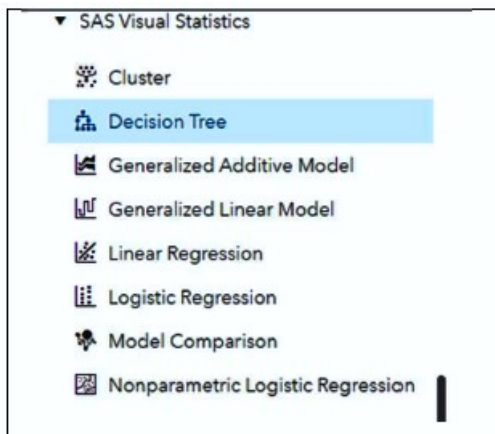


Figure 6: Statistical Models

C.

### Decision Tree

- We start with a decision tree, as it is generally the easiest model to understand. To build the tree we'll add a new tab to our exploration, go into the objects menu and select the decision tree in the list we saw in Figure 6.
- Next, we go to the Roles tab on the left menu and set the Response variable to Status and the Partition ID to Partition.
- Finally, we will select the Predictors we want to consider. Decision trees can take both category and measure types of variables. It is a best practice to not select multiple variables that address the same thing.
- For example, in Figure 7 we are selecting Smoking (measured in number of cigarettes) but not Smoking Status, which is a category variable made from the smoking variable.

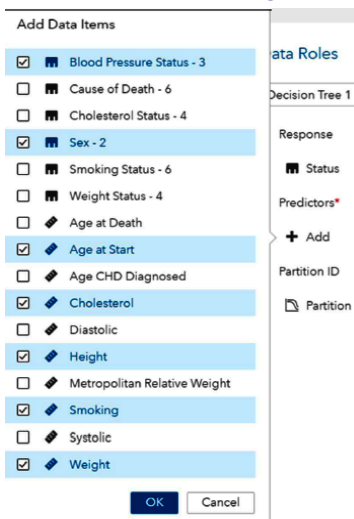
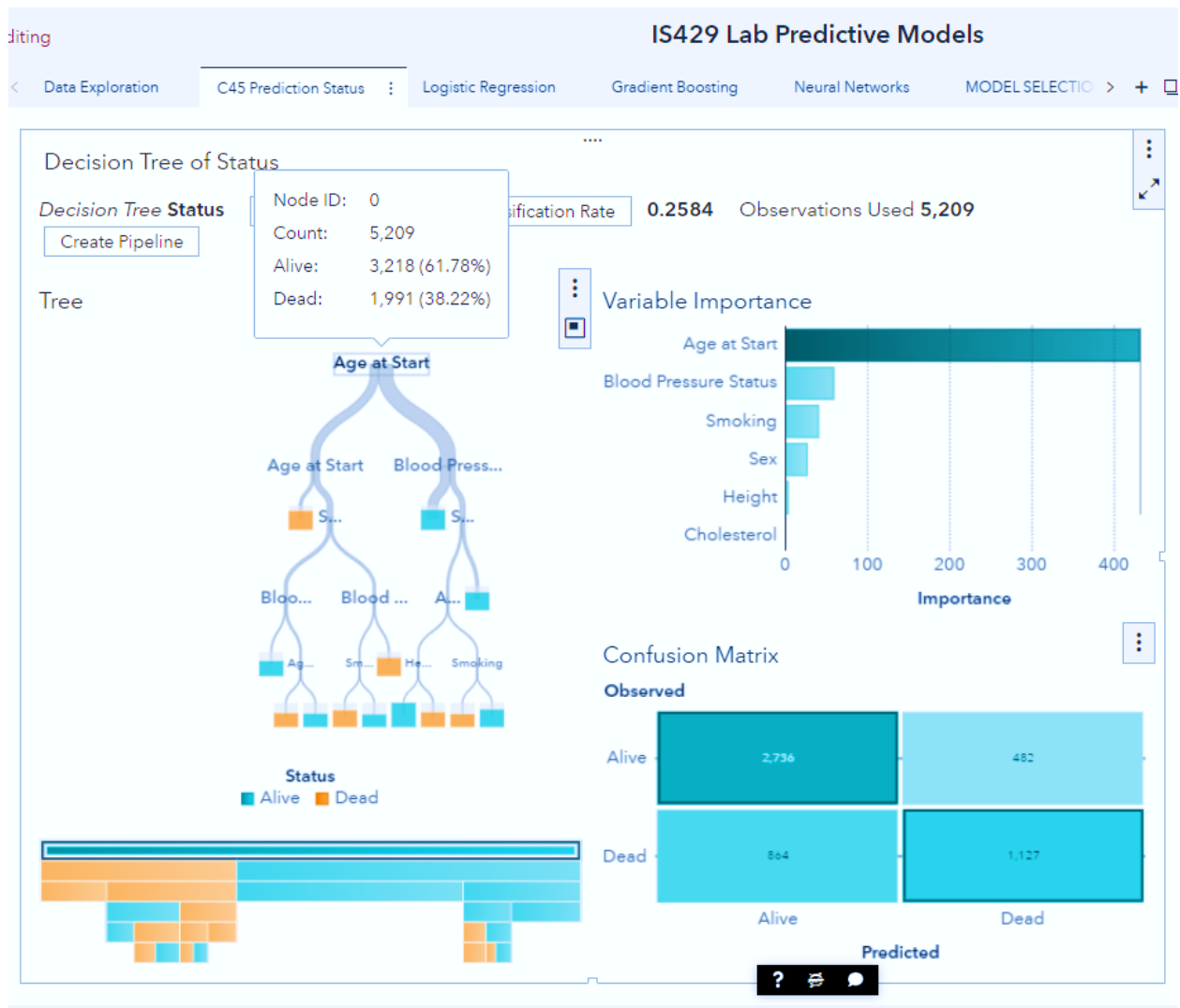


Figure 7: Variable Selection

- h. As soon as the variables are selected, the decision tree is generated and output is produced in the tab. Figure 8 shows the tree. The first split was on age, indicating that this variable was the most important in predicting the status. The data was cut on the value of age of 48, with 62% of patients older than that having a status of Dead (compared to the overall 38%).



- ▶ Note that we can choose the model comparison statistic we want to look at. Here we have chosen to look at the misclassification rate in the validation partition.
- ▶ This is telling us that the model was built on the training data and when it was used to predict the status of the validation subjects, it was wrong 25.84% of the time.
- ▶ This statistic will be useful when we are comparing all four models at the end of the process select OK.

### Logistic Regression

- ▶ The next model we will build will be a logistic regression. We could add a new tab and then assign role like we did with the decision tree, but we have a shortcut available to us that makes it easier.
  - ▶ Figure 9 shows the way to do this; to get to that menu, we will click on the three vertical dots we see at the top right on the decision tree we see in Figure 8 (this three-dot menu is sometimes called the snowman menu).
- i. When the menu pops up, hold down the ALT key and the "Duplicate as" selection becomes "Duplicate on New Tab as". Here we will select Logistic Regression, and all the Response, Predictors and Partition selections we made on the Decision Tree will be copied onto the Logistic Regression roles in a new tab. If we did not hold down the ALT key, Logistic Regression output would have been added below the Decision Tree, which makes it hard to see both models.

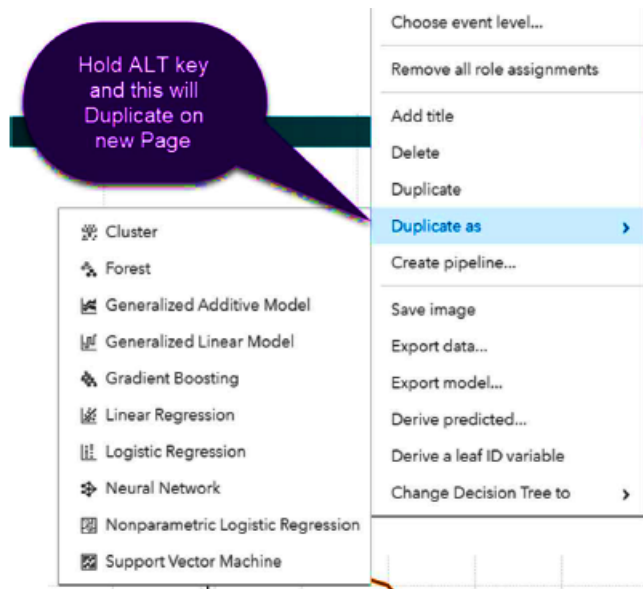


Figure 9: Duplicate Models

- j. The Logistic Regression output is shown in Figure 10. We have selected the same validation misclassification rate to be the model statistic; it's a little better than the Decision Tree. Remember that a logistic regression model requires that there be no missing values in any of the selected predictors; any patient with a missing value is excluded. In this case, that meant we had to throw away 170 of the patients. That's a small number here, but it is important to consider when we are looking at other datasets.

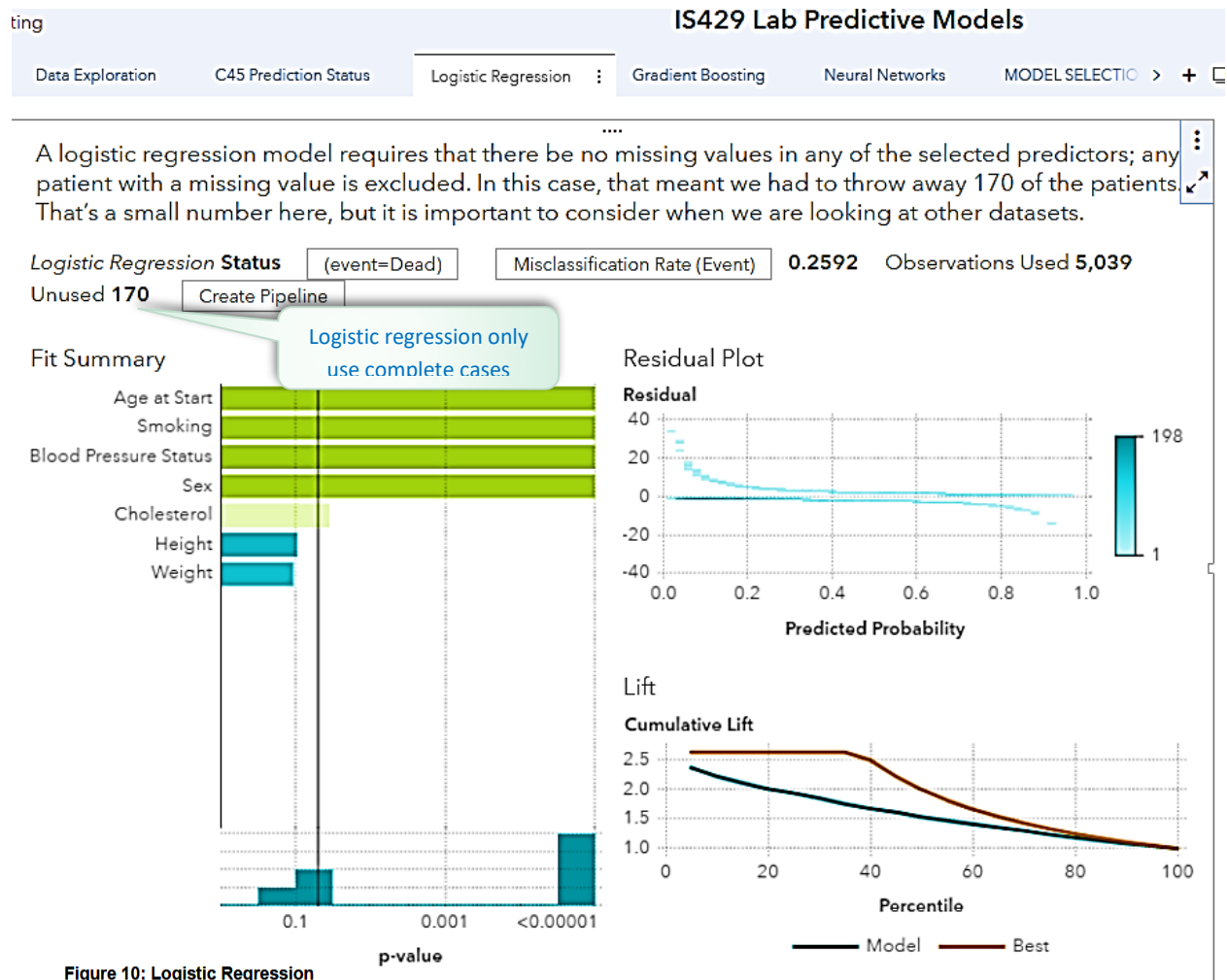


Figure 10: Logistic Regression

## MACHINE LEARNING MODELS

- ▶ In addition to the traditional statistical models, we also can use some of the more advanced machine learning models.
- ▶ Figure 11 shows the types of machine learning models that are available. They will often have less restrictive assumptions than do the statistical models, and it is helpful to read up on how they work before considering them as candidate models.
- ▶ Here we will use two of the most commonly talked about machine learning models: Gradient Boosting and Neural Networks.



Figure 11: Machine Learning Models

## Gradient Boosting

- ▶ Gradient Boosting is a popular machine learning technique for this type of classification problem.
- ▶ It's a complex model, but for our purposes, we can think of it as a technique that uses a bunch of different decision trees to make a better, combined predictive model.
- ▶ We won't get into the discussion of how to tune parameters here, but the Options menu gives us a great deal of control over how the model is implemented.

- We'll just stick with the defaults and look at the outputs. Figure 12 shows that output; note that we see the same sort of variable importance that we have in the Decision Tree, and that we can still look at the Validation Misclassification Rate.

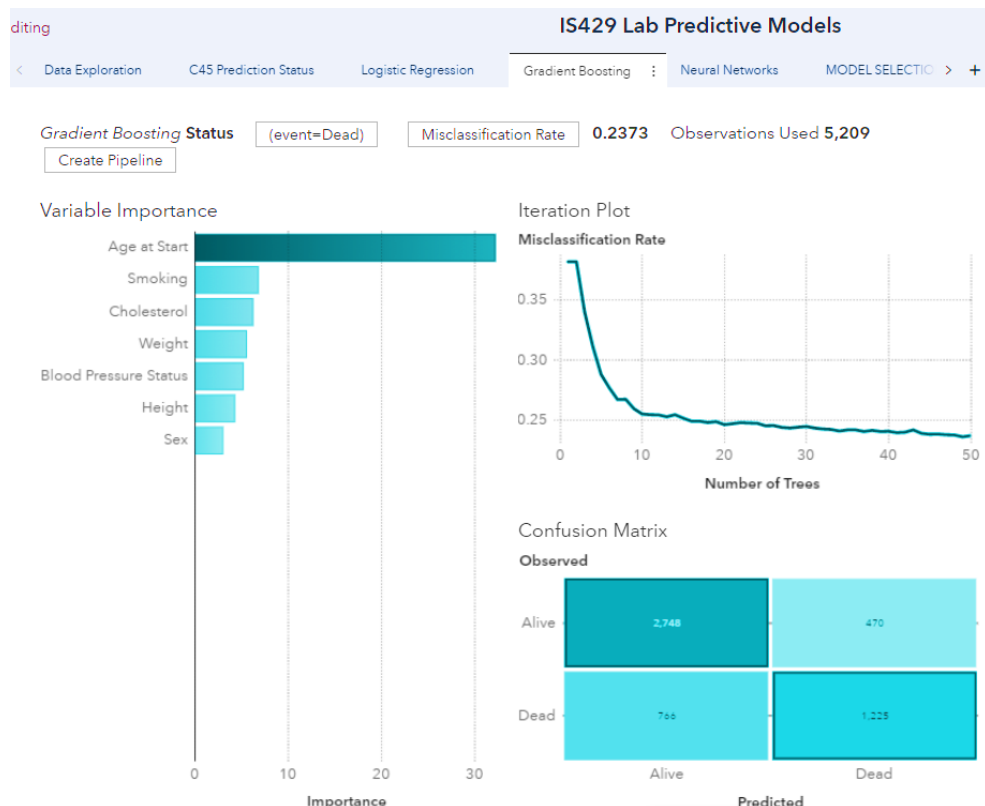


Figure 12: Gradient Boosting



## Neural Networks

- b. The final model we will add is probably the most complicated: the Neural Network. A high-level view of a Neural Network is that it is a collection of nested regressions to get to the predicted variable. Like the Gradient Boosting model, we can just use the model defaults for this exercise, but there are a variety of options we could adjust to try to fine-tune the network.
- ▶ As before, we can see the Validation Misclassification Rate in Figure 13. Notice that this model also
  - ▶ requires complete cases, and 170 observations were not used.

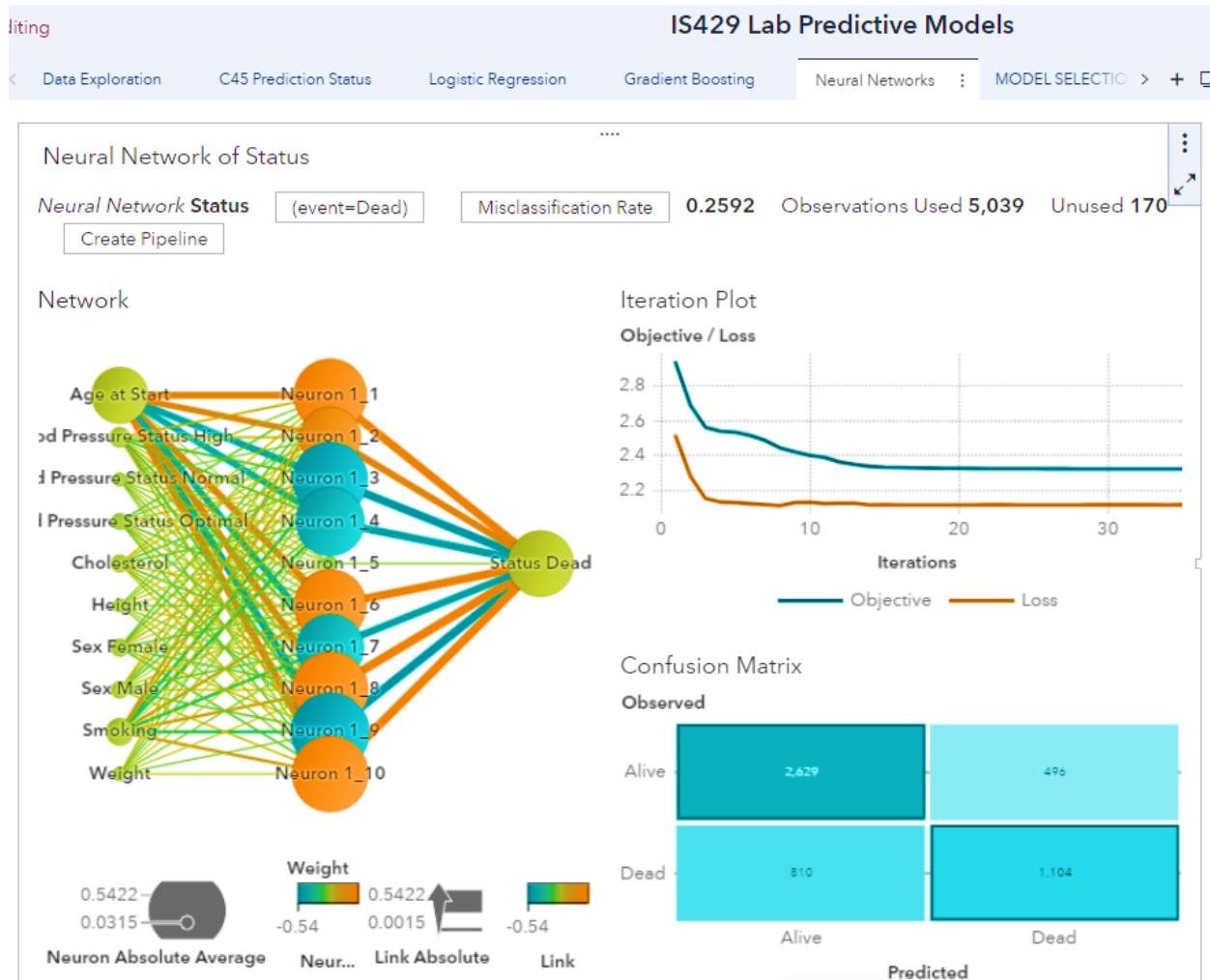


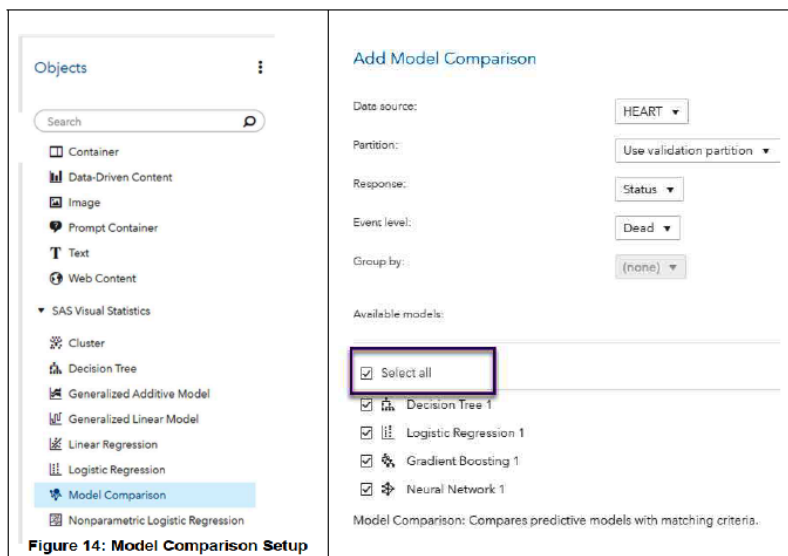
Figure 13: Neural Networks



### 3. MODEL SELECTION & SCORING

- ▶ Now that we have created four candidate models, it's time to do a model comparison.
- ▶ Figure 14 shows the process. First, we add a new tab to the exploration, then select the Objects menu on the left. **STATISTICAL MODELS**

- In the Visual Statistics menu, select Model Comparison, and a dialog box pops up. The data source, Partition, Response, Event Level and Group By selections are all made by default to match the first model. All models that match that criterion show up in the Available Models section (you cannot compare a model that has a validation partition with one that does not, for example). Here we selected all, and the results were added to the tab.



- The model comparison results are shown in Figure 15. Also note that by right-clicking on any of the model bars, we can change the statistic we use to compare the models.
  - ▶ There are several available, on both the training and the validation partition.
  - ▶ We've chosen the Validation Misclassification Rate, since an accurate prediction is how we want to select the final model.

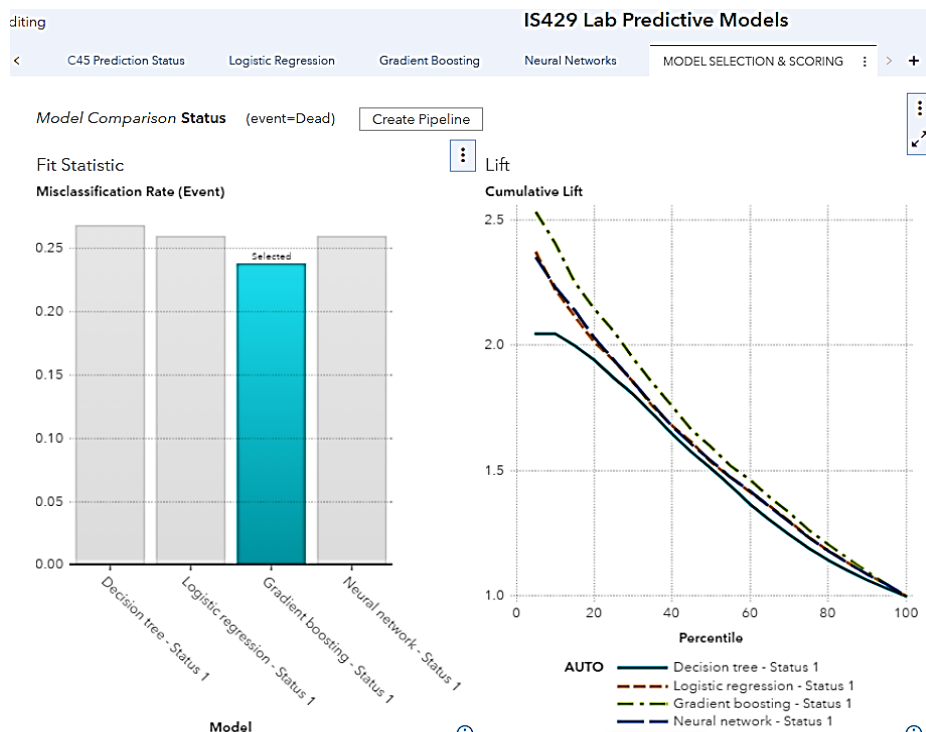
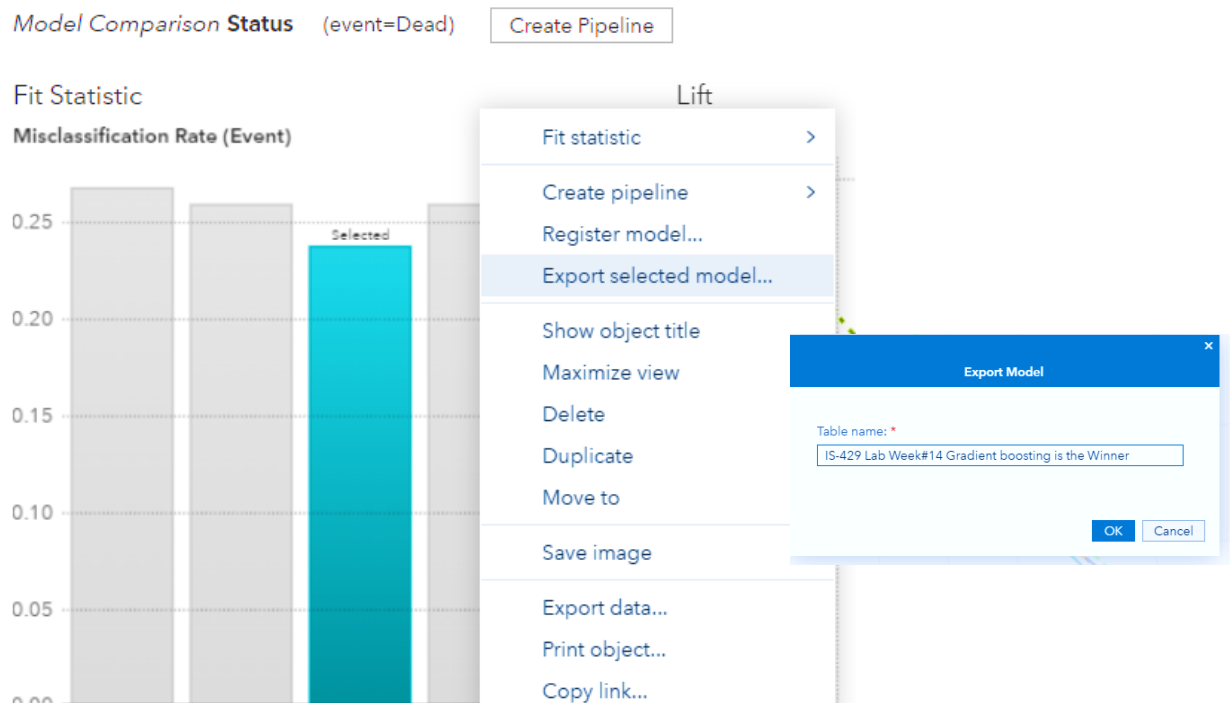
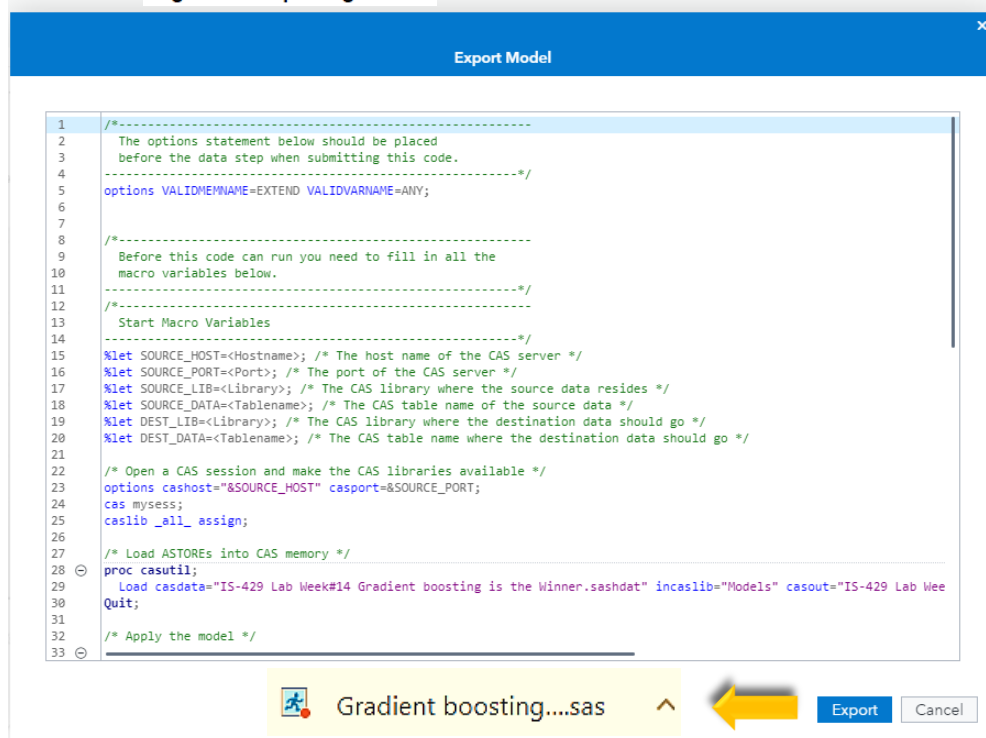


Figure 15: Model Comparison Results



### Figure 16: Exporting Models



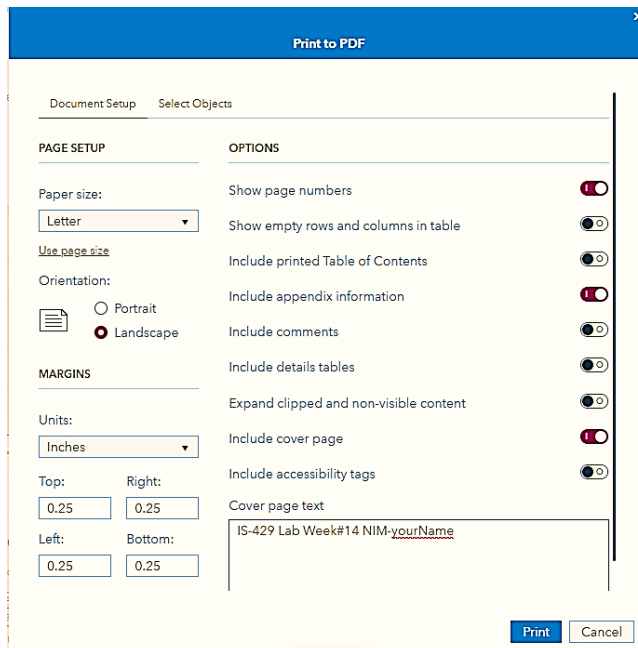
### Figure 17: Exporting Code

- ▶ Once the export has been selected, the box in Figure 17 appears.
- ▶ This contains the SAS code that would be applied to new observations and would generate a predicted value for Status.
- ▶ We would use this to evaluate new patients who were being seen and evaluated for risk of death.
- ▶ Once we have observed several more outcomes, we could redo the entire process with the old and new data and refresh our models.

## CONCLUSION

- ▶▶ The SAS VDMML makes it very easy to produce, evaluate and export models.
- ▶▶ There are options for traditional statistical methods and for newer, more complex machine learning models.
- ▶▶ Users can have a lot of control over the model options, or they could accept the defaults provided.
- ▶▶ There are other ways to build models without code in this interface Visual Exploration method is the easiest way to interact with complicated models.

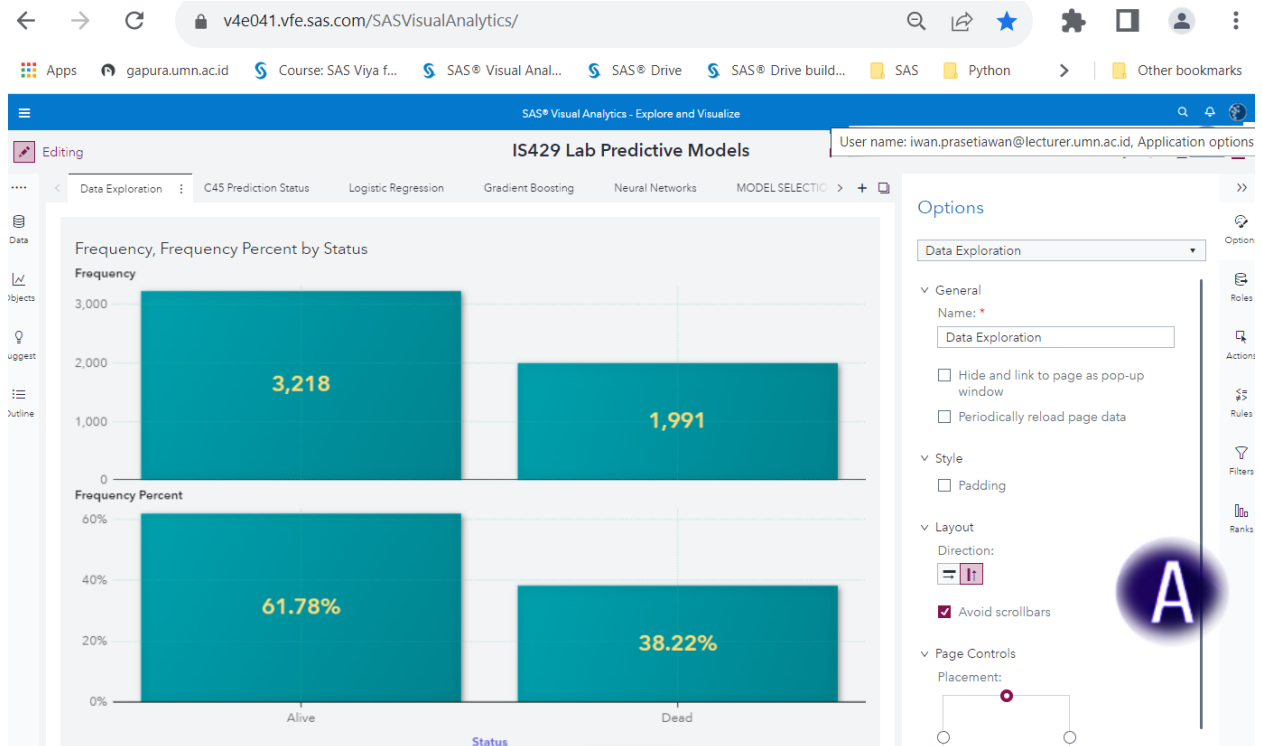
- ❖ Save your VDMML report as "IS-429 Lab Week#14 Models Build & Comparison NIM".
- ❖ Print your VDMML report as shown below:



- ➔ Finally, today's practicum is over, Zipped your VDMML report.pdf and SAS program produced on exercise #3.C and submit immediately today to e-Learning IS-429 Practicum Week#9 with the naming format IS-429 Lab Week#14 NIM yourName.zip.

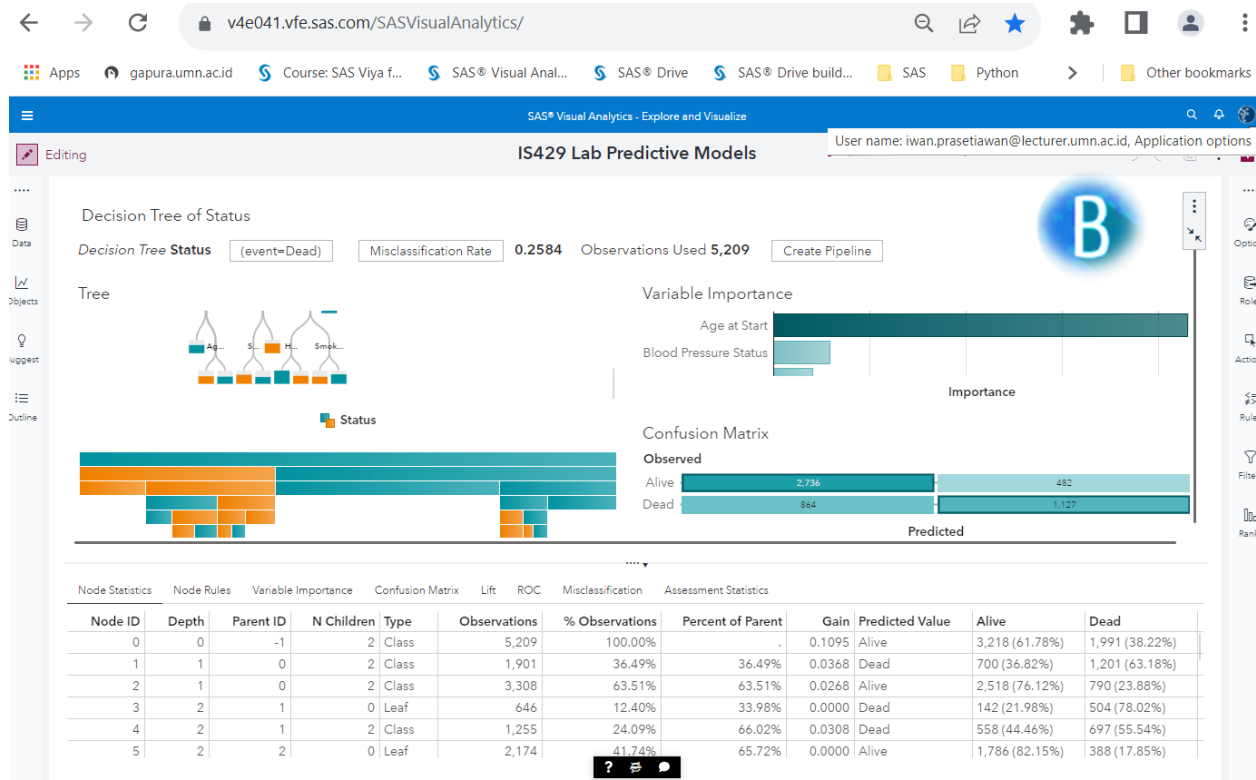
## RESULTS/ OUTPUT

### 1. Data Preparation



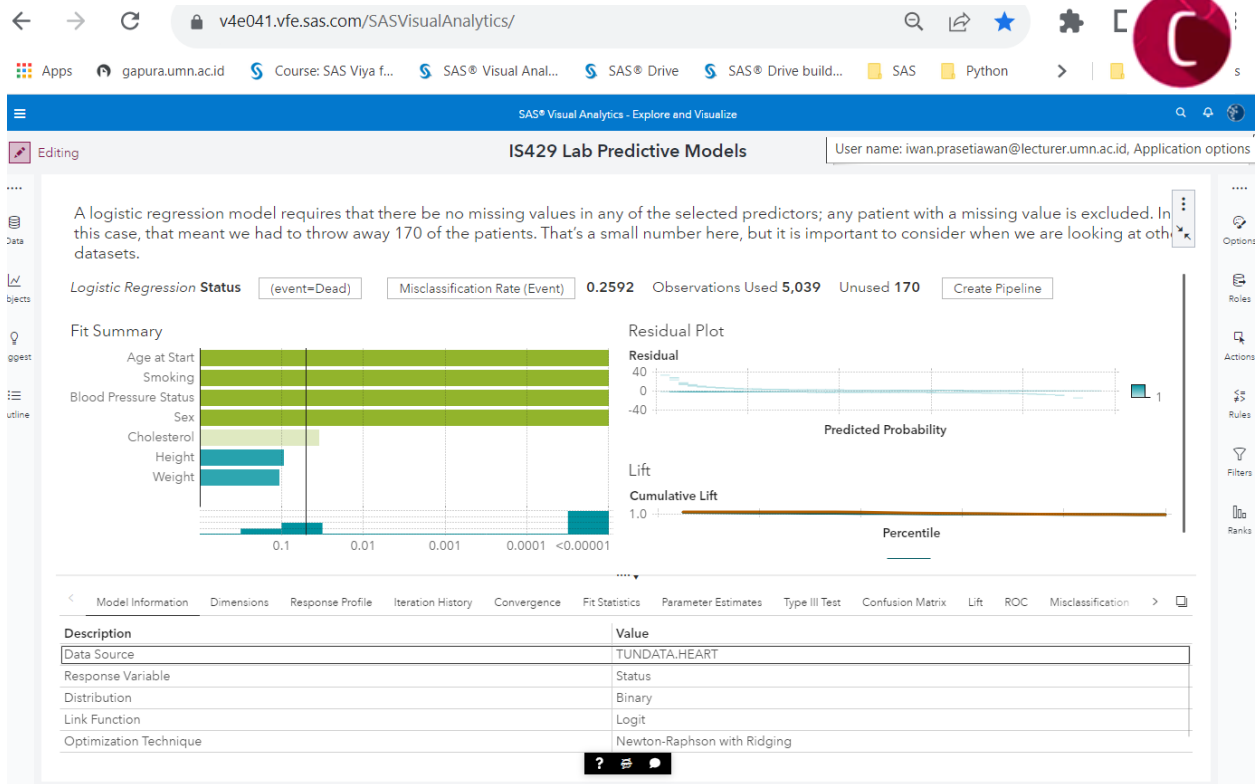
### 2. Model Building

#### a. C45 (Decision Tree)

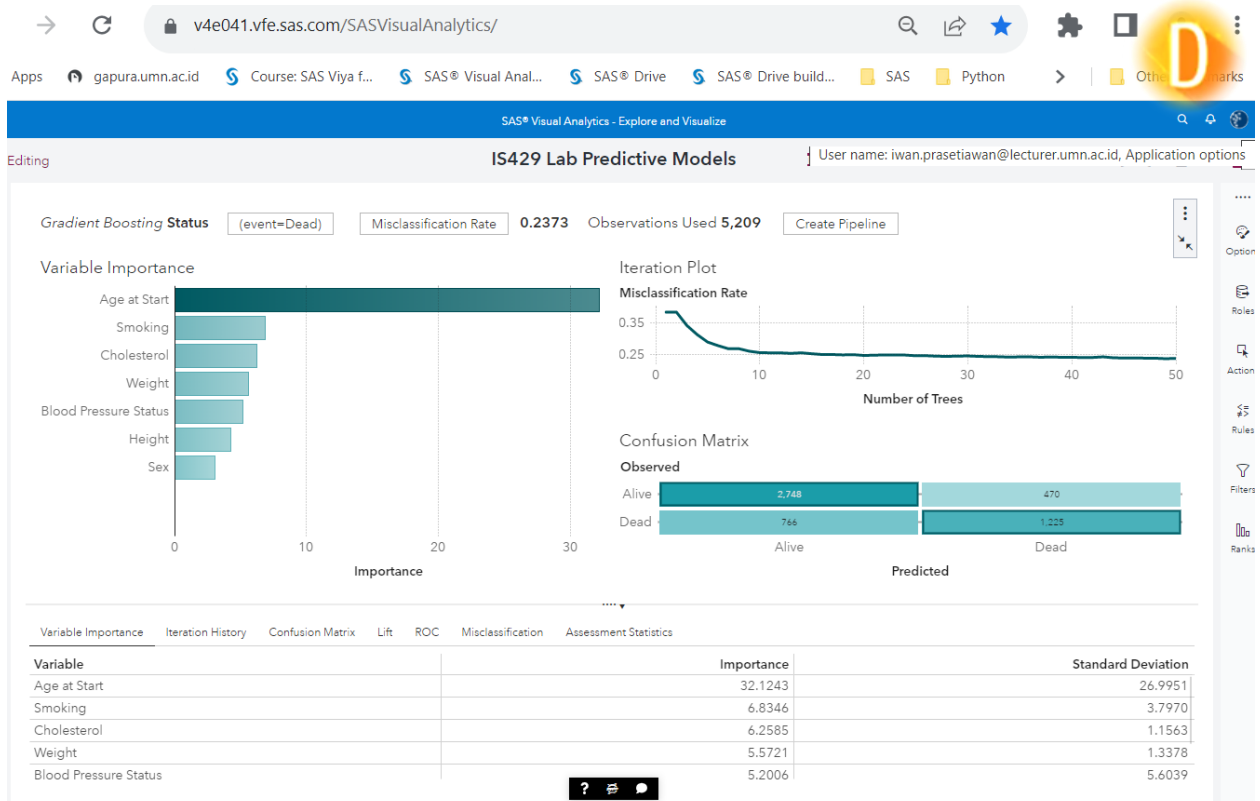


## Model Building

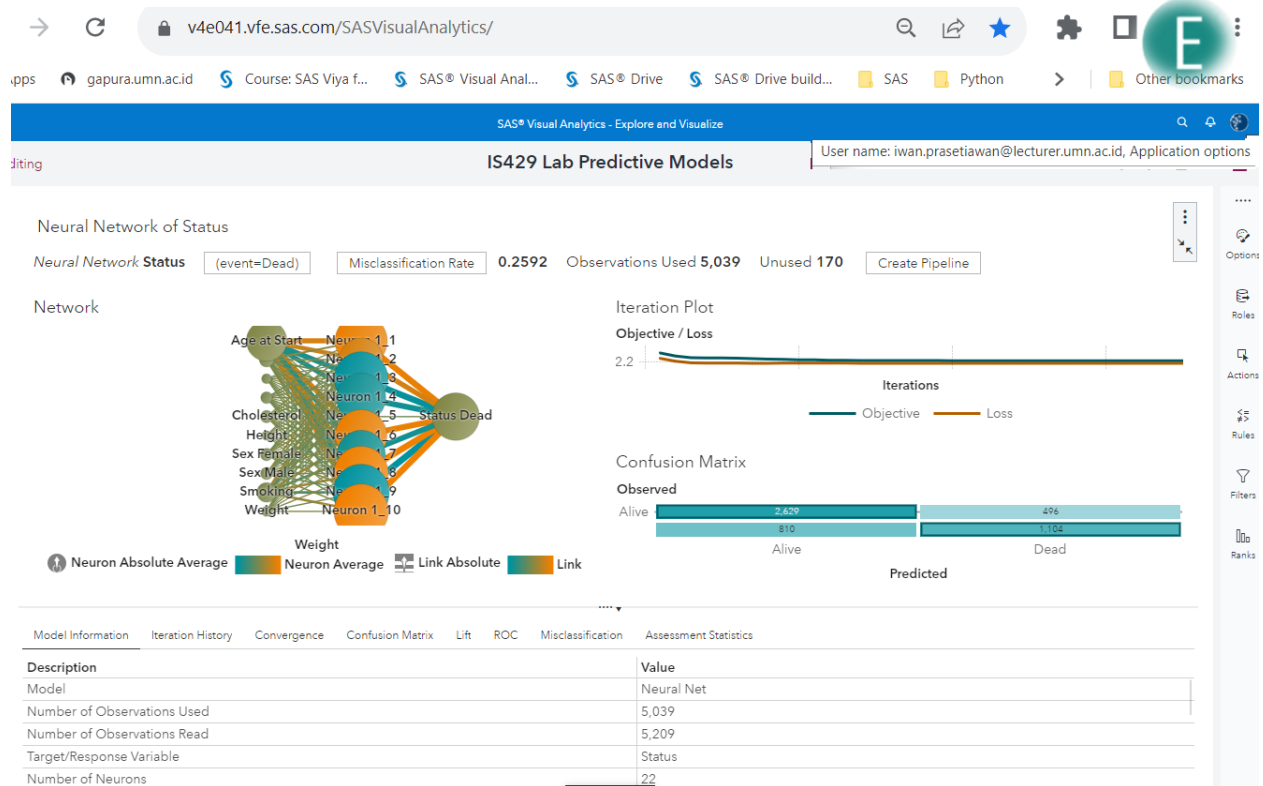
### b. Logistic Regression



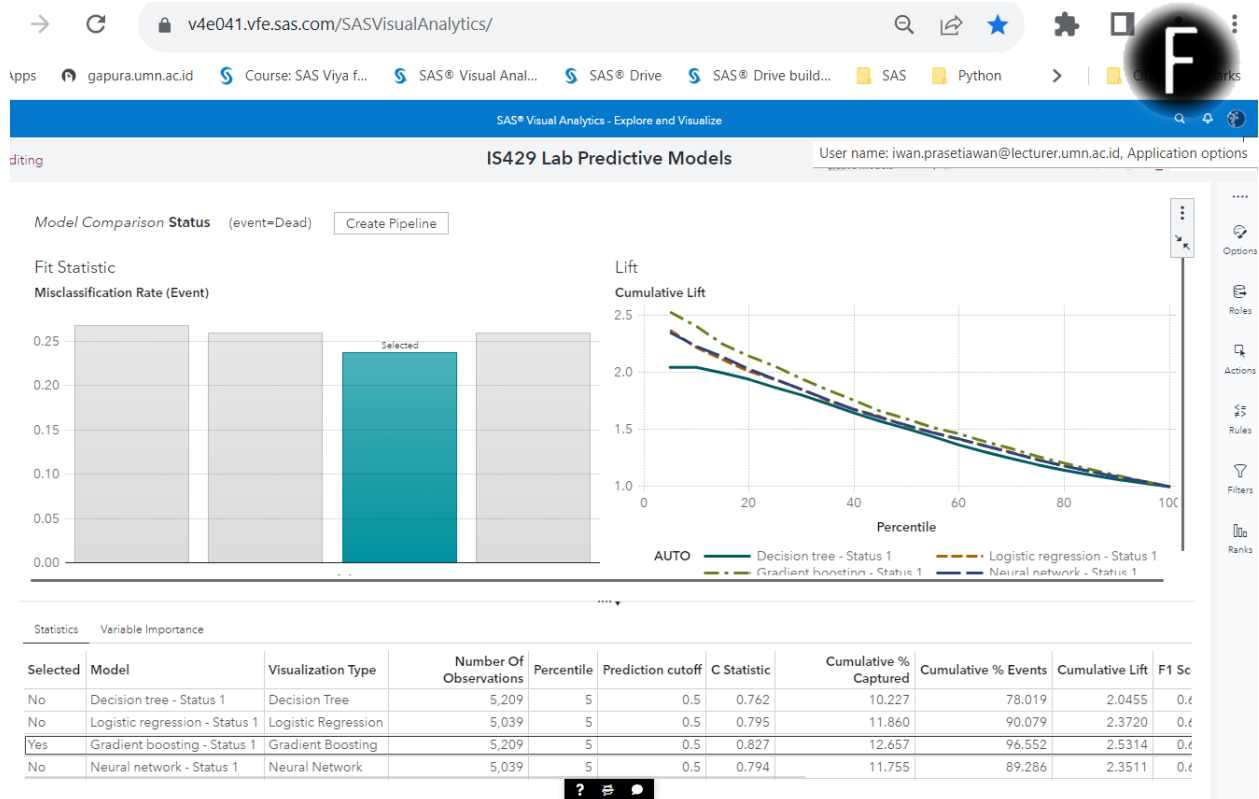
### c. Gradient Boosting



## d. Neural Network



## 3. Model Selection and Scoring



The  
End

## REFERENCE

1. SAS Institute Inc. 2020. SAS® Visual Data Mining and Machine Learning in SAS® Viya®: Interactive Machine Learning. SAS Institute Inc. Cary, NC, USA.
2. SAS Institute Inc. 2020. SAS® Viya® Programming: Getting Started. SAS Institute Inc. Cary, NC, USA.
3. SAS® Support | Documentation
4. Other additional references are excerpts from various Online Learning/websites.