***Chapter 6***

1. Explain the difference between supervised and unsupervised methods. Which data mining tasks are associated with unsupervised methods? Supervised? Both?

In supervised methods, there is a particular target variable, and the algorithm is given many examples where the value of the target variable is provided, so that the algorithm may learn which values of the target variable are associated with which values of the predictor variables.

Most of classification methods such as decision trees, neural networks, and k-nearest neighbors and regression models are supervised methods.

In unsupervised methods, no target variable is identified as such. Instead, the data mining algorithm searches for patterns and structure among all the variables.

The most common unsupervised data mining method is clustering.

The data mining method, which may be supervised or unsupervised, is association rule mining.

3- Should we strive for the highest possible accuracy with the training set? Why or why not? How about the validation set?

Maximizing the accuracy of the data model against the training set leads to a complex model (overfitting) and also, Increasing the complexity of the model leads to a degradation in the generalizability of the provisional model to the test set and validation dataset

***Chapter 7***

1. Clearly describe what is meant by classification.

classification is the process of sorting and categorizing data into various types, forms or any other distinct class.

In classification, there is a target categorical variable which is partitioned into predetermined classes or categories. The data mining model examines a large set of records, each record containing information on the target variable as well as a set of input or predictor variables.

First, the algorithm examines the data set containing both the predictor variables and the target variable. In this way, the algorithm learns about which combinations of variables are associated with which target variables. This data set is called the training set. Then the algorithm would look at new records, based on the classifications in the training set, the algorithm would assign classifications to the new records.

1. What is meant by the term instance-based learning?

Instance-based learning is a family of learning algorithms which the training data set is stored, so that a classification for a new unclassified record may be found simply by comparing it to the most similar records in the training set.

1. Discuss the advantages and drawbacks of using a small value versus a large value for

K?

By choosing a small value for k, it is possible that the classification or estimation may be unduly affected by outliers or unusual observations, noise.

the algorithm will simply return the target value of the nearest observation, a process that may lead the algorithm toward overfitting, tending to memorize the training data set at the expense of generalizability.

On the other hand, choosing a value of k that is not too small will tend to smooth out any idiosyncratic behavior learned from the training set. However, if we choose a value of k that is too large, locally interesting behavior will be overlooked.

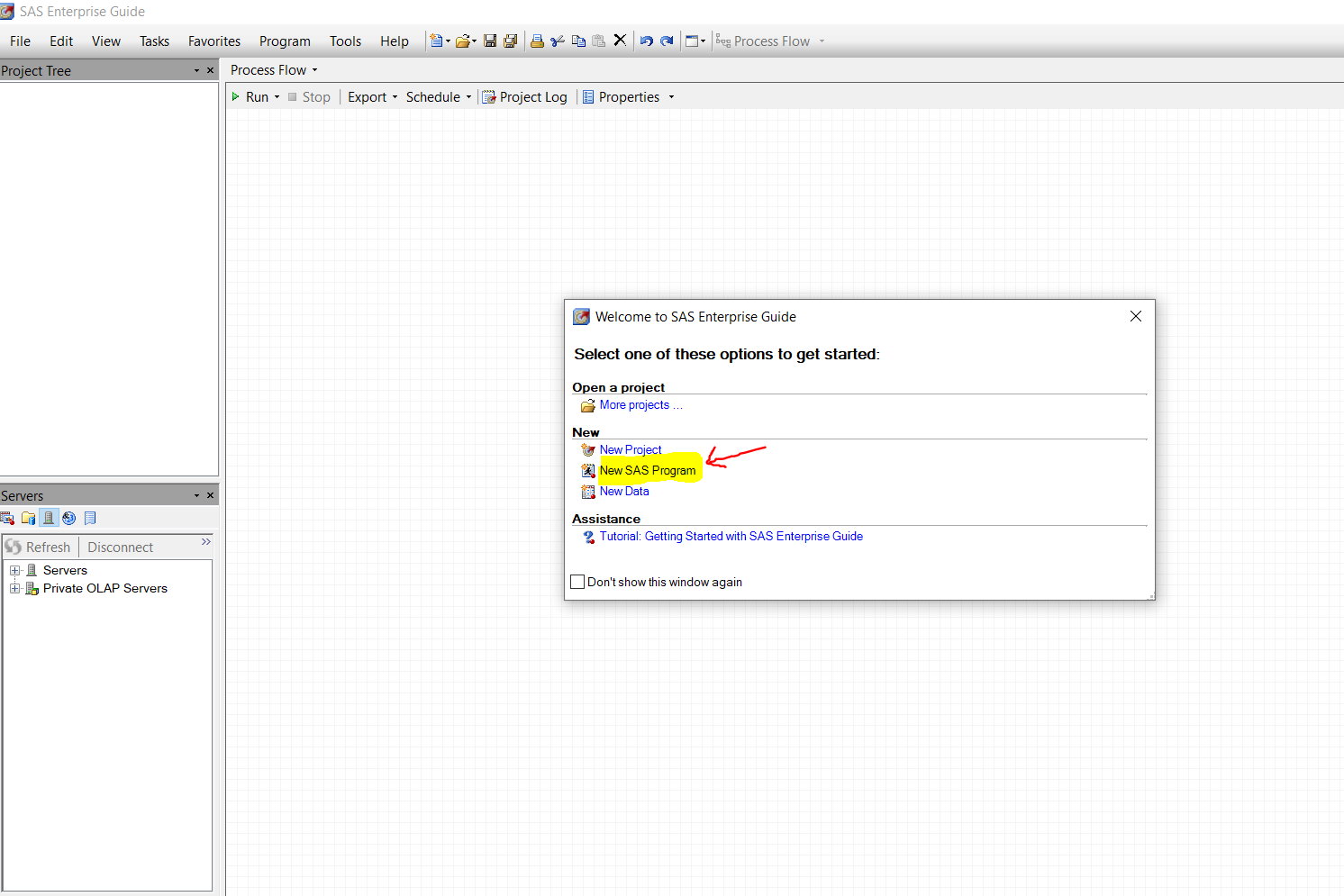
12- Using the data in Table 7.5, find the k-nearest neighbor for record #10, using k = 3.

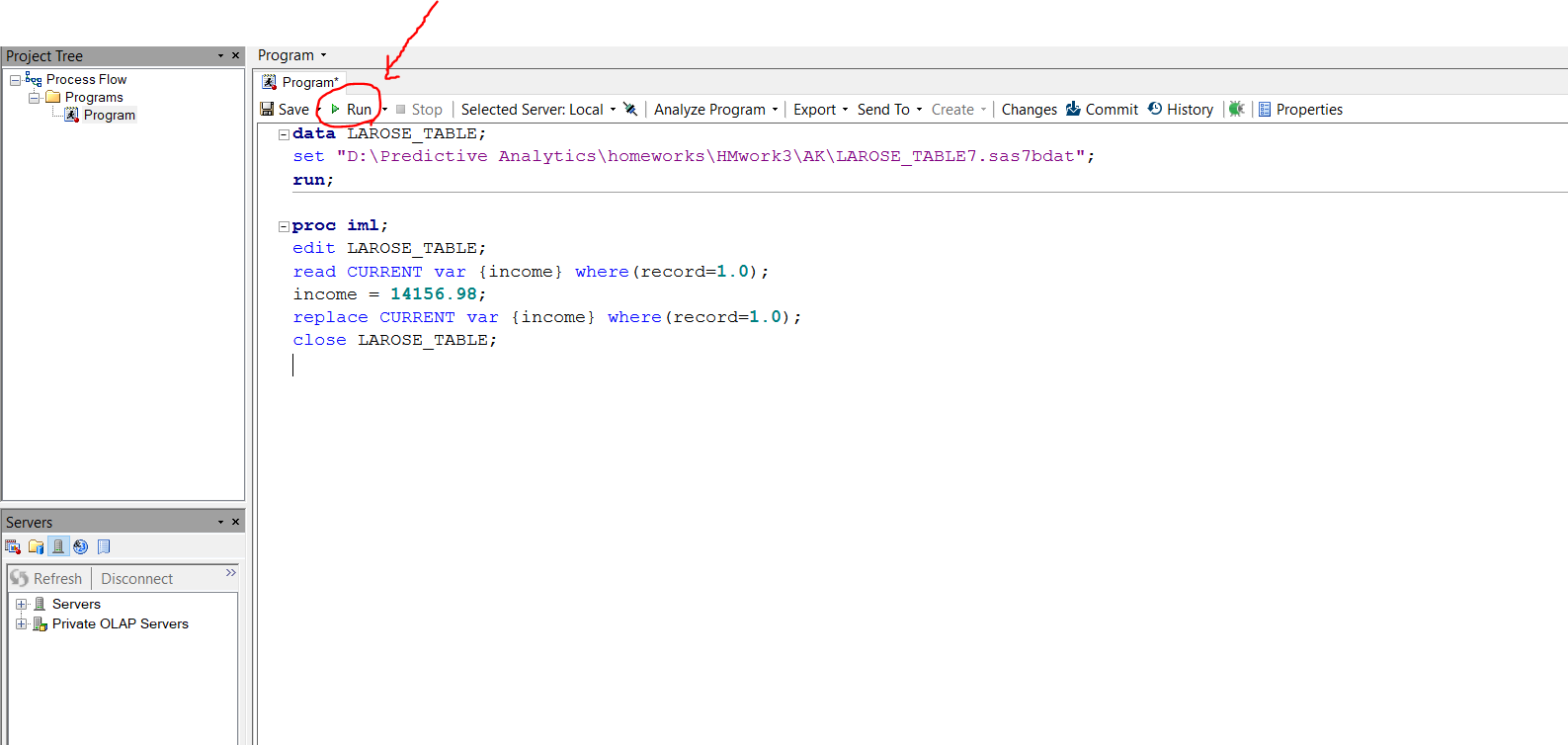
By comparing the given Table and the table 7.5 in the book, we’ll see that the

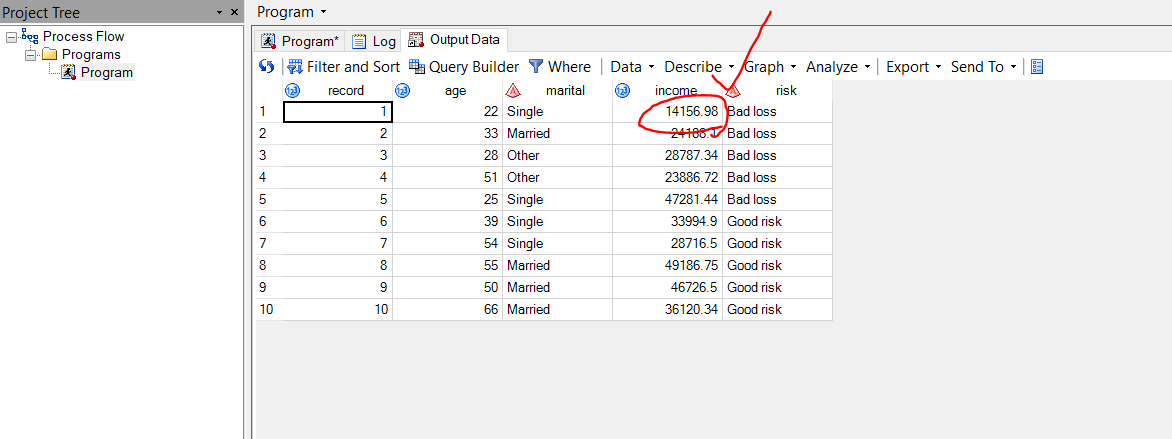
income for record #1 has a different number. In order to fix that first, I use the excel

file to create SAS table (Laros\_table7) in SAS Enterprise Guide then use the code

bellow to replace the right value for income for record #1. then export the table.







After exporting the table, we use SAS Enterprise Miner to find the K-nearest neighbor. After creating our diagram, we use data node and then statexplore to explore the data. And we used transform node and finally MBR node to find k-nearest neighbor.

