**4-4– Machine Learning 4**

Aside from the feature preparation tools that we’ve discussed in the past lessons, 2 other tools help clean and fit your data to be model ready. These tools work hand in hand with the assisted modeling tool and are automatically inserted into the workflow once you’ve applied the output of the assisted modeling tool. For this lesson, we are going to discuss how the transformation tool, the fit tool, and predict tool work in an assisted modeling setup.

The **Transformation tool** is used to perform these data-prep tasks: Set data types. Clean up missing values. Select features. And finally, encode data. You can perform each of those tasks in any order. The order depends on how you want to prep the data. Do take note that this tool only handles 1 task at a time, so if you need several of its tasks, you need to insert more than 1 transformation tool. The tasks are called “transformers” in the tool’s configuration.

To demonstrate, we will use the 4-1 workflow with the same assisted modeling setup that classifies milk into different grades. After clicking “Add Models and Continue to Workflow” on the assisted modeling window, we are presented with the complete set of tools that composed the machine learning pipeline. After the assisted modeling tool, 4 consecutive transformation tools are placed**. This tool is always placed between an Assisted Modeling tool and a Classification or Regression tool. Tru or False.** If there is no assisted modeling tool present before the tool, it will generate an error. The first transformation tool’s task is to “Set Data Types”. The UI for the configuration window also matches the step 3 tab of the assisted modeling window where we are required to set the data types for each feature. Since this tool has already applied the recommended data type, the features were already converted to Numeric and Boolean values. You can always change the type by choosing a different one from the dropdown. You can choose between Numeric, Categorical, Boolean, or ID. The next transformation tool cleans up missing values. The current configuration is set to “Replace with Mode” if there are missing values on the checked features. **Replace with mode** will replace null and missing values with the number or string that occurs most often. It is recommended to choose this option if a feature contains categorical values and you don't want to drop it. Other ways of replacing missing values are listed on the method dropdown.

**Replace with Mean** replaces the missing values with the average or the sum of all values in the column divided by the total number of rows. This method should only be used for numeric data and is recommended if your data is normally distributed and has no outliers.

**Replace with Median** replaces nulls with the number that is the midpoint of the feature distribution. This is best used for data that is skewed and contains outliers.

**Replace with Constant** reads empty fields as missing values. Select this option if you think the modeling algorithm could find meaning in the missing values themselves because sometimes it can find patterns in the absence of data. You can also select this option if you think other methods of handling missing data could bias your model.

The 3rd transformation tool is tasked to “Select Features”. From the name itself, it simply removes unchecked features so that they will not affect the model. The total number of features selected is also shown in the upper portion of the parameters section. The 4th and last transformation tool is set to “One Hot Encoding”. **One Hot Encoding** takes categorical variables and encodes them into numeric variables to make them suitable for modeling. Predictive models are more suited for numeric values (that includes binary and Boolean values) and this also makes categorical values unsuitable. In order to use categorical features,One Hot Encoding creates a Boolean response for each categorical value. For example, if you have a categorical feature called “Color” which contains 3 values “Red, Yellow, Green”. One Hot Encoding will convert it to 3 columns; Red, Yellow, and Green which will be tagged either 0 for false or 1 for true. For this workflow, since our features are not categorical, one hot encoding is not needed. You can also hide the features that do not need this task by toggling “Hide Un-encodable Features”.

The **Fit tool** is used to output a model. It is the final tool in a machine-learning pipeline. This tool has the job of taking input from your dataset and the other machine learning tools, then "fitting" the model to the data. Fitting a model entails teaching your algorithm the relationship between predictors and outcomes in order to predict future values of the outcome. To use the fit tool, it must be connected to either a regression tool or a classification tool. The tool is also sensitive to changes you make in upstream tools, including any Transformation tools.

From the workflow, we can see the fit tool connected after the classification tool. This tool is automatic and does not require any configuration, so it simply takes in the model and fits it to the data.

After the fit tool is the Predict Tool. The **Predict tool** is used to make predictions about new data using a machine learning pipeline you've built. The tool has 2 input anchors and 1 output anchor. D input anchor is connected to the test data, in this case, the Validation output of the create sample tool. M input anchor is connected to the model. This takes in input from any model objects which can either be the output of a Regression tool, classification tool, image recognition, topic model, or named entity recognition tool. The output of the predict tool has the predicted value and the probability of each value. For example, in this classification model, the predict tool appended the predicted value under “Grade\_predicted” and it also included 3 new columns which show the probability that the data will be classified into a specific grade. Our first row is predicted to be almost 94% probable to be High Grade and .5% probable to be Low Grade.