**4-1– Machine Learning 1** 

Aside from data preparation, Alteryx has also started innovating their AI and Machine Learning toolset by releasing the Intelligence Suite. The **Intelligence Suite** offers several features which ranges from improving the range of data that you can import by including the option to import Pdf and images, to no-code advanced analytics guided by assisted modelling. This is great news specially for data enthusiasts that have just started learning about ML models and analytics. Machine Learning by itself can be pretty taxing and calculation heavy, but with a tool that guides you every step of the way, creating an ML Model can be a walk in the park.

**Assisted Modeling tool is used to create a machine-learning pipeline. In the pipeline, you select a target, set data types, clean up missing values, select a machine-learning algorithm, and train the model. After that, you can connect your pipeline to the Predict tool, which allows you to make predictions. Assisted Modeling tool has 2 options. Assisted & Expert Modeling. Use the Assisted option to get help building machine-learning models. It guides you through a step-by-step process, which includes selecting a target and machine-learning method, setting data types, cleaning up missing values, choosing features, and selecting the best algorithm. If you're unsure what algorithm is best, Assisted Modeling allows you to compare algorithms in the leaderboard. Use the Expert option to build models without assistance. All you have to do is select a target from the dropdown, then you can connect other Machine Learning tools to build the pipeline.**

We are going to demonstrate how assisted modelling can be applied to this dataset which contains milk characteristics. It consists of 7 independent variables, pH, Temperature, Taste, Odor, Fat, Turbidity, and Colour which will be used to predict the Milk grade of either high, medium or low. We’ve already inserted a select tool to check the fields available as well as a Create Sample tool that takes 60% of the data for training the dataset and the remaining 40% for testing. From the Machine Learning toolset, drag an Assisted Modeling Tool and connect it to the E anchor of the create samples tool. Before using the tool, it is essential to run the workflow first for it to recognize the new data. Once done, you can now select from either Assisted or Expert Modeling. Select Assisted and “Start Assisted Modeling”. The new window will take you through each step of transforming your test data into a model, with instructions and examples to help you decide the best configuration. Click the “Start Building” button to proceed with step 1. The first step is to select the target variable and the ML method. The fields available in your dataset are listed under “Available Targets”. You need to select the “Target” or the column you want to predict. In this case, we will point it to “Grade” since it contains the milk classifications. Once selected, the available machine learning methods will show up. For the milk grade, we can choose between Classification or Regression.

**Classification predicts a discrete class label. Meaning, the target field should be categorical or a string. In the assisted modelling window, alteryx gave 2 examples.1: You can use classification if whether or not a customer will buy a product with the values being yes and no. 2: Categorize credit risk where the values might be high, medium, or low. Both examples have discrete type of class labels, which were yes/no or high/medium/low. Regression predicts a continuous quantity. Meaning, the target field must be a numeric field which can be an integer or with decimal values. The 2 alteryx examples given here is 1: You can use regression to predict the response a patient will have to a drug, with the response measured on a scale from 1 to 10. And 2: What the stock price will be for a particular stock. Both examples show that the target variable for regression can be a numeric value within a specific scale or without.**

**Since our milk grade dataset has the class label of either High, Medium or Low, it is considered a multiclass classification so we are going to select “classification”.** Hit Next to proceed with step 2. This prompts us to verify whether we really want to use “Grade” as our target variable. Click continue to go to step 2. Step 2 lets us select the automation level. **Step-by-Step** shows you the whole process, while **Automatic** will make alteryx build the whole model and will not show you steps 3 to 5. Let’s select step by step so we can see how each process works. Hit next to proceed to the next. Step 3 automatically sets the data type. From the select tool earlier, all of our fields were originally strings, but alteryx recommends to convert them to Numeric and Boolean. You can always change the recommended data type by using the dropdown. **Numeric for integer, double and float, Categorical for strings with limited number of values, ID for columns with unique string or numeric characters that are intended as keys, and Boolean for columns that has one 2 values, usually true or false. You can check each feature or colu**mn’s contents by selecting one and navigating to its column details on the right side. The upper portion contains the data type probabilities, while the lower section has the preview of its data. If you are a bit confused as to which data type to use, you can always refer to the glossary tab. All of the recommended data types are correct so hit next to go to the next process. For step 4, alteryx simple checks and cleans any missing values on the features or columns we are going to use. This is essential as nulls can directly affect the outcome of our model. Since this dataset is already preprocessed and clean, no missing values were found. Hit next to go to step 5.

Step 5 is where you select the feature**s. Feature is the term used for the fields or columns that you can use in order to predict the target variable.** Alteryx automatically checks each and every column available and measures whether it is a good predictor or not. Simply uncheck a feature if you don’t want to include it as a predictor in your model. Selecting a feature also shows its predictor details on the column details tab. All features are tested using the Gini and GKT to know if the feature associates too much or too little with that of the target column. **Gini Impurity measures each feature’s contribution to the whole model, while GKT or Goodman-Kruskal Tau measures the association or how well it can predict the target.** You can see the amount of Gini Impurities and GKT under the predictor details. If the feature falls within a specific range, it will be classified as a good predictor. All our 7 features are classified as good so we can proceed with the next and final step.

Step 7 lets you select the algorithm to be used. Each algorithm is listed with its own pros and cons to help you decide which ones to keep and which ones should be left out. The list of algorithms will change depending on what kind of dataset you’ve used, as well as the target variable. We want to see how all 3 performs so let’s click the “Run Selected Algorithms” to generate the leaderboards.

Once all models are finished, its accuracy and comparison are shown on the leaderboard window. According to the results, our best model in terms of accuracy is the first XGBoost followed by the decision tree. To view more information about a specific algorithm, select one of the models and navigate to the rest of the tabs. Overview shows individual performance by displaying the exact accuracy or the number of correct and incorrect output as well as its confusion matrix. The Interpretation tab shows what features the algorithm valued the most and the features that can be removed or negligible. The Configuration tab shows all of the setting from step 2 to step 6.

If you are satisfied with the results, you can now apply the model to the workflow. Check the box for XGBoost 1 and click the button “Add Models and Continue to Workflow”. The workflow contains a new set of tools from the machine learning toolset that was automatically applied by assisted modeling. Each tool contains annotations that you can refer to understand how the process works. Each tools configuration matches the options that we’ve selected upon using the assisted modelling tool. We can now test the applied XGBoost model using the rest of our data. Connect the V output anchor of the create samples tool to the D anchor of the predict tool. Once done, run the workflow. The output anchor of the predict tool has 4 new columns. Grade\_predicted for the grade output of the model, Grade\_high which shows how likely the milk is high grade, Grade\_low which shows how likely the milk is low grade and Grade\_medium which shows how likely the milk is average or medium grade.

Assisted Modelling makes Machine Learning accessible even for beginners. In order to use the 3 tool palettes from intelligence suite, you need to have an active Alteryx Intelligence Suite License, apart from the license of your alteryx designer. The suite also has the same system requirements to that of alteryx designer. Before installing, make sure that you have downloaded the same version that matches your current alteryx designer. For example, if you have Designer version 2021.3, your Intelligence Suite version must also be 2021.3, and so on.