## Training Methodology

In this project, first the Adult Census Income dataset (available in Azure) is uploaded.

Subsequently, in the next step, appropriate features (age, work class, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income) suitable for training were chosen using ‘Select Columns in Dataset’.

Subsequently three nodes- “Summarize Data”, “Edit Missing Data”, and “Filter-based Feature Selection” were created.

After appropriate features were chosen, the applicable data was summarized and downloaded in CSV format. Some of the features which had missing values were replaced with mode values (since it’s a classification features) using “Edit Missing Data” function in Azure ML.

To observe the features having large impact on the label (Income), Chi-square test was carried out and visualized using in-built feature in Azure ML. Chi-square test indicated a relationship in the following manner with respect to the income: relationship > marital-status > education-num > occupation > age > hours-per-week > capital-gain > sex > work class > capital-loss > race > native-country.

From the “Edit Missing Data” node, a “Split Data” node was created for splitting data existing data for “training” (70%) and “testing” (30%) using the classification regression algorithms (Logistic Regression, Naive Bayes Classifier, Boosted Decision Tree, Decision Forest, and Support Vector Machines).

Subsequently, all the algorithms were scored and evaluated. Based on the evaluation metrics (Accuracy, Precision, and Recall scores), appropriate model was chosen as model for predicting the incomes.

## Prediction Methodology

1. “Tarrant Dallas Tenton County dataset” was applied and transformed based on the training model described above using “ Apply Transformation” mode in Azure ML.

2. Subsequently, the transformed data is scored in comparison to the data used for Training using "Score Model" function and converted to csv using " Convert to CSV" function.