**Exploratory Data Analysis**

A best practice in the lifecycle of a machine learning model is to start with exploratory data analysis (EDA) with the objective of extracting information that unveils trends, feature correlation, group distributions, demographics, biases, etc. Specific to our research, our intent is to consider different scenarios with respect to bias. One scenario of interest is to explore if the data used in training our model contains any biases. Several techniques can be applied to identify biases in the data prior to training a model and metrics can be produced to measure data bias with respect to a protected feature, age, for example. If bias is found, there are several techniques such as reweighing, disparate impact remover and optimized preprocessing to mitigate or eliminate bias in the data. We are running 2 scenarios, applying pre-processing debiasing prior to running the ML model with adversarial debiasing and a second without pre-processing debiasing step.

The data set we are initially exploring is employee data collected from IBM to indicate attrition of an employee1. There are approximately 1500 entries with 35 features with ‘Attrition’ being the dependent, target feature. We are interested in the ‘Age’ feature specifically as the protected variable.

We performed the preliminary EDA on the data set as follows:

1. EDA was executed in Python language and libraries.
2. The Python Jupyter notebook is found in the project’s GitHub2
3. Data was extracted in Excel format and converted to CSV format
4. The data was loaded into a Data Frame to calculate simple statistics for all features
5. Explored the protected Age feature by finding unique age value counts and producing a bar chart to learn the distribution of the employees ages.
6. Calculated categorical frequency grouping counts to unveil trends and distributions, for example we found that 1,043 employees out of the 1,470 total employees rarely traveled. This type of frequency groupings is useful to unveil insights and important knowledge within the data.
7. As a simplification step, the levels on the ‘Education’ feature were reduced from 5 to 4 by combining levels 4 and 5, since level 5 had a reduced number of employees.
8. The data set was one-hot encoded to prepare the data as input to the:
   1. ML model and removed original features
   2. Correlation Matrix function
9. A correlation matrix between features is also provided for analysis purposes
10. A Matplotlib graph chart of the correlation matrix was built for easy visualization
11. Built a preliminary ML model based on the XGBooster algorithm to explore preliminary classification results

According to IBM’s Artificial Intelligence Fairness 360 toolkit and framework, there are three (3) main points in the lifecycle of a ML model where bias can enter the model and where one can intervene to control bias, see figure 1 below.

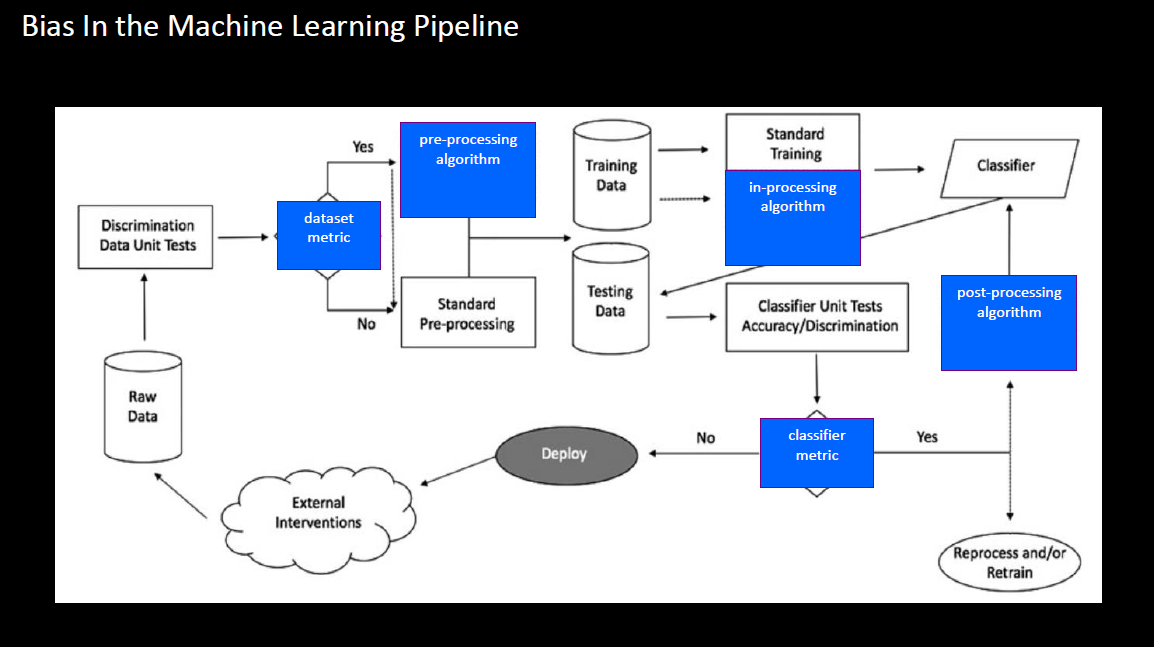


Figure 1 – courtesy of AIF3603

**References:**

1. <http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html>
2. Project GitHub: <https://github.com/skhayden/SMU-Capstone-Age-Bias-in-Predictive-Modeling->
3. <file:///C:/Users/solan/Downloads/Removing%20Unfair%20Bias%20in%20Machine%20Learning%20Webcast%20Slides%20-%20IBM%20Community%20-%2011.13.19.pdf>