**Project: Capstone Project 2: Project Proposal**

**Mechanical Fitting Failure Classification**

****

**Problem Statement:** Code of Federal Regulations (49 CFR Parts 191, 192) requires gas distribution pipeline operators to submit reports on an annual basis of all hazardous leaks that involve a mechanical fitting (DOT Form PHMSA F-7100.1-2).

Our goal is to classify mechanical fitting failure in the gas pipeline so that we can identify in- advance reason for the leak, when it happened and may be how often since installation? If there any fittings that failed during the first year of operation, then that issue could be either design related or material related from the manufacturer, which should get reported to CFR.

**Benefit to Client:** This problem is relevant to all pipeline operators who are responsible for transporting oil or gas to various locations using their pipeline network safely, reliably and long-term.

By training a model, if we could identify in advance which mechanical fittings leak frequently, what causes the leak to happen and after how many years of operation the leak usually occurs, then this will help our client to adjust maintenance/ repair schedule sooner and on specific fittings that are impacted. It will also allow client to engage manufacturers to research and develop improved fittings that offer better life and provide advance nortification to pipeline operators before a fitting leaks.

**Data Source:** <https://www.kaggle.com/binovi/mechanical-fitting-failure-data>

The dataset is available on Kaggle in csv format.

* **Problem Solving Approach:** This dataset is entirely in text format. I will need to transform the data into numerical values as well as replace missing values such as manufacturer information, cause of leak with information available in in other columns. I will also utilize dates to understand difference between installation, failure, reported and filing dates.

Once, the data is cleaned and convreted into numerical values, I will apply various classification algorithms to see which model can identify leak accurately based on the information we are able to model.

* **Project Deliverables:** I will submit a notebook showing data pre-processing steps, model evaluation and selection reasoning. A detailed project report and a slide deck to discuss steps taken and communicate results to the client.