**PROJECT REFLECTION**

Name – Akashdeep Singh, Vanshnoor Singh Narang

1. **Business Problem**

Given the NPRI dataset, our objective is to predict for the companies that are a part of an given industry that are expected to experience the highest quantity of releases for any substances over the next year.

1. **ML Problem Definition:**

In this scenario, the goal is to predict the quantity of substance releases for a given set of industries over the course of one year. The target column for prediction is the quantity of releases. Our approach involves an aggregate strategy to determine the change in quantity of releases for substances within a predefined set of ten industries.

Strategy-

To calculate this change, we'll employ a straightforward method: subtracting **the quantity of releases in the previous year from the quantity of releases in the current year**. This difference will indicate whether there has been growth or decline in the quantity of releases for each substance within the specified industries. If the result is positive, it indicates growth; if negative, it signifies a decline.

1. **Feature Requirements:**

The following features could have an impact on the target variable:

Category - Sub-Category - Current - NAICS - ReportYear - IssueType - PriorityYears - Years

Minus 1 - Years Minus 2 - Years Minus 3 - Years Minus4-PercentageModification -

DistinctivenessVariation - Four-Year Average - NAICSAverage Substance, Standard

Deviation, NumStandardDeviation.

1. **Does it make sense for the target value to be affected by what you listed? Explain. If you answer with yes, argue why?**

These features provide context about the industry, time, and specific substances being released. They are relevant because they help us understand the historical trends and variations in releases over the period from 2014 to 2022.

They make sense because they show **positive correlation** and can help identify patterns, outliers, and trends that impact the quantity of releases.

1. **Based on the answer to the last question, if some things do not make sense, how can you construct variables that do make sense in being used to make predictions about the (different) targets?**
2. **How should you frame your problem then?**

According to me, I have already reframed the problem so there is nothing to do with that.

1. **How are you planning to assess your results? Be minute, detailed and exact.**

In this project, we'll use **linear regression** to forecast substance release quantities over time. To evaluate our model, we'll use R-squared (R2) to measure the proportion of explained variance and root mean squared error (RMSE) to quantify prediction accuracy. These metrics will help assess the model's performance in capturing trends and making accurate forecasts.

R-squared (R2): R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (quantity of substance releases in our case) that is explained by the independent variables (features used in the model). A higher R2 value indicates that the model effectively captures the variability in the data and provides a better fit.

Root Mean Squared Error (RMSE): RMSE is a measure of the average deviation of the predicted values from the actual values. It provides a measure of the model's accuracy by quantifying the difference between predicted and observed values. A lower RMSE indicates that the model's predictions are closer to the actual values, implying better performance.

1. **Does your assessment make sense? If not, how can you update it?**

Yes, the assessment makes sense. R-squared (R2) and root mean squared error (RMSE) are commonly used metrics for evaluating the performance of regression models.

We can also try different models such as random forest, KNN regressor which can show better accuracy. Even we can update feature selection and do normalization to update are successful model.