**Topic** –

Predictive algorithms for different subjective aspects of questions-answers

1. **Introduction**

Computers are really good at answering questions with single, verifiable answers. But humans are often still better at answering questions about opinions, recommendations, or personal experiences.

Humans are better at addressing subjective questions that require a deeper, multidimensional understanding of context - something computers aren't trained to do well…yet... Questions can take many forms - some have multi-sentence elaborations; others may be simple curiosity or a fully developed problem. They can have multiple intents, or seek advice and opinions. Some may be helpful and others interesting. Some are simple right or wrong.

**1.1. Problem Statement**

Google Quest Q/A labelling is a Kaggle competition where participants are expected to develop predictive algorithms for different subjective aspects of question-answer from Stack Exchange. Here, Subjective aspects are human-labelled to reflect whether the question-answer is well-written, relevant, helpful, etc.

**1.2. Dataset Overview**

Unfortunately, it’s hard to build better subjective question-answering algorithms because of a lack of data and predictive models. That’s why the [CrowdSource](https://crowdsource.google.com/) team at Google Research, a group dedicated to advancing NLP and other types of ML science via crowdsourcing, has collected data on a number of these quality scoring aspects.

* 1. **Real-world/Business Objective and Constraint**

This model will help the questioner clearly frame the question based on the scoring from 21 aspects calculated from the question title and body. In addition to this pop, the best answer related to the question and its intent, based on the other 9 scorings.

The latency of the model is moderate to high.

* 1. **Evaluation Metrics**

Submissions are evaluated on the mean column-wise [Spearman's correlation coefficient](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient). The Spearman's rank correlation is computed for each target column, and the mean of these values is calculated for the submission score.

1. **Dataset Details**

**Train data** has total of 6079 rows and 41 columns where 11 columns are features and 30 columns are labels.

**Test data** has a total of 11 columns only for features.

The data is taken from the [Kaggle competition](https://www.kaggle.com/competitions/google-quest-challenge/data) which includes questions and answers from various StackExchange properties. Our task is to predict the target values of 30 labels for each question-answer pair. Target labels with the prefix question\_ relate to the question\_title and/or question\_body features in the data. Target labels with the prefix answer\_ relate to the answer feature.

Each row contains a single question and a single answer to that question, along with additional features. The training data contains rows with some duplicated questions (but with different answers). The test data does not contain any duplicated questions.

Target labels are aggregated from multiple raters, and can have continuous values in the range [0,1]. Therefore, predictions must also be in that range.

1. **Proposed Solution**

We will select the 30 best models for each of the target labels, after experimentation between the traditional supervised learning algorithms namely KNN, SVM, Random Forest, and XGBoost.

Finally, will combine all the 30 model’s output.