## Capstone Project Report

The data is fully in hand.

EDA

The data is a classification type data with its **target variable being categorical**. After uncovering the distribution of the target variable, It is discovered that that the data is a **multi-class classification type data** which consist of **‘Yes’ , ‘No’ and ‘Maybe’**. There are a decent amount of missing value in the dataset and all of it seems to be data about whether the person is in a self-employement or not. Through EDA I have discoveed that instead of doing the normal classification , I need to do a multi-class classification ways of modeling which require a tweak and adjustment on the modeling which difer from usual. The EDA is probably already fully completed if I didn’t make any mistake.

Modeling

I have already begun my modeling process. At first I **started with logistic regression model** thinking that this model is the most suitable for this project. However, after fitting the model and doing evaluation , the **accuracy of the model is only a meagre 0.355** which is abysmal for something like my project. Thus I **switch to a different model called Random Forest**. After fiting and evaluating the model , **the accuracy of Random Forest is as high as 0.976** which is more preferable. It is noted that this is maybe not the most optimal perfomance as it has not done any hyperparameter tuning and bootstraping and doing this may increase the perfomance more in the future.

Blockers

As this is related to people health, the most important blocker is probably the value of accuracy and its ability to predict on unseen data.

Status , Timeline and Questions.

-No changes in topic happened since my lightning talk. I do believe I have the necessary data in hand.

-In this week and a half I would like to finish this project. I probably need to optimised my model in this week and a half.

-Correctness of my model. How to improve my model more.