Original email:
Hi Team,

I am Harshwardhan Yadav, you can call me Hersh. So, I am your assigned TA for the project. I went through the proposal and believe it has great potential. Still, there are a few things I wanted to get a clear understanding of and some standard questions that we must answer, they are as mentioned below:

1. To what extent have they clearly articulated a problem (i.e. TEP definitions or the like)?
The goal seems to be clearly defined i.e. to find the precipitation by month. The problem seems to be a standard ML problem of determining output based on the input features. It would be great if an explanation about the line “Since the existing models were to help predict flood or crop yield, the absolute rainfall mattered more than the percent error” could be provided.

2. To what extent are the extensions they defined clearly articulated and feasible?
The contribution defined in the project is taking an open-source project and aiming at creating a model to provide better accuracy than them. This is a valid contribution but I believe if you could compare it with any standard research work then that would be nice. Please make clear to me if there are no such works done in the field. Again, beating their work is not what we are looking for, we want a genuine attempt at it. I am still very confused about the second contribution that the team has proposed. It would be nice to get more clarity on the second contribution based on which the team can be graded. The project seems feasible to me since there are no GPU requirements as the team is aiming at traditional ML algorithms for the project.
Since acquiring the data is hard, I want to know if the scope can be increased from India to some other country’s data as well.

3. What are the strengths of the proposal?
a. Good Ethical considerations of the project
b. Computational feasibility
c. A clear framework of what the team wants to achieve

4. What are the weaknesses or risks of the proposal?
a. Second contribution is vague or not given
b. Some clarity is required on their first contribution

Reply:
For point 1. There is a threshold of rain that needs to occur in order for crops to be watered. If it does not rain enough then other methods may be utilized in order to water crops. This will take time, energy, and money. If the farmer knew that it would rain above this threshold for a given month they would not have to spend as much time preparing watering systems. (They would care less if it will rain one vs two inches. (Where this would be a 100% error if they were expecting one inch and it rained two.)

On the other hand of the scale, if it rains over a certain “flooding threshold” then citizens will need to know in advance so that they can prepare accordingly. If it will flood, there is a range of rainfall that will make citizens make the same preparation decisions. For example, if the flooding will not impact homes they may stay home. However, if the flooding is so bad that homes may be destroyed they would want to evacuate. Therefore, we will be using mean absolute error (MAE), and RMSE to evaluate our models.

<u>For point 2:</u>
Aside from open-source projects that we found. There is a paper,
“Mohammed, M., Kolapalli, R., Golla, N., & Maturi, S. S. (2020). Prediction of rainfall using machine learning techniques. International Journal of Scientific and Technology Research, 9(01), 3236-3240.”
that attempts to predict rainfall. They use PCA, and then SVR in order to predict rainfall. They predict monthly, seasonal, and yearly rainfall.

Another paper:
“Singh, P., & Borah, B. (2013). Indian summer monsoon rainfall prediction using artificial neural network. *Stochastic environmental research and risk assessment*, *27*, 1585-1599.”
, uses a neural network to predict rainfall. They discretize by year however.

<u>Further explanation of second point:</u>
We are laying out what machine learning methods and techniques we will use in order to make our predictions. We will use a few machine learning methods, and possible ensemble methods in order to make predictions. We will start with linear regression with and without regularization. Additionally, we expect to use a random forest method, specifically something like sklearn’s RandomForestRegressor. For the open-source projects there are some key techniques that were not utilized that we could implement in a straightforward manner such as k-fold cross-validation. We believe we do not need to go through the effort of using a neural network in order to predict monthly rainfall and that simpler statistical methods will work.

Last sentence of point 2:
As we are in the process of getting enough data, if we feel that we do not have enough data for India, we could expand to regions in China, as it appears there is also data available there. However, as it stands there appears to be enough data in India.