Project Proposal 

#### *Rohit Roy Chowdhury*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | Healthcare system is becoming more expensive day by day, in this scenario this product will help not only doctors but also patients to diagnose pneumonia in children from x-ray images on their own in few seconds.  Machine Learning in this product is going to reduce the cost significantly & brought down all the hassle & time to get if the symptoms are positive or not. |
| **Choice of Data Labels**What labels did you decide to add to your data? And why did you decide on these labels vs any other option? | Here in the Appen platform I choose three data labels to help the annotator to get the job done fast. They’re following:   1. Yes. 2. No. 3. Maybe.   This minimal approach to data labeling will make the data annotation job fast but may also arise confusion in some cases. |

# Test Questions & Quality Assurance

|  |  |
| --- | --- |
| **Number of Test Questions**Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job? | I developed 13 test questions to prepare for launching the medical annotation job. Considering 3-4 examples for all possible scenarios. |
| **Improving a Test Question**Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question? | I’ll Improve & consider simplifying the rules section for the annotator. |
| **Contributor Satisfaction** Say you’ve run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.) | Considering given results I’ll look into –   1. Improve the steps & rule section then, 2. Add more examples for all different scenarios. |

# Limitations & Improvements

|  |  |
| --- | --- |
| **Data Source**Consider the size and source of your data; what biases are built into the data and how might the data be improved? | The data size don’t have such bias as far my concern but it should contain enough data to gather chest x-rays from children from all demographics. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | I’ll provide more diverse cases of pneumonia symptoms in the examples & elaborate the test questions in depth to get most out the data labelling platform. |