Project Proposal 

#### *<Maha Mohammed Alshamrani>*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | Medical Industry. In this task, we help doctors diagnose pneumonia, the use of machine learning here helps doctors quickly identify cases where pneumonia is not present and spend more time on symptomatic cases. |
| **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? | I used the three labels "Yes", "No", "Unknown". The first two Yes and No labels were chosen because we need to decide if there are pneumonia symptoms given the image or not. The third label "unknown" is chosen to leave room for uncertainty. |

# 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 11 test questions. |
| **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? | Rephrase the question to remove any ambiguities that might be present. Take a step back and see if the rules specified are clear and unambiguous. Also provide a more detailed description so that the annotator knows why it was labeled the way it is. |
| **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.) | Provide more questions for each label and try to see if the rules and questions stated are clear and unambiguous. |

# 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 size of the dataset is not large enough for a machine learning model to learn patterns. If we need to deal with more possible scenarios, then we must need more data. Anyways, if there are biases in the dataset you need to account for it either by augmenting the class that does not have more labels or throwing away some data from the class that has more data. The data source should also be more diverse and have more variety like images with different lighting conditions, illuminations, cropped, etc. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | Keep updating and improving the data, test questions can be improved as you come across new data, Rules and Tips might also need to be updated to reflect that. |