Project Proposal 

#### *<Chidimma Dominic Nweke>*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | The industrial problem which I am trying to solve here is that of providing a means to help doctors to quickly and more accurately diagnose cases of pneumonia in children.  ML is used to solve this task because it provides a more accurate means to predict outcomes without being programmed explicitly by using historical data. |
| **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? | The labels used in this project are as follows:   * 0 – Normal * 1 – Pneumonia * Not sure   These labels are used because we want the model to be more accurate in flagging down positive cases of pneumonia and also reduce the uncertainty which might be created by incorrectly labeled and unclear x-ray images. |

# 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? | Eight test questions were created to prepare for the launching of this annotation job. |
| **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? | First, identify what is going wrong in the question, augment the job to improve the clarity, include more and diverse examples to capture all possible cases, and possibly, redesign the job if necessary. |
| **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.) | Test Questions and Instructions needs to be improved. |

|  |  |
| --- | --- |
| **Data Source**Consider the size and source of your data; what biases are built into the data and how might the data be improved? | Considering the source and size of this data, the most common bias that is built into it is that of the patterns and distribution of normal and abnormal cases.  For the data used in this annotation job, it is biased because the size is too small and possible is not a good representation of the study space. This bias can be mitigated by added more data points. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | To design for longevity, a dynamic model should be used so it continuously trains on new data and keeps learning.  The data is a medical data so there are chances that new cases which are slightly different from the existing ones will arise, hence the need to change or update the annotation job and data to include more relevant definitions or examples. |

# Limitations & Improvements