Project Proposal 

#### *Magido Mascate*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | In this project we at helping **Smallholder Farmers** to quickly identify **Maize Plant Diseases** and find recommended Pesticides and Treatment Instructions offered by local Co-Operatives.  Plant diseases diagnosis process is suitable to apply **Computer Vision Solution** as there are significant stock plant leaf images either as proprietary and public repositories to train and test Machine Learning Models. Also, there are significant research resources from scholars and professionals, and NGO[[1]](#footnote-1) in Agroindustry.  Actually, we have access of **public Plant Leaf Dataset** account **up to 2000 images**, and we plan to have more from **End-Users**. |
| **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? | We decided to implement a Classification Systems, which can support mostly binary independent variable; either a Plant Leaf is (**NORMAL**) or Infected (**INFECTED**). All uncertainties are strictly flagged as **SUSPICIOUS**. |

# Test Questions & Quality Assurance

|  |  |
| --- | --- |
| **Number of Test Questions****Business Requirements:**Number of Test Questions must bind to Pareto (80/20) Principle. | We plan to prepare and submit 200 Test Questions – representing up to **10%** of actual images’ dataset, in **Four** separate **Annotation Jobs**.  However, we start with a minimum of **20 Test Questions** per Job while ensuring **Cost Overrun Control** with our Minimum Viable Product (MVP).  Every Job shall ensure a minimum of **20% Test Questions Balance** between Classes: **NORMAL**, **INFECTED**, and **SUSPICIOUS**. |
| **Improving a Test Question****Business Requirements:**Test Question which almost 10% of annotators missed, statistically, further steps shall be taken to improve or redesign this question. | To ensure a High-Quality Annotations Results, all four (4) Jobs are reviewed by Expert Matter Personnel and Improvement of Annotation Ontology will be handled accordingly.  Failing annotation jobs will be improved and resubmitted to annotators. |
| **Contributor Satisfaction** **Business Requirements:**  Rating results from annotations on instructions and test questions are must be **3.5 or above**. Otherwise, Instruction document or test questions must be reviewed for improvement. | This project as any other project in our company is driven with Total Quality Management Practices, so when required we’ll:   * Improve the Instructions and Test Questions * Re-submit Test Questions to Annotators |

# Limitations & Improvements

|  |  |
| --- | --- |
| **Data Source**What biases are built into the data and how might the data be improved? | Considering the source of the data we could **Human Introduced Bias.**  Factors such as Camera Resolution, light, shadows, and Images formatting could Impact on Plant Leaf Image Quality.  Also, future images uploaded by End-Users (Farmers) can be **Very Bias** for Future Model Training. |
| **Designing for Longevity**How might we improve our data labeling job, test questions, or product in the long-term? | To ensure a better data labeling, test questions We shall apply different approaches for Data Collection, Annotations. For test quality, its import to recruit Expert Matters to supervise the outcome.  Also, the system shall have a mechanism to strict uploading Images with Specific Quality and Resolution. |

1. NGO = Non-Government Organization [↑](#footnote-ref-1)