Capstone Project Proposal

Magido Mascate

Business Goals

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	Our company will be developing a **Plants (Maize) Diseases Diagnosis and Recommendations System**. We are adopting Machine Learning (ML) Models as core of engine of our systems. Computer vision capabilities in Artificial Intelligence (AI) Technology are vastly adopted in other industries and we aim at leveraging this technology to support fast remote plant diagnosing. The design, development and implementation of Digital Disease's Diagnosis Systems is not new and industries such as Healthcare are excellent example in using AI for diagnosis.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	Maize is one of the crops that most impoverished families, around world, rely on to fight famine. Also, many NGOs[1] in agriculture sector around world are working hardly in line with **UNICEF** - fighting hunger due poor harvesting and climate challenges. However, **it is easier to prevent Plant diseases than preventing droughts**. Therefore, **we are providing a Cloud-based Plants' Diseases Diagnosing solution** –which by agriculture industry stakeholders in developing countries, can upload crop leaf images through Co-Operatives Communities Platform. **Our platform aims at 1.000.000 (one million) End-users - smallholder farmers - per year in the first five years**. End-users are charged through community co-operatives**, On Per-use Basis**. Through co-operatives, end-users are **recommended registered pesticide and local suppliers, and technician**, and final well managed yields are **recommended to registered supplies buyers**.

[1] Non-Goverment Organization

Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	We will be using ML/AI in **maize diseases identification and treatment recommendation**. Herein, we'll ensure **Fast farmers assistance** by community technicians. Also, we will ensure **High harvesting ratio** and **High-quality products**. Moreover, we ensure **Reduce Farming Costs**.

Success Metrics

Success Metrics What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.	Following are Expected Diagnosis Engine Metrics: **Percentage of Accurate Diagnosis** • Diagnosis accuracy shall be 80% or above **Recommendations Acceptance Rate** • End-user's recommendations acceptance rate shall be **4 stars or above** **Percentage of Diagnosis Completed Successfully** • Diagnosis Request Processed Successfully shall be 95% or above. **Percentage of Farming Costs Reductions** • Co-ops Reported Farming Losses shall be less than 2% per Square Meter.

Data

Data Acquisition Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	Our system relay on open-source and proprietary maize leaves images. We are collaborating with industry stakeholders to acquire reliable Maize Leaves Images around world, while ensuring no PII and sensitive data are included in our dataset. Data will be available on per-use base - mean - every time an End-user submit a request, we'll have an additional image in our stock. We have access of up to **6000 (six thousand) open-source Maize Leaves Images**, and we plan to acquire more ongoing basis.
Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	We are aware of potential **Human Introduced Bias** and **Sampling Bias**. Factors such as Camera Resolution, light, shadows, and Images formatting could Impact on Plant Leaf Image Quality. Also, ongoing images uploaded by End-Users (Farmers) can be very biased for Future Model Training.

Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels versus any other option?	Our Classification System will support mostly binary independent variable; either a Plant Leaf is (**HEALTHY**) or Infected (INFECTED). Infected leaves are flagged as (**RUSTY**) or (**BLIGHT**). All uncertainties (Unknow) are strictly flagged as **SUSPICIOUS**. Suspicious flagged leaves fall into Expert matters evaluation for further Diseases Categorization (such as physiological diseases) in Product Capabilities Incrementation. These labels are related to Maize (corn) common pathologies and symptoms. Relying on common pathologies will help in accelerating the systems adoptions and easy data acquisition.

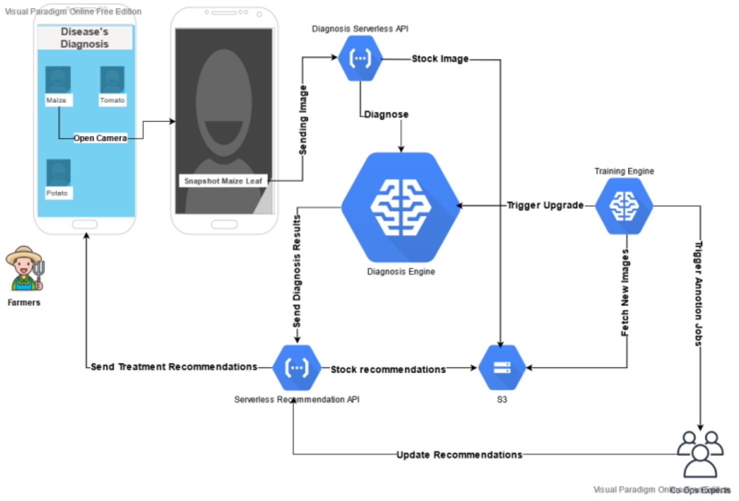
Model

Model Building How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	We are planning to outsource MVP model building and deployment. We are adopting **Google Cloud AutoML** Service to provide a faster and reliable **Prototype**, while shortening time for product-market-fit and proof-of-concept assessment. Alternatively, we our considering adopting **AWS Rekognition Service**. The product will be developed incrementally based on factual research results, and a continuous User Experience Testing. Complex Capabilities (**Features**) will be developed incrementally by in-house team through outsourced service, such as **AWS SageMaker** and **GC BigML**. Prior the building the ML model, we a planning to outsource the data annotation to **Appen**, and alternatively to **Google Data Labeling Service**.
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	Following evaluation metrics are relevant for MVP Model at **Threshold = 0.5:** • **Precision** at 90% or above • **Recall** at 90% or above • **Accuracy** at 90% of above (F1 Score metric) The indicated metrics above, will be monitored for any handled diagnosis request, and evaluated in-line with specified success metrics above. Whenever, the evaluation result fall bellow success metric, it will trigger a retraining workflow.

Minimum Viable Product (MVP)

Design

What does your minimum viable product look like? Include sketches of your product.



Use Cases

What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?

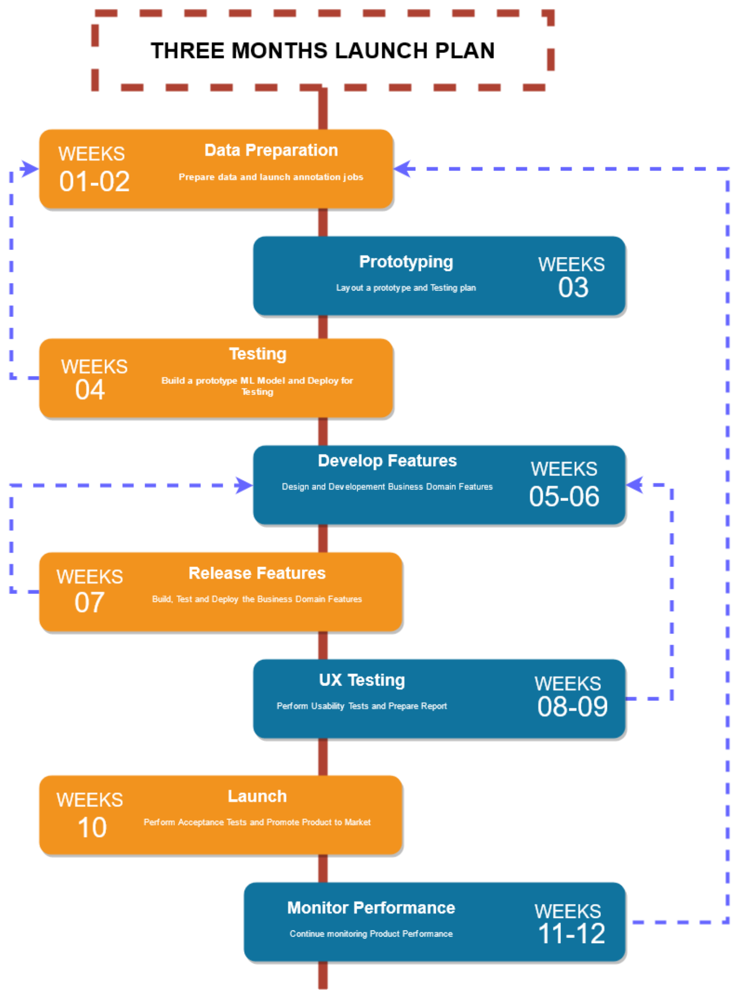
USE CASE 01: I am José and live-in remote area of Mozambique. I do use an affordable mobile phone with a camera, and access to internet. As farmer, I grow crops like, **Maize, Beans and Peanuts**. I work in collaboration with a community co-operative to receive periodic technical assistance. Whenever I need assistance on a problem, I'd like to share crops' images, so quickly figure-out what I can do – while waiting for one to two to weeks for a technician.

USE CASE 02: I am João, the local farmers community assistant. I live at main town 120 miles away and working **USDA** sponsored project. I assist 10 (ten) co-operatives with 150 farmers each, per season. While I can't be in all co-ops on same day, I'd like to jump into the co-ops platform to know what is going with farmers' plants and quickly find and give valuable knowledge-based guidance.

USE CASE 03: (MVP Future Incrementations) I am Mabida, **ZeroHunger Program Director**. Our project manages 50 pilot Co-operatives in fighting Famine, by ensuring increased production and access to market. We connect farmers with farm's goods suppliers and crops buyers. Also, our project, aim at assuring procurers have access to products with high-quality in a desired volume, any-time. Therefore, we would like to easily match well-managed crops to our procurers – through co-ops platform.

Roll-out

How will this be adopted? What does the go-to-market plan look like?


THREE MONTHS LAUNCH PLAN
WEEKS 01-02
Data Preparation
Prepare data and launch annotation jobs
Prototyping
Layout a prototype and Testing plan
WEEKS 03
WEEKS 04
Testing
Build a prototype ML Model and Deploy for Testing
Develop Features
Design and Developement Business Domain Features
WEEKS 05-06
WEEKS 07
Release Features
Build, Test and Deploy the Business Domain Features
UX Testing
Perform Usability Tests and Prepare Report
WEEKS 08-09
WEEKS 10
Launch
Perform Acceptance Tests and Promote Product to Market
Monitor Performance
Continue monitoring Product Performance
WEEKS 11-12

Post-MVP-Deployment

Designing for Longevity

How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?

Our product will be developed incrementally. Hence, we are planning follow to Quarterly Upgrade Release. Meanwhile, whenever, the success metrics fall short, will trigger a updates development process.

Success metric and Performance metric play an important role in triggering a retraining on new data. Also, retraining workflow will be triggered when, following threshold are reached:

- New Data represents 1% of dataset on which the model was trained previously.

Our company implements DevOps Practices in Software Development Lifecycle Process, and we're aware of A/B

	Testing. So, every feature update, upgrade will be handled through <=20% traffic to new version and >=80% to production version.
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	We aware of human and sampling biases in our dataset, especially when come to collect data from End-users. Therefore, we will implement a screening mechanism to validate collected data, so to ensure only valid data are stored and if possible curated prior using in model retraining. The validation process can be extended to Labelling Jobs, where we ensure the data is correctly categorized.