Project Scope and Deliverables | Fall 2023 AI Studio

This document is designed to help your team understand, internalize, and align on the scope, goals, and technical aspects of your AI Studio Project Challenge.

Complete all 4 sections as a team based on information you have gathered through:

* The project overview doc(s) provided in your team’s Project Folder in Google Drive (e.g., company video or slides);
* Insights gained during your first Challenge Advisor meeting during Bridge to Studio;
* Referring back to your Machine Learning Foundations summer course modules;
* Additional research done by your team related to the project/industry

Once you’re done, one team member should submit it through the assignment page in your AI Studio course in Canvas (“Business Understanding” module) by **September 3rd**. Your team’s AI Studio TA will review your submission and provide some initial feedback.

During your team’s first “Full Group” meeting during the week of September 4th with your Challenge Advisor (and AI Studio TA if they’re available), review your completed Project Scope and Deliverables document together and make updates / fill in any gaps as needed.

| **Project Title:** | We Inspect - The Home Health Impact |
| --- | --- |
| **Team Members:** | 1. Ayelet Kleinerman (She/her/hers) 2. Ilda Martinez (She/her/hers) 3. Tiffany Truong ((She/her/hers) 4. Max Ju (He/him/his) 5. Megan Kang (She/her/hers) |
| **Challenge Advisor(s):** | 1. Corey Levy, Co-Founder of We Inspect, [clevy@yesweinspect.com](mailto:clevy@yesweinspect.com) |
| **AI Studio TA:**  *(aka Tutor or Course Support)* | Swagath Babu (swagathb18@g.ucla.edu) |

**PART 1: PROJECT OVERVIEW**

**Project Description**

In your own words, what are you trying to accomplish? What type of ML problem is this? (e.g., “Supervised Learning: Classification”, “Unsupervised Learning: Clustering”, etc.)

| * Unsupervised Learning: Clustering   + Handling unlabeled data   + Trying to find similarities/patterns within the dataset * We are trying to identify any patterns or correlations between different mold types in people's homes with the symptoms people are facing and their diagnoses * The goal is to predict what diagnoses someone can have given the mold, symptoms present, and location (these can be interchangeable) * The goal is to predict what are the chances that some type of mold exists in a person’s household, and what type of mold it is, given the location and the health symptoms/previous diagnoses. If time permits, we would model given a mold and location, what are the possible symptoms to that mold type and given the mold and symptoms, what is the location. |
| --- |

**Purpose of Project**

Why is this project important or relevant to your AI Studio host company/org?

| * This project is important because it helps people find out what can be potentially causing their health issues. Through this project, our goal is to bring awareness to people that the different types of pathogens/mold within their homes may be affecting their health in some way. |
| --- |

**Ethics Considerations**

Are there any potential ethics-related considerations to take into account for your project?

| * Since we are dealing with people’s health, it’s our responsibility to provide the most accurate information.   + Avoid misdiagnosing, or misidentifying any molds * We should be transparent to users and make sure our research and models are interpretable by others as they should be allowed to understand how we arrived at our conclusions about their health and home. * Since we are working with personal, private data, making sure we understand what the model is doing is important for interpretability. * We need to be aware that the data we have is coming from people who were able to pay for the ERMI test, and who were willing to share this information voluntarily based on their test, which may indicate different socioeconomic status; It may also indicate susceptibility to health conditions. * We need to be aware of class imbalance, especially since we are dealing with many types of mold and many locations. These imbalances may be correlated to different socioeconomic status, or even specific genetic dispositions. * Ensure we have permission to use people’s medical data |
| --- |

**PART 2: PROJECT SCOPE**

**Project Requirements**

What is your Challenge Advisor expecting your team to deliver by December? Are there specific algorithms that you might use as part of model training/testing? (e.g, Linear Regression, KNN) How might you evaluate your model(s)? (e.g., F1 Score, RSME)

| A machine learning model for estimating the probability of mold presence, determined by a combination of location and symptoms. Before any specific models can be created, we will work in an unsupervised manner to find clusters. Since our label is going to be a continuous value with numerical and categorical features, currently we believe that Random Forest Regression would be a good fit. MSE and RMSE are good tools for evaluation of the model. |
| --- |

**Python Libraries**

What Python libraries do you expect to use? (e.g., Pandas, NumPy, Scikit-learn, NLTK)

| The Python libraries we expect we would use are Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn |
| --- |

**Other Resources**

What resources (e.g., online forums, recommended research papers, example code) does your team plan to consult while working on the project? Be specific where possible (e.g., listing a specific research paper relevant to your project)

| <https://www.cdc.gov/fungal/outbreaks/healthcare-associated-mold-outbreaks.html>  <https://yesweinspect.com/buy-ermi-interpretation-step-two/>  <https://www.emlab.com/services/ermi-testing/>  <https://drive.google.com/file/d/1Gyt5p96Fg7THzxQ571-oRMMjr-jgVyQP/view?usp=drive_link> |
| --- |

**Timeline and Deliverables**

What tasks and outcomes do you plan to accomplish in the first few weeks? The first couple of months? List out specific steps for achieving your objectives.

| **Task**  (what will be done) | | **Outcome**  (expected result of task) | | **Start Date** | **Target Completion Date** |
| --- | --- | --- | --- | --- | --- |
| **Research Unsupervised Learning** | | **Share some concise bullet points on what each of us have learned about unsupervised learning** | | **8/30** | **9/18** |
| **Business Understanding** | | **Specify the variables that we will use to serve in our model and plan out the project in order to succeed.** | | **9/6** | **9/13** |
| **Data Understanding** | | **Review Data and run it under different models, graphs, and tables to see how data behaves.** | | **9/13** | **9/20** |
| **Data Prep** | | **Produce clean data which relates correctly to the target variables.** | | **9/20** | **11/1** |
| **Modeling** | | **Create a ML model that predicts the target accurately and focus on implementing Feature Engineering or training model.** | | **11/1** | **11/15** |
| **Evaluation** | | **Question the final model and evaluate if it performs efficiently for users and check if accurate** | | **11/15** | **11/30** |

**PART 3: DATA UNDERSTANDING**

**Data Structure and Source**

What is the source of the data? What is the data type? (e.g., numerical, time series, text, images, etc.) What is the data format? (e.g. tabular, nested, array, etc.) How much data has been or will be provided? Where will it be stored and in what format? (e.g., csv files)

| The data type provided to us is numerical and in text format. The data we were provided with currently has almost 600 data points and as time passes, more data will be added to the current data we have. The data is stored in a spreadsheet as a csv file. Some of the cells are missing; Some cells contain long texts, which will need to be separated in some way during data preparation. The data comes from people who shared this information with WeInspect voluntarily, based on the ERMI test they did in their homes. |
| --- |

**Data Understanding**

What are some of the variables/features of the dataset(s)?

| Some of the variables/features of the dataset are location, the different types of molds, and the health symptoms. |
| --- |

**Data Preparation**

What data preprocessing steps will be required? (e.g. cleaning, missing value imputation, feature engineering, etc.)

| The data preprocessing steps that will be required are cleaning up the data, and any missing values, determining whether how we will encode ND values in the dataset based on the model we choose, finding and dealing with possible outliers, determining how to encode the many symptoms we have who are currently in long texts. We will need to figure out class imbalance and find ways to deal with it. We believe data preparation will be the longest step. |
| --- |

**PART 4: WAYS OF WORKING**

**Biweekly Meeting Details**

What will be the recurring meeting day and time for your 2 monthly virtual check-in meetings with your Challenge Advisor (“Full Group” meeting in week 1 and Challenge Advisor meeting in week 3)? Please note if these meetings will not happen in weeks 1 and 3 because of scheduling difficulties or preferences.

| The recurring meeting day and time for our 2 monthly virtual check-in meetings with our Challenge Advisor are on Wednesdays, 9:30-10:30 PM. We will have a “Full Group” meeting in week 1 and a Challenge Advisor meeting in week 3. |
| --- |

**Challenge Advisor Communications**

How will you communicate with your Challenge Advisor outside of your biweekly virtual check-in meetings - do they prefer Slack and/or email? How will you share your meeting agendas with them 48 hours prior to each meeting - Slack, email, or a Google Drive link?

| Outside of our biweekly virtual check-in meetings, we will communicate with Corey, our Challenge Advisor, through Slack if any questions arise. We will share our meeting agenda with them 24 hours prior to each meeting via our [Meeting Minutes](https://docs.google.com/document/d/1A4Q6hUONnalaNvkzhmarRxGbZg197Kwi30i4_fB74jA/edit) document. |
| --- |

**Additional Project Stakeholders**

Are there any other stakeholders from your host company/org that your Challenge Advisor mentioned, and who your team might want to connect with to discuss the project?

| Not that we are aware of |
| --- |