Data-Driven Soil Analysis and Evaluation for Agriculture Using Machine Learning Approaches

A Project Report
Presented to
The Faculty of the Department of Applied Data Science
San Jose State University
In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Data Analytics

By

Yixin Huang

Chloe Ngo

Rishi Srivastava

December 1, 2022

APPROVED FOR DEPARTMENT OF APPLIED DATA SCIENCE

Dr. Jerry Gao, Project Advisor

Dr. Lee C. Chang, Department Chair

ABSTRACT

Data-Driven Soil Analysis for Agriculture Using Machine Learning Approaches

By

Yixin Huang, Chloe Ngo, Rishi Srivastava

Healthy soil is vital for daily human nutrition, growing plants, and filtering water (Falmouth, n.d). To raise the productivity of the crop product, it is necessary to choose the crop type wisely for different kinds of soil and to maintain the soil carbon and hydration since the ground is the base to provide nutrition to the crop. As a result, soil analysis is very important which can help us maximize the soil's productivity and maintain the soil's sustainability in producing food and crops regularly. The main goal of this project is to develop data-driven machine learning models to support soil analysis and evaluate the farmlands using satellite images and remote sensor data. As machine learning and deep learning techniques advance, we incorporate these techniques into our soil analysis including crop identification, irrigation system recommendation, and fertilizer recommendation. Classification machine learning models are deployed for crop identification system to classify different crop types. Random Forest with XGBoost provides the best performance with an 86.5% accuracy. The Linear Regression model with Lasso Regularizations is deployed for the irrigation system with 93.9% R-squared. For the fertilizer analysis, Multi-Layer Perceptron is used to find out the best performance with a 93.3% accuracy. We evaluated these classification models using evaluation metrics as described in the web paper Gulati (2020), such as Root Mean Squared Error, Mean Squared Error, Accuracy, and F1-Score. A web-based interactive map is developed for farmland users to check the crop type, irrigation supply and demand, and fertilizer recommendation from our platform.

Keywords: crop identification, irrigation cycle, fertilizer recommendation, machine learning models.

ACKNOWLEDGMENTS

Group 3 would like to express our special gratitude to Professor Jerry Gao who has always given us valuable advice throughout the whole project. Professor Gao's insightful and valuable instruction has led us to the completion of this research. We would not be able to complete this research without his help. The research was completed from spring 2022 (Spring 2022) to fall 2022 (298B), he always goes through with us bi-weekly to ensure that we are on the right track. His willingness to help us every time we need him is much appreciated.

We also would like to express our gratitude towards all the previous team's information which has been very helpful to us. With all the resources that we have received, we were able to complete our research. As a result, group 3 would like to send our thankfulness to all the resources and all the contributions that everyone else made to help us.

TABLE OF CONTENTS

DATA 298B MSDA Project II

Project Report

Team 3

Data-Driven Soil Analysis for Agriculture Using Machine Learning Approaches

December 1, 2022

Chapter 1 Introduction

1.1 Project Background and Execute Summary

Soil is a common material seen on Earth, a natural material composed of solids (e.g., minerals and organic matter), liquids, and gasses according to the definition by Natural Resources Conservation Service (USDA, n.d.). Due to its composition and properties, it has been widely applied in different areas such as construction, energy, or agriculture, etc. Especially the contribution in agriculture leads to the fact that the soil is highly coupled with human daily life. According to the Food and Agriculture Organization of the United Nations in 2015, there is about 95 percent of food is directly or indirectly produced from the soil, which raises the importance of soil since food is the essential energy and nutrition source for humans. Healthy soil is important for human nutrition, growing plants, and filtering water (Falmouth, n.d). There are six types of soil such as clay soil, sandy soil, silty soil, peaty soil, chalky soil, and loamy soil (Barton, n.d.). 65 percent of the human body absorbs the calories from stable crops which are maize, soy, wheat, and rice which are produced from soil (Silver et al., 2015). Crop products are the main food source of human society. There are 6 basic types of crops: food crops, feed crops, fiber crops, oil crops, ornamental crops, and industrial crops. Inside food crops, there are other kinds of crops such as cereal, seeds, pulses, fruits, vegetables, and spices. There are various types of crop products in agriculture and each crop product has a different soil

environment. To raise the productivity of the crop product, it is necessary to choose the crop type wisely for different kinds of soil. In addition to making a good choice of crop for each soil, we need to maintain the soil carbon and soil health since the soil is the base to provide nutrition to the crop. As a result, soil analysis is very important which can help us maximize the productivity of the soil and maintain the sustainability of soil in producing food and crops regularly.

Since there are various kinds of crops, and each crop usually requires different kinds of nutrients to maintain stable status, it is hard for farmers or people without sufficient agricultural background to perform the irrigation. Therefore, we want to combine state-of-the-art deep learning and machine methods to analyze the soil properties and further assist in irrigation production recommendation (or optimization) and fertilizer recommendation. The main goal of this project is to build deep learning models for the purpose of recommending the irrigation product by using the spectral indices extracted from satellite images. Besides, we also want to develop data-driven machine learning models to support soil analysis. To analyze the characteristics of soil for precision agriculture, we first extract the vegetation features including the RGB-based Normalized Difference Vegetation Index (NDVI) (GISGeography, 2021), Normalized Difference Red Edge (NDRE) (MicaSense, 2021), Modified Soil-Adjusted Vegetation Index (MSAVI)(Sarparast,2019), Normalized Difference Moisture Index (NDMI) (Landsat Mission, n.d.), Red-Edge Chlorophyll (RECl) (Earth Observing System, 2022), etc. We take them as inputs and train machine learning models to perform the irrigation production recommendation. For the soil analysis, we also include geographical information such as weather information, farmland properties information (crop type, pH index), etc. to assist in analyzing the soil's chemical composition.

Using remote-sensing data, we want to divide the project into three main focuses: (1) crop identification recommendation, (2) irrigation cycle, and (3) fertilizer recommendation model. We will have the crop identification model which is based on the spectral bands and sensor data; we will analyze demand of irrigation and fertilizer according to the crop it is applied to. To explore the choice of the methods in both soil analysis and irrigation product recommendation tasks, multiple machine learning models are considered including Logistic Regression (Giasson, 2006), Random Forest (R, 2021), Naive Bayes (Ray, 2021), and K Nearest Neighbor (KNN) (José, 2021), Linear Regression (Gandhi, 2018), Support Vector Machine (Gandhi, 2018), Polynomial Regression (Abhigyan, 2020). Besides, we also consider artificial neural networks (Goyal 2021) or other neural network / deep learning-based methods in fertilizer recommendations specifically Nitrogen content in soil. Since the irrigation cycle can be considered as the classification problem, we utilize the standard classification metrics including accuracy, mean squared error (MSE), root mean squared error (RMSE), F-1 score, precision, and recall (Data, 2021).

1.2 Project Requirements

Our goal is to provide an expert system in both soil analysis and irrigation product recommendation for people involved in agriculture fields and is user-friendly without requiring prior agriculture knowledge. The system is a web-based application. The backend is composed of machine learning models to perform the tasks: (1) soil analysis, (2) irrigation product recommendation, and (3) fertilizer (nitrogen) recommendation. The front-end is a web interface for users to input their geolocation (e.g. GPS) of the area of interest, and the analysis result and irrigation product recommendation list will be outputted.

Functional requirements include the features to maintain the expert system operation. Functional requirements are mainly related to (1) how the frontend interacts with the users, (2) how the frontend communicates with the backend, and (3) how the backend retrieves the input from the servers. The functional requirements can be summarized as follow:

1. Allow the users to input the geolocation of the geography soil in GPS format.
2. Send the geolocation information to the backend through HTTP requests.
3. Deploy the search functionality to find the corresponding satellite images from NASA servers according to the geolocation information.
4. The satellite images are sent back to the backend through HTTP requests.

AI-powered features, on the other hand, focus on enabling the system to be intelligent and expert. AI-powered features are closely related to how the backend is able to perform the soil analysis and the irrigation product recommendation. The AI-powered features can be summarized as follow:

1. Reform (or unmix) the image pixels by combining more spectral bands.
2. Extract the vegetation index features for both soil analysis and crop product recommendation.
3. Preprocess the input by combining vegetation index features and geographical information (e.g. climate, land type, etc.).
4. Analyze the soil (chemical) composition given the preprocess input using the machine learning models.
5. Perform the suitable irrigation production recommendation given the preprocess input using the deep learning models.

To construct the expert system, we need the following datasets:

1. Satellite image datasets contain farmland images from the far distance with geolocation (e.g. longitude and latitude) information.
2. UAV image datasets also contain farmland images with geolocation information but are collected at a closer distance.
3. Geographical datasets contain soil properties, weather information, the chemical composition of the farmlands.

1.3 Project Deliverables

The project deliverables can be summarized as follows:

Datasets:

1. Collected datasets including satellite image dataset
2. Satellite remote sensor dataset, climate dataset, and agriculture dataset.

Models:

1. A crop identification model is able to provide an in-depth crop property understanding. The machine learning model will be trained, tested, and evaluated based on satellite remote sensors datasets.
2. An irrigation recommendation model is able to provide suitable irrigation guidance. The linear regression model will be trained, tested, and regularized based on satellite agriculture and climate datasets.
3. A fertilizer recommendation model to predict the Nitrogen content of soil and recommend fertilizer (Nitrogen) application. The Random Forest regression model will be trained, tested, and regularized based on training data obtained from drone-based sensors.

4. A complete project report covering all the details and progress with a demonstration video.
5. A web-based application with a user-friendly interface allows users to check the crop type and soil properties in one specific land area by accessing our crop identification model.
6. A web-based application with a user-friendly interface irrigation recommendation model, to tell the users the moisture situation on the county level.
7. A web-based application with a user-friendly interface fertilizer recommendation model, to tell the users the nitrogen level of one specific land area.

Figure 1

Web-Based Application for Vegetation Index

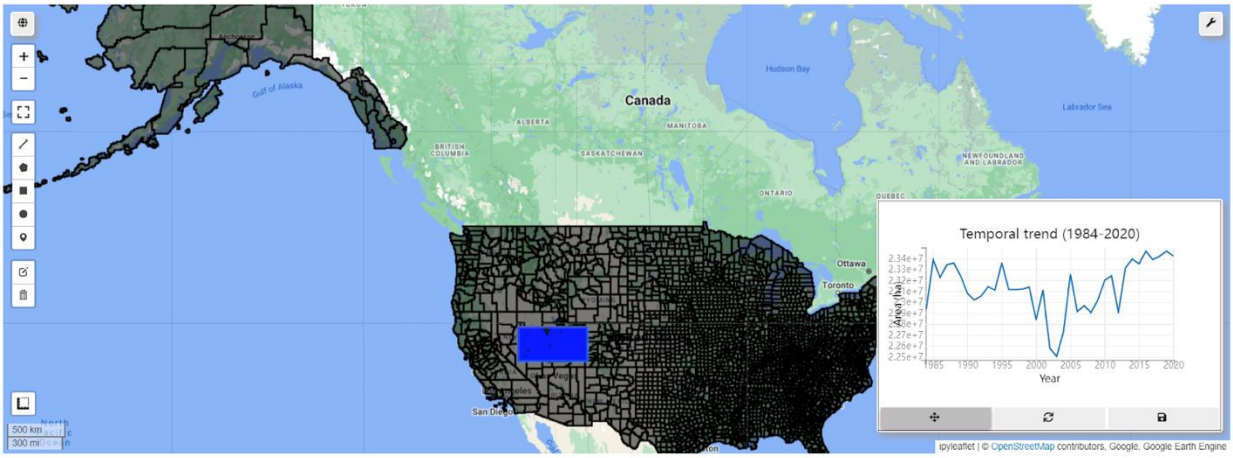


Figure 2

Web-Based Application for Irrigation Recommendation System

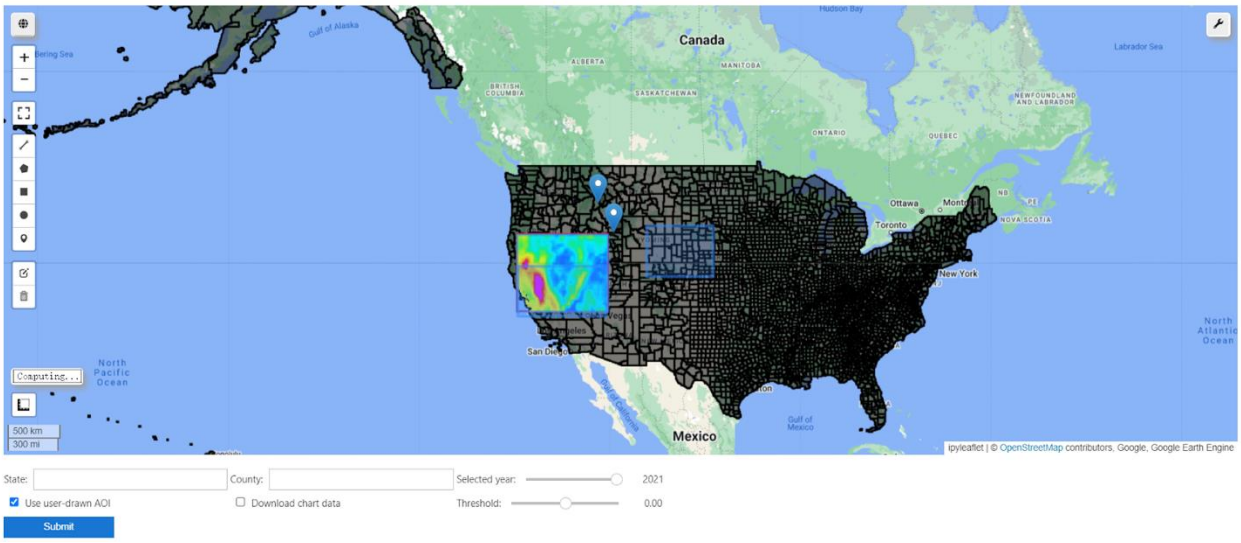
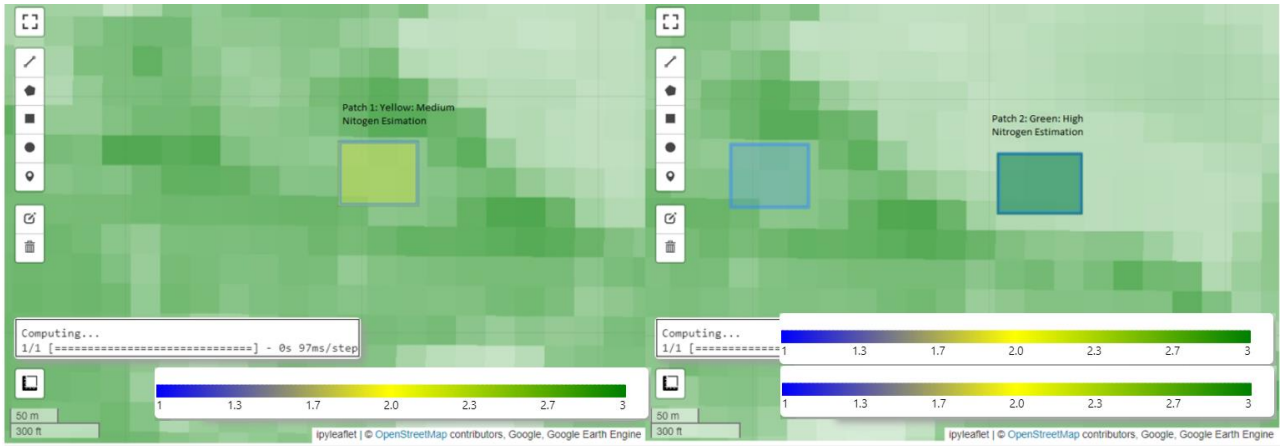


Figure 3

Web-Based Application for Fertilizer Recommendation System



1.4 Technology and Solution Survey

In the first part, we compare the functions of different satellites to understand the suitable use cases in the aspect of launch time, sensor, several bands, spatial resolution, revisit time, and base location. After thoughtful consideration, we prefer images downloaded from Landsat 8 for two reasons. The first is that Landsat 8 has the most spectral bands to analyze spectral indices. Another reason is that the study of Landsat 8 satellites is USA land, where our project focuses. Although the revisit frequency is 16 days with daytime and night-time updates, the frequency is

enough for our data usage. The comparison of our findings on the satellite functions is shown in Table 1.

Table 1

Comparison of Satellite Functions

Satellite Name	Launch Time	Sensor	No. of bands (nominal resolution)	Spatial Resolution	Revisit Time	Base
Landsat 7	1999	ETM+	6	15m, 30m, 60m	18 days	US program
Landsat 8	2013	OLI/TIRS	8	15m, 30m, 100m	16 days	US program
Terra-MODIS	1999	MODIS	36	250m, 500, 1km	Daily	US program
NAIP	2003	Lidar	4	0.6m, 1m	Yearly	US program
Sentinel 1	2014	Radar sensor	A single C band	5m-400m	6 days	European program
Sentinel 2	2015	Radar sensor	13	10m, 20m, 60m	5 days	European program

The decomposition of mixing pixels from hyperspectral images is the main process of our project. Different optimal algorithms of unmixing pixels have distinct changes. The commonly used approach is using the single value decomposition method to extract features from image pixel subgroups. We optimize the unmixing model by deploying neural networks, by which the

noise of the model reduces apparently. Table 2 will show the dimension reduction and pixels unmixing based on our findings

Table 2

Dimension Reduction and Pixels Unmixing

Algorithm	Mechanism	Advantages	Use Case
Bayesian Self-Organizing Map (BSOM) neural network (Liu et al., 2007)	Decomposition of pixels of multispectral sensor image. The algorithm is to minimize the information metric by normalizing the range of Gaussian distributions.	The noise is less than the original model.	Hyperspectral pixel unmixing
Pixel unmixing with SVD (Ball et al., 2004)	First is to use the PCA method to deduct image dimensionalities by recognizing crucial eigenvalues. Then create the mixed pixel groups used to perform pixel unmixing. The mean value is calculated from the testing spectra.	The linear unmixing model is efficient.	Decomposition of mixed pixels from remote sensing images
Modified NMF algorithm and genetic algorithm (GA) (Tao et at., 2007)	The final matrix obtained from MNMF is the initial input to GA, and the optimal result got from GA is the new input to the next round of MNMF running. Within the max iterations, the global optimal is achieved based on the cost function least square method.	Quicken the convergence and find the global optimal instead of local optimal	Fast decomposition of mixed pixels in hyperspectral images and better rate-to-noise result.

1.5 Literature Survey of Existing Research

Table 3 compares the models, algorithms, and input datasets of literature with a similar research topic. The most common method to identify crop types is supervised, learning models. The most commonly used inputs are bands and images from satellite remote sensors. Table 4 lists the methods of analyzing nitrogen in the soil, by summarizing the crucial points and contributions of the literature. The most common method used nowadays is to calculate the nitrogen value based on image band values. But the drawback is that the accuracy of the calculation is not so satisfying. Then new indices improved the correlation between nitrogen and spectral bands by adding or creating new features into the band calculation. In recent years, neural network models have been trained to find out the nitrogen amount in the soil.

Table 3

The Literature Comparison of Models and Input Datasets

Model	Supervised Classification	Unsupervised Classification	Deep Learning	Inputs
Monitor leaf nitrogen in coffee (Parreiras, 2020)	Maximum likelihood classifier	N	N	Vegetation index
Fertilizer analysis from UAV imagery (Escalante et al., 2019)	N	N	Convolutional neural networks	RGB images captured from low-cost UAVs
Estimating biomass and nitrogen amount (Nasi, 2018)	Y	N	N	RGB images captured from drones' camera

Estimate nitrogen status (Caturegli, 2020)	Y	N	N	New index HSB (Hue, Saturation, and Brightness)
Crop water use (Abuzar, 2013)	Linear regression model	N	N	Crop water supply, Crop water requirement, NDVI
Growth stages classification (Liu, 2020)	Correlation analysis, successive projection, random frog	N	N	Spectral bands, feature coefficients
Growth stages of crops using mobile images (Rahima, 2021)	SVM, RF, DT, K-NN, NB	N	N	Crop images from mobile phone
Wheat and barley growth stage (Rasti, 2020)	N	N	CNN	Crop images
Total nitrogen estimation in agricultural soils via aerial multispectral imaging and LIBS Hossen et al.(2021)	Random Forest	N	Neural Network	Spectral images from Drone LIBS images from Drone Sensors

Table 4

The Approaches of Crop Identification Model

Literature 1: Crop identification using Landsat temporal-spectral profiles (Odenweller, 1984)	
Summary	Green vegetation indicated from single label can indicate that numerous segments determined that individual corps is determined by individual corps different from the other. This can support crop identification at multiple levels.
Contribution	Using temporal-spectral profile of Landsat data Annual crops separated into groups based on thresholds Individual crops identified based on different amplitude and shape of the temporal-spectral profile.
Literature 2: Evaluation of Deep Learning CNN Model for Land Use Land Cover Classification and Crop Identification Using Hyperspectral Remote Sensing Images (Bhosle, 2019)	
Summary	Using deep learning convolutional neural network and land use land cover with remote sensing data for crop identification has optimized around 97.58% accuracy
Contribution	Tested on two dataset using deep learning convolutional neural network by tuning the parameters based on the optimizer, activation function, filter, learning rate, and batch size. Gather data from EO-1 Hyperion sensor
Literature 3: Crop classification of upland fields using Random Forest of time-series Landsat 7 ETM+ data (Tatsumi, 2015)	
Summary	Crop classification using multi-temporal with different frequencies classified by Random Forest. The authors identified had a comparison between Random Forest and Kappa statistic which yields the accuracy of 81% and 0.70. They concluded that Random Forest has good performance enhanced by appropriate classifiers.
Contribution	Using enhanced vegetation index and the statistics obtained from Landsat 7 based on the sensitivity of the dataset size, the variables, and mapping accuracy.

	Analyze using Random Forest algorithm and Kappa statistic.

Table 5

The Approaches of Irrigation Cycle Model

Literature 1: An approach of cost-effective automatic irrigation and soil testing system (Mahmud, 2020)	
Summary	The shortage of the rainfall and water has pushed the farmers into critical period. The authors wanted to integrate irrigation cycle in order to keep track of the macronutrients. Based on the results, they will build a smart system that can analyze the chemical elements, rate of soil, and the moisture.
Contribution	Used TCS230 color sensor instead of pH meter. Build a mobile app to monitor the device instead of hardware display Compare the results and the color diagram of N, P, and K in the sample
Literature 2: Research on soil moisture prediction model based on deep learning (Cai, 2019)	
Summary	Existing soil moisture prediction models are inaccurate, and the generalization, process capability, and performance need to be improved. Using Beijing dataset, they summarized that deep learning is more feasible and effective. Good data fitting and generalization capability is improved with higher accuracy.
Contribution	Proposed Deep Learning Regression Network with big data to build a soil moisture prediction model. Analyze time series of predictive variables and find out the connection between features and variables using Taylor diagram
Literature 3: A comprehensive review on automation in agriculture using artificial intelligence (Jha, 2019)	

Summary	The high demand in food correlates with the rapid increasing of the population. What farmers using right now isn't so efficient but to use harmful pesticides. The authors have used different automation has shown that the gain from soil has created a significant improvement on the soil fertility.
Contribution	Different automation like IOT, wireless communicators, machine learning, artificial intelligent, and deep learning. Using IOT to implement a botanical farm identification and watering system.

Table 6

The Transition Process of Analyzing Nitrogen Method

Literature 1: Using the unmanned aerial vehicle and machine learning algorithm to monitor leaf nitrogen in coffee (Parreiras, 2020)	
Summary	In order to prevent unnecessary fertlizer , the paper tried to analyze the vegetation indices based on the bands of UAVs. Then evaluate the correlation between nitrogen and the vegetation indices. Image classified by using the color index of vegetation and maximum likelihood classifier.
Contribution	Analyzing the vegetation indices by calculating bands of UAVs. The approach used to separate coffee lines from the image background is CIVE from supervised classification.
Literature 2: Barely yield and fertilizer analysis from UAV imagery: a deep learning approach (Escalante et al., 2019)	
Summary	Try to optimize the fertilizer doses by analyzing the nitrogen volume in the farmland. Features are extracted based on deep convolutional neural networks. The data resource is RGB images captured from low-cost UAVs
Contribution	Use deep convolutional neural networks to monitor rates of nitrogen.

Literature 3: Estimating biomass and nitrogen amount of barley and grass using UAV and aircraft-based spectral and photogrammetric 3D features (Nasi, 2018).	
Summary	Use machine learning models to estimate nitrogen amount in crop fields. The approach to collecting spectral features is using RGB and hyperspectral cameras installed on drones.
Contribution	Combine RGB spectral data with other features to analyze the nitrogen.
Literature 4: Normalized difference vegetation index versus dark green color index to estimate nitrogen status on bermudagrass hybrid and tall fescue (Caturegli, 2020).	
Summary	Converse the RGB values into a new index HSB (Hue, Saturation, and Brightness), which has a higher correlation to the nitrogen value than RGB has.
Contribution	Create a new index HSB with a close relationship with nitrogen content.

Table 7 presents the evolution of soil fertility analysis, from the aspects of moisture level, fertilizer amount, nutrient content, and so on. The challenge for farmers to feed the soil is to manage the soil web properly. By monitoring multiple soil parameters and utilizing tools to improve soil health, the living soil factory will function in a more effective way.

Table 7

Evolution of Soil Fertilizer Analysis

Literature 1: Soil nutrient analysis using colorimetry method (Madhumathi et al., 2020).	
Summary	This report monitored the amounts of soil moisture and nutrients such as NPK (nitrogen, phosphorous, and potassium) based on an IoT system. Farmers can increase crop yield by following the recommended usage of water and fertilizer. The amounts of NPK are extracted from a color sensor,

	by calculating the RGB values. The recommendation is achieved by comparing the amounts to the standard ranges of NPK.
Contribution	It states that the irrigation and fertilizer values are related to amounts of soil moisture and nutrients.
Literature 2: Application of colorimetry to determine soil fertility through Naive Bayes classification algorithm (Agarwal et al., 2018).	
Summary	A soil fertility system is detected in the report by analyzing RGB from colorimetry and testing the result based on a Naive classification algorithm. The accuracy of the classification model depends on the soil types, texture, and structure (such as black soil with clayey texture).
Contribution	The reports shows that soil fertility is not only dependent on soil RGB but also on other features of the soil.
Literature 3: IoT-based real-time soil nutrients detection (Patil et al., 2021).	
Summary	The chemical contents are monitored in the soil with the utilization of an RGB color sensor and soil doctor plus kit. The chemical contents include nitrogen, soil pH, and humidity. All the RGB data has been analyzed after uploading to the cloud system.
Contribution	Real-time, cloud-based soil parameters (soil pH, nutrients, temperatures)

Chapter 2 Data and Project Management Plan

2.1 Data Management Plan

The data we collected include satellite spectral bands, numerical data about soil properties, and underground utilities. Satellite images are used for an irrigation recommendation system, soil property data is used for soil chemical attributes analysis, and ground-penetrating radar data is used to detect the distribution of underground utilities. Data will be stored in the Google cloud platform.

A neural network based on satellite images is used to acquire the land features, such as vegetation index NDVI and NDMI. The resources of satellite images come from Landsat 8 and the EOS crop monitoring system of NASA. We collected the Landsat 8 data each two weeks of 2021, which is updated every 16 days, with 36 spectral bands.

Soil chemical properties, which influence the availability of plant nutrients, are extracted from WSS (web soil survey) of USDA, including calcium carbonate (CaCO3), cation-exchange capacity (CEC-7), pH to water, and NPK. We picked the expected value by averaging the low value from the deep soil layer and the high value from the shallow soil layer.

Climate data online provided all the historical weather and climate data globally. We collected the temperature, precipitation, evaporation, and wind direction in 2021. The frequency of collected data is daily. Table 8 shows the dataset sources that we have collected.

Table 8

Data Sources Collected

Data Source	Type of Variables	Collected Variables	Description
EOS Crop Monitoring: (Crop Monitoring, 2022)	Satellite Image	Satellite images NDVI, NDMI, RECI Farmland objects	Landsat 8 satellite Update in 16 days 36 spectral bands
Web Soil Survey: (Nrcs, 2022)	Soil chemical properties	CaCO3, CEC-7, NPK, pH	Low value for deep soil layer, high value for shallow soil layer, expected value
Climate Data Online: (National Centers for Environmental Information, 2022)	Climate data	Temperature, precipitation, evaporation, wind direction	Daily weather data from Jan 2021 to Dec 2021

All the datasets are split into three subsets which are training dataset, validation dataset, and testing dataset, at the weight of 60%, 20%, and 20% respectively. The training dataset is used to build the deep learning and machine learning models, the validation dataset is used to validate the effectiveness of models, and the evaluation and accuracy results are based on the testing dataset

2.2 Project Development Methodology

There are two main development cycles in our project, one is a neural network deep learning model, another is a machine learning model. The first self-learning AI model will start from the labeled images with separated pixels subgroups. Then the model will be enhanced gradually by feeding more labeled satellite images. The extracted data from the image will be considered as input features to our machine learning model. In the meantime, an iterative method will improve the accuracy of the AI model continuously. The detailed process is shown in the following Figure 4.

Figure 4

Soil Analysis Development Cycle

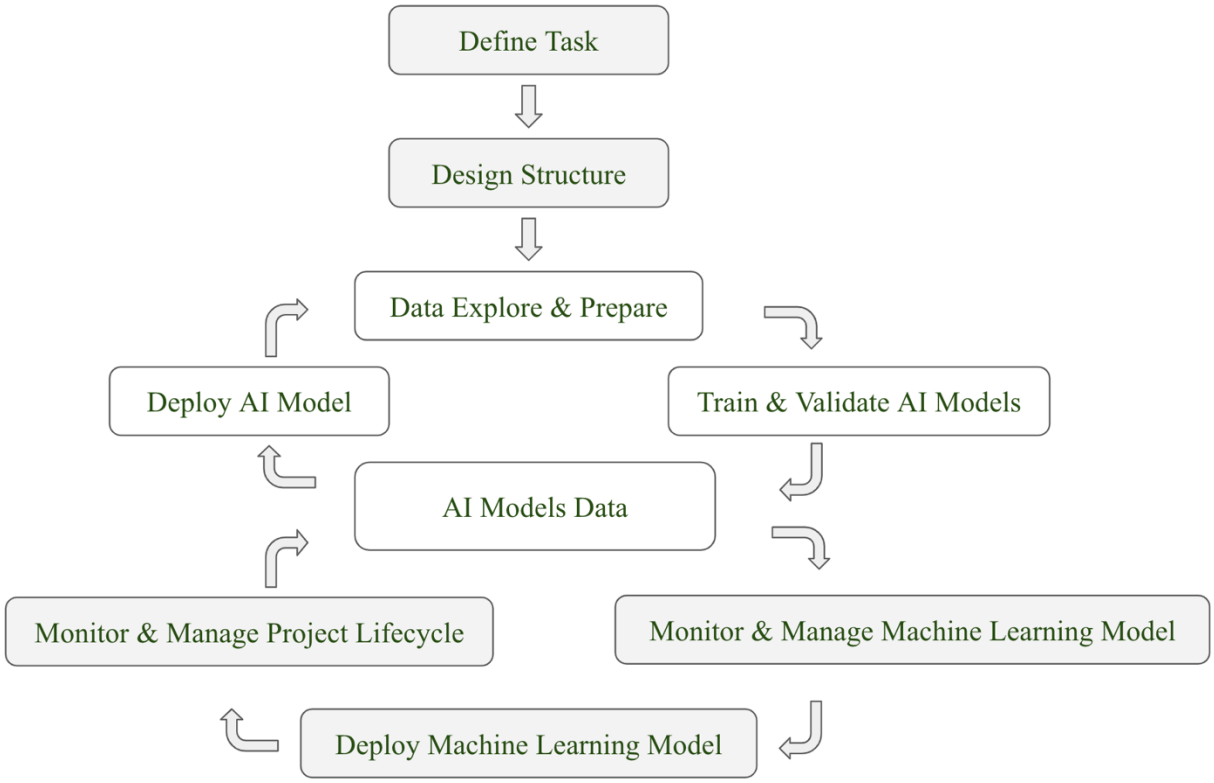
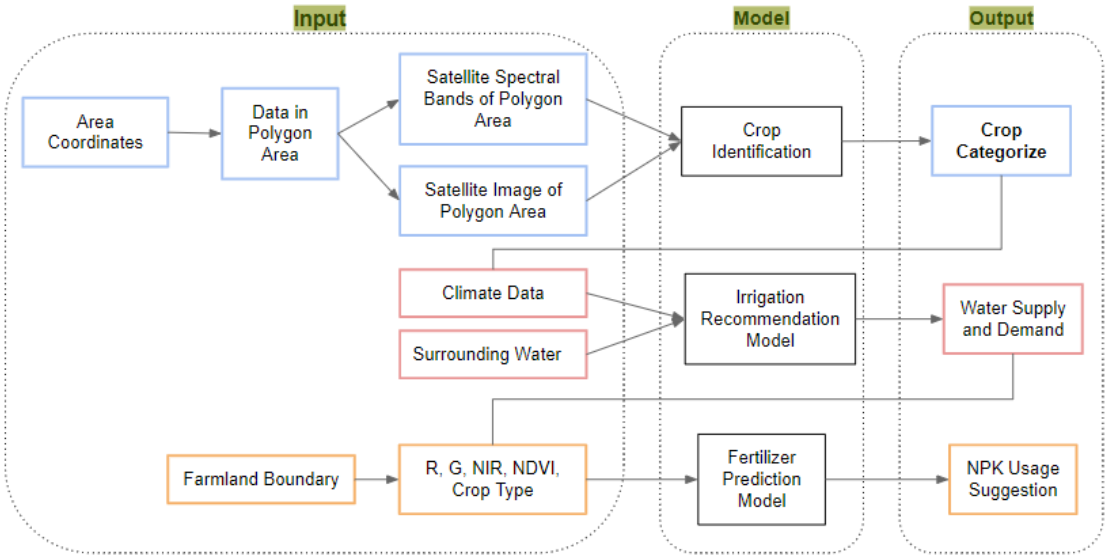


Figure 5 presents the data flow from crop identification platform to irrigation platform and then to fertilizer platform in detail. The output from former platform is used as the input of the next following platform. For example, the crop categories predicted from crop identification platform flows into the irrigation model, and the water quantity calculated from irrigation model is used as the feature for the third platform.

Satellite bands based on coordinates are viewed as the training data. Then the small groups are delineated based on polygon boundary. By deploying the machine learning category model, the vegetation detection and classification are trained from the model. After meeting the accuracy requirements, the vegetation indices calculated from bands are used in the machine learning model to predict the irrigation recommendation system. When we get the water supply and demand from irrigation model, the results will be used to predict NPK usage in the fertilizer prediction model, combined with land features and climate features.

Figure 5

Data Flow and Management



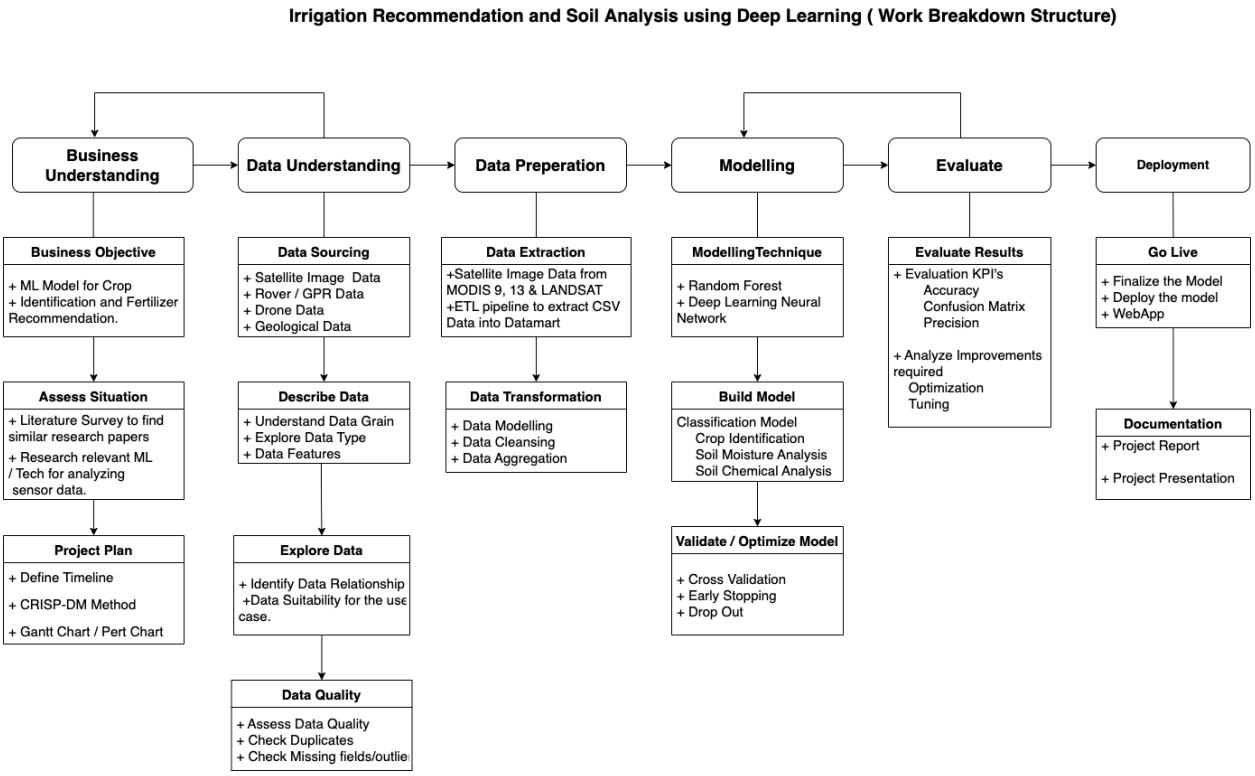
2.3 Project Organization Plan

The project will be executed in six Steps. The first step 'Business Understanding' involves gathering and documenting project requirements and goals of the project. The goal of this project is to come up with a machine/deep learning model that can predict soil quality and recommend irrigation for farmland. We will be doing literature surveys of existing similar work and assessing relevant technologies that can be helpful in current projects. A detailed project plan and timeline will be finalized in this step. In the 2nd step 'Data Understanding', we will take the exercise to find relevant data sources required for this project. The data will be checked for its quality and features relevant to the project. In the next step 'Data Preparation', we will be processing all the data sources in the previous step, which will be cleaned, transformed, and aggregated to be further fed into a machine learning/deep learning model. The next step is 'Modeling', which involves training the machine learning model to do soil analysis and irrigation recommendations. In the next step 'Evaluate', we will evaluate the machine/deep learning

models on evaluation metrics to get the desired accuracy and optimize the model for performance. In the last step 'Deployment', we plan to deploy the model into a web application and complete writing the technical paper for presentation. The work breakdown structure is represented in the following flow chart (see Figure 6).

Figure 6

Work Breakdown Structure for Irrigation Recommendation and Soil Quality Prediction using Machine / Deep Learning



2.4 Project Resource Requirements and Plan

Various hardware and software required in this project are listed in the table below. The modeling part of the project will be done in an in-house computer system with 4 GB GPU Memory. The crop identification model is trained on Razer Blade 13 with 1.8GHz CPU and GeForce MX150 GPU. The irrigation cyclE is trained on MacBook Pro with Inter i7 and Intel HD graphics 630 GPU. The fertilizer recommendation is trained on HP Omen PC with Inter i7

processor and Nvidia 3080ti GPU. The whole models are deployed in a web app hosted in the google cloud environment. The cost of these resources is shown in Table 7 below. The other software required for this project such as Postgres DB, Jupyter Notebook, Python 3.7 with Sklearn, Pandas, Numpy, Tensorflow and Keras, SNAP, and ArcGIS is free software available for non-commercial purposes. Table 9 will show all the specifications and costs of our project resources.

Table 9

Specifications and Cost of Project Resources

Function	Resource Type	Resource	Time Duration	Cost
Crop Identification	Hardware	Razer Blade13 Processor: Intel i7-8565U CPU @ 1.80GHz Memory:16 GB GPU: GeForce MX150	02/2022 - 11/2022	$700.00
Irrigation Cycle	Hardware	MacBook Pro Processor: 2.8GHz Quad-Core Intel i7 Memory: 16GB 2133MHZ GPU: Intel HD Graphics 630	02/2022 - 11/2022	$800.00
Fertilizer Recommendation	Hardware	HP Omen PC Processor: Intel Core i7 Memory: 32 GB GPU: Nvidia 3080ti, 12 GB	02/2022 - 11/2022	$1100.00
Database	Software	Postgres Database Google cloud	02/2022 - 11/2022	$50.00
Data Exploration	Software	Jupyter Notebook SNAP Arc GIS Google Earth Engine	02/2022 - 11/2022	$0.00
Machine Learning Tools	Software	Python 3.7 Sklearn, Pandas, Keras TensorFlow, NumPy, Google Collab.	02/2022 - 11/2022	$30.00
Licenses	Software	ArcGIS Google Earth Engine	02/2022 - 11/2022	$0.00
GUI Application	Software	Voilà (Jupyter Notebook)	02/2022 - 11/2022	$0.00

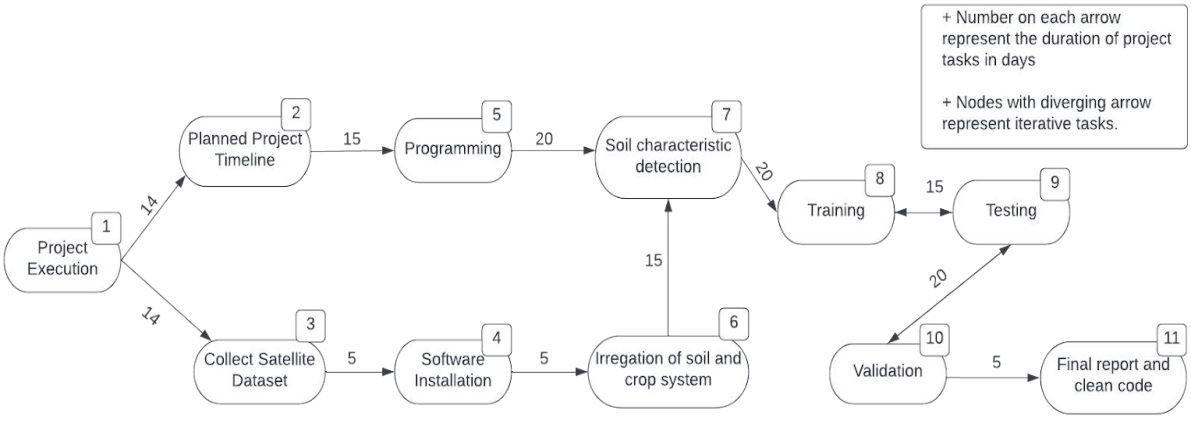
2.5 Project Schedule

In order to guarantee the best results and deliverables, we create a Gantt chart (Figure 7) to keep up to date throughout the whole project. Moreover, we also create a PERT chart (Figure 7) to keep track of the project analysis and independent tasks. All timelines are the dates that are currently planned to help the team organize tasks and deliver the application on time.

Our project is separated into three main tasks. The first task, the crop identification cycle, is assigned to Yixin, which is our project manager. The second task, the irrigation cycle, is assigned to Chloe. Finally, the last task which is the fertilizer recommendation model is taken care of by Rishi. Below is the static image of the PERT Chart and Gantt Chart (Figure 8) which shows the status of our project.

Figure 7

PERT Chart

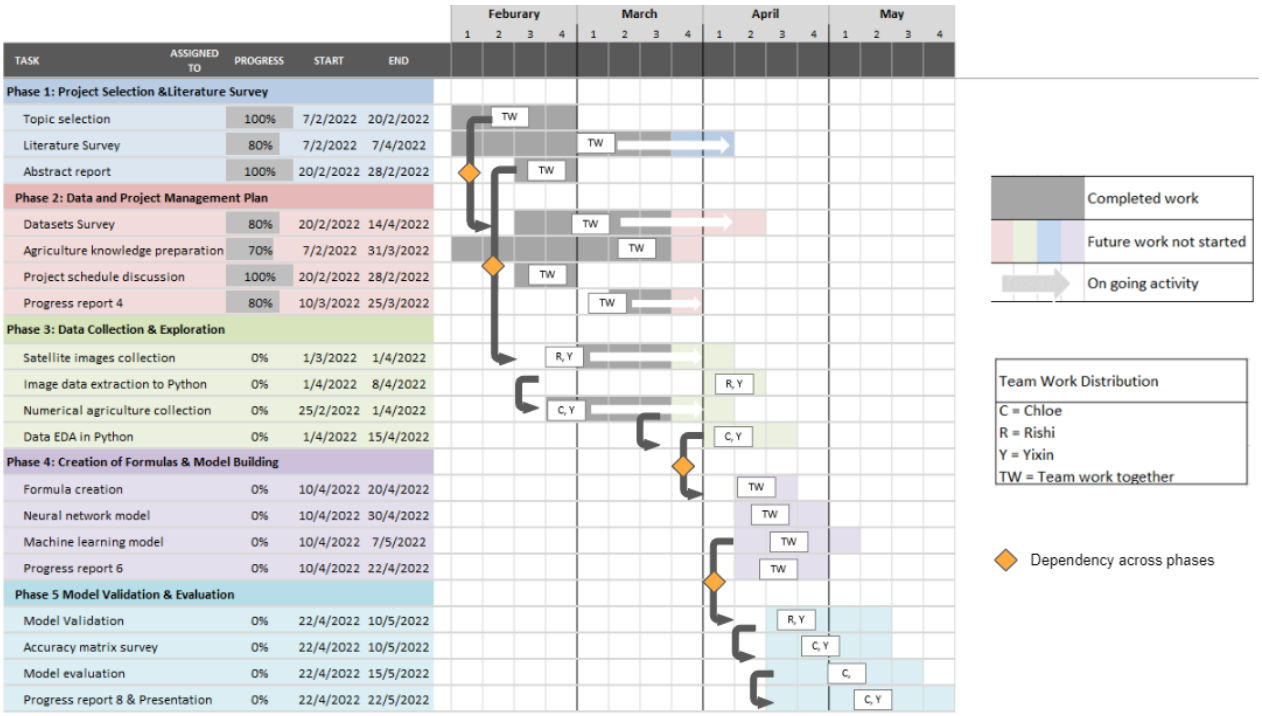


All of us three worked on Phase 1 project selection and literature survey (topic selection, literature survey and abstract report) and Phase 2 data and project management plan (datasets survey, agriculture knowledge preparation, project schedule discussion and progress report) on February 2022. Then we split into subgroups to collect and explore data on Phase 3 April 2022. During the summer of 2022, our group focused on build machine learning and deep learning

models to analyze soil properties. On Phase 5 model validation and evaluation, we spent August and September to compare the model results. Lastly, we built the web application allowing the clients to interactive with our platforms on Phase 6 web application.

Figure 8

Gantt Chart



Chapter 3 Data Engineering

3.1 Data Processing

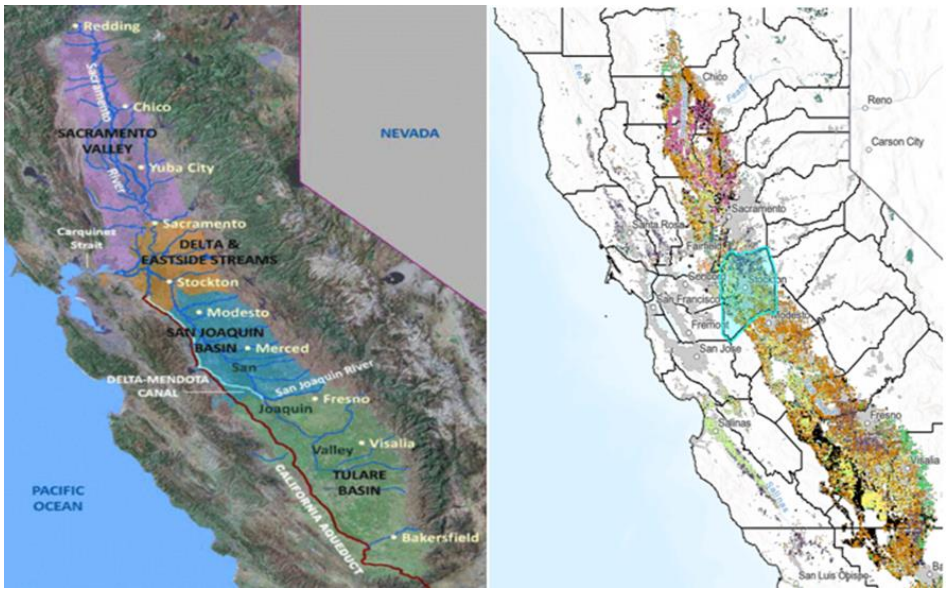
Step 1: Determine the Study Area

The study area of the project is San Joaquin County, one part of the Great Valley of California, US. The main reason to select this county is that San Joaquin is located in the middle valley with multiple agricultural productions. The whole area covers about 20,000 square miles, accounting for approximately 75% of the cropland of California and 17% of the whole country. The terrain is flat without so many mountainous areas in the west valley, also the soil is in a good

condition for agriculture. The predominant crop types include citrus, grains, cotton, vegetables, nuts, and grapes.

Figure 9

Area of Interest: San Joaquin County



Step 2: Determine the Needed Raw Datasets

The datasets related to the project are mainly focused on two aspects: one is the on-the-ground task of crop identification and irrigation cycle system based on soil moisture analysis; another is the underground task about fertilizer recommendation based on soil nutrition analysis.

Each objective needs different raw data and analytics tools. For the crop identification section, we build the models by utilizing satellite images and land use maps; for the irrigation cycle system, the crop identification results are used as part of the inputs, combined with climate and soil data; for the fertilizer recommendation section, the crop types and soil data are used.

Table 10

Main tasks to Develop Models

Task	Content	Raw Data

Crop Identification	Identify the types of crops	• Vegetation indices • Geolocation (longitude and latitude) • Date of the satellite flights • Land use map collected by California natural resources agency
Irrigation Cycle	Recommend irrigation system	• Precipitation and ET • Vegetation indices • Energy flux (Abuzar, 2013) • Surface-water • Soil water (4 layers) • Underground water
Fertilizer recommendation	Recommend the usage of fertilizer (NPK)	• Soil Chemical properties analysis • Nitrogen Content Prediction

Step 3: Dataset Partition

We separated our crop identification dataset and irrigation system dataset into 80% data for training and 20% for testing. As the crop count in each group is an abnormal distribution, the stratified method is used to confirm the same weight in each splitting dataset.

Figure 10:

Data Partition

split the datasets

```
# stratify=df_target

x_train, x_test, y_train, y_test = train_test_split(df_features, df_target, test_size = 0.2, random_state = 11, stratify=df_target)
```

3.2 Data Collection

The data collection sources are categorized into five categories: 1) vegetation indices, 2) geolocation data (longitude and latitude), 3) land crop data; 4) soil data (moisture content and fertilizer content), and 5) climate data

3.2.1 US Census Counties Dataset

In order to find the boundary and compute the area of the interested region, we use the US census bureau TIGER dataset from Google Earth Engine (GGE). This dataset contains up to

Level 2 (counties) boundaries of all the US states (Figure 11). The contents of the dataset shown in Table 11 include land area, water area, county code, county name, internal point latitude, internal point longitude, etc. Then we have the polygon for our region of interest (San Joaquin County) by filtering the feature collection ‘county code’.

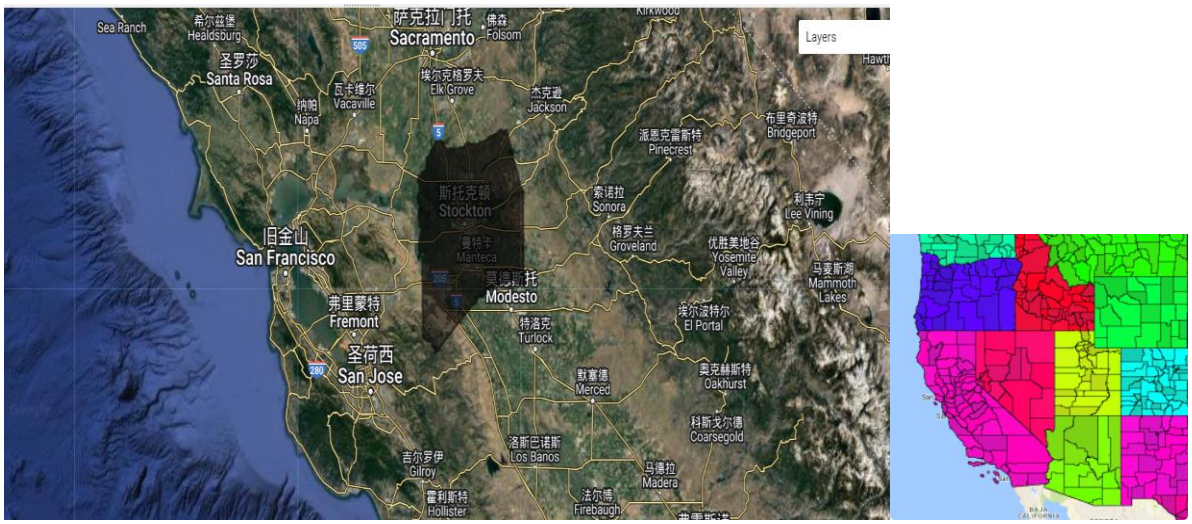
Table 11

Contents of US Census County Dataset

County Code	Type	Description
Aland	Double	Land area
Atwater	Double	Water area
IRR_TYP1PB	String	Irrigation status for the first land-use/type of the irrigation system
COUNTYFP	String	County FIPS code
COUNTYNS	String	County GNIS code
GEOID	String	County identifier; a concatenation of state FIPS code and county FIPS code
INTPTLAT	String	Internal point latitude
INTPTLON	String	Internal point longitude
Name	String	County name

Figure 11

Sample of Layers Gathered by Google Earth Engine



3.2.2 Land Use Map

The land use map from DWR gives the public access to statewide land and crop datasets for the last 30 years. It is a Legislature requested investigation in the aspects of agricultural land use and water needs. Most of the survey data are recorded into geographic information system (GIS) software by surveyors, who visited more than 95% of agricultural areas in the state and recorded the land usage.

We select the dataset for San Joaquin County. Overall, there are 74,382 polygons delineated in the region, within which 34 objects are categorized. For example, acres, area_meter, class1, class2, class3, croptype1, croptype2, croptype3, irr_type1… The general types of crops are labeled on the map with different colors, and the second and third level categories are present in the shapefile. The time of the survey focuses on summer which is the driest season of the whole year. Water irrigation system loads a heavy burden in summer as it is the growing season for most plants. There are 12 categories of the general crops, except U stands for urban and X stands for unspecified.

We picked ‘Class1’ as the classification for land use, which is the first level of land use identification. The codes for the classification are all capital letters. And each polygon in the

region has a code for Class1 type. Since too many categories will influence the accuracy of the classification model, we delete or combine the 17 categories in Class1 into 8 categories which are G, R, F, P, T D, C, V (Table 12 and Figure 12).

Table 12

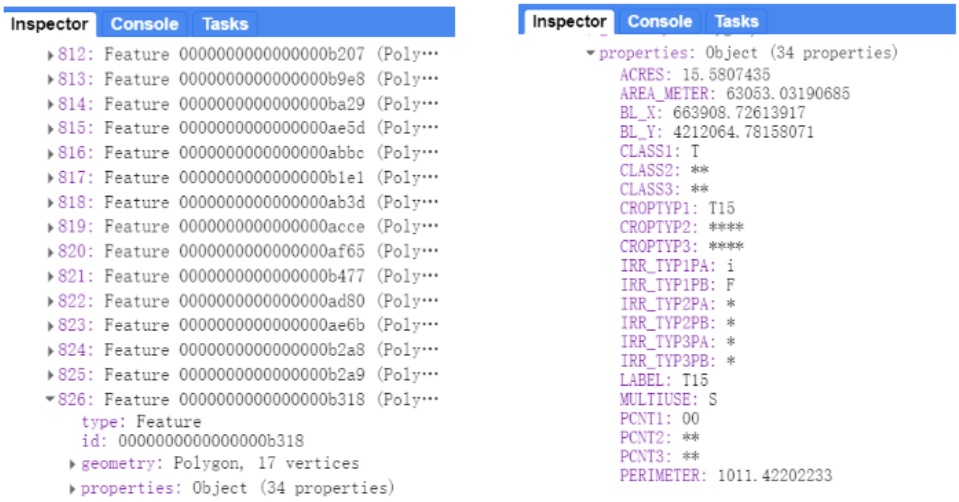
Description of Class1

Category code	Description	Details
R	Rice	
P	Pasture	
G	Grain and Hay	
T	Truck, nursery, berry crops	
F	Field Crops	
C	Citrus and Subtropical	
D	Deciduous fruits and nuts	
V	Vineyard	
S	Semi-agricultural	Other land use
I	Idle	Other land use
U	Urban-residential, commercial, and industries	UR - Urban residential UC - Urban commercial UI - Urban industrial UV - Urban vacant
N	Native classes	NC - Native classes unsegregated NV - Native vegetation NR - Native riparian vegetation
NW	Water surface	Other land use
NB	Barren and wasteland	Other land use
NS	Not surveyed	Other land use
E	Entry denied	

Z	Outside of the study area	

Figure 12

Feature Inspector Using Google Earth Engine



3.2.2 Vegetation Index

We collected satellite images from Google Earth Engine (GGE), which provides high-resolution images with free access and download. The imagery we selected comes from Landsat 7 collection 1 Tier 1 and Landsat 8 OLI data. Landsat 7 Tier 1 has the high available data quality with a consistent scene (less than 12-meter RMSE). The resolution is 30 -60 meters, and the global coverage repeats after 16 days.

The bands used for the vegetation index calculation are shortwave bands from band 1 to band 6. B1 represents blue color with 0.45-0.52 µm, B2 represents green color with 0.52-0.60 µm, B3 represents red color with 0.63-0.69 µm, B4 represents near-infrared with 0.77-0.90 µm, B5 represents shortwave infrared with 1.55-1.75 µm.

Since the crop reference results are based on the whole year, we also analyze the multi-seasonal data of the spectral bands and put all the bands into an input vector. The number of the

total spectral bands N = 5 bands * 4 seasons. Moreover, clouds could influence the value of bands, producing noise in the bands' results. We removed those images with cloud coverage of more than 5 percent. Table 13 shows the description of the total bands.

Table 13

Total Bands Based on Satellite Remote Sensors

Name	Description	Resolution (meters)	Wavelength (μm)
B1	Blue	30	0.45 - 0.52
B2	Green	30	0.52 - 0.60
B3	Red	30	0.63 - 0.69
B4	Near-infrared	30	0.77 - 0.90
B5	Shortwave infrared 1	30	1.55 - 1.75

1. NDVI

NDVI is an agriculture index to estimate the density of green on land. It is a popular measurement method since the inputs for NDVI come from the public remote sensing data, most are spectral bands of satellites. The logic behind the NDVI is based on the differences in plant reflectance. The pigment in plant leaves absorbs red light (red band in satellite image) but reflects near-infrared light. It means the greener segments on the land, the more difference between red waves and near-infrared waves. Moreover, NDVI is one way to detect the health of vegetation. Healthy vegetation absorbs abundant amounts of visible light and reflects a big portion of the near-infrared light, while unhealthy or sparse land absorbs less visible light and reflects less near-infrared light, making the light difference smaller.

Band 3 of the satellite image represents the visible light which is also taken as the red band, and Band 4 is called for the near-infrared band. The formula of NDVI is near-infrared band minus visible band divided by the near-infrared band plus visible band (Sruthi et.al, 2015).

$$NDVI = \frac{NIR - VIS}{NIR + VIS}$$

2. EVI

Enhanced vegetation index was adopted from two MODIS satellites for the purpose of eliminating the environment and background noise shown in NDVI. The formula of EVI is similar to the NDVI's, but with a correction to the mistook reflected light which is caused by land ground and air particles (Matsushita, 2007).

$$EVI = G \times \frac{(NIR - Red)}{(NIR + C1 \times Red - C2 \times Blue + L)}$$

where:

NIR / Blue / Red are surface reflectance extracted from satellite images

L is the background adjustment

C is the coefficients to correct the red band reflectance

3. LSWI

Land surface water index is a vegetation index response to the total water content in the land surface, by using near-infrared and shortwave infrared bands. The soil drought and moisture level depicted by evaporative stress can be measured from the LSWI index.

$$LSWI = \frac{\rho_{858} - \rho_{1640}}{\rho_{858} + \rho_{1640}}$$

where:

ρ858: NIR band at 858 nm

ρ1640: SWIR band at 1640 nm

4. NDSI

Normalized soil index outlines the landscape feature of the land exposure level, which is measured by the difference between the soil mineral composition (short wave infrared and the red band) and the vegetation presence (near infrared and blue band).

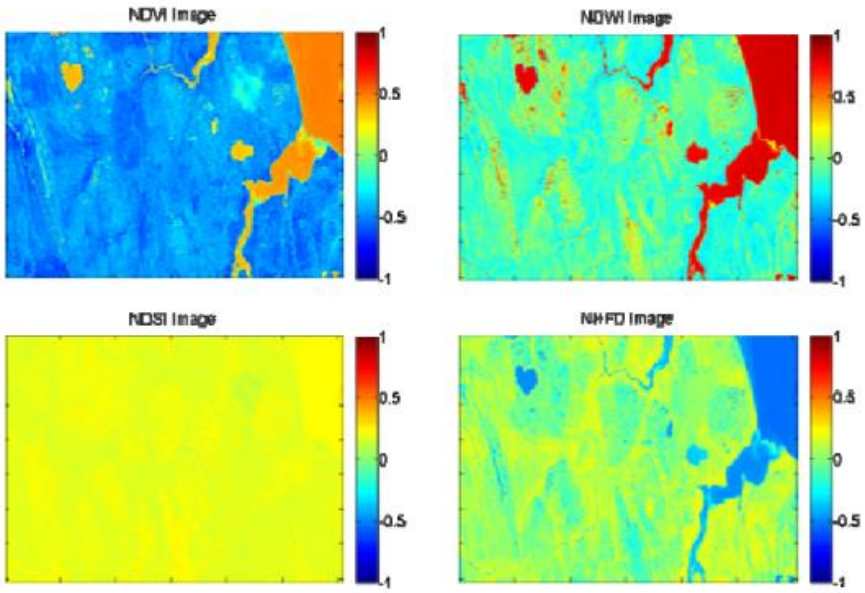
$$BSI = \frac{\big((Red + SWIR) - (NIR + Blue)\big)}{\big((Red + SWIR) + (NIR + Blue)\big)}$$

3.2.3 Reasons for Vegetation Index Deployment

Each spectral index has its specific use case and interpretation. As the example of the following figure 13, for the barrow area the NDVI (top left) separate the vegetation and moisture in blue and orange color, the NDWI (top right) distinguish the moisture level more apparently than the NDVI's. The NHFD (bottom right) index presents the split of dry and moist section moderately, while the NDSI (bottom left) index scaly catches the moisture feature in the land.

Figure 13

Comparations Among Different Index Images with the Same Color Scheme



Different crops have different index distributions as showed in figure 14. The horizontal axis represents the type of crop, the vertical axis is the value of index ‘NDVI’, ‘LSWI’, and ‘NDSI’. The upper line of each box means the top 25% value, the middle line in each box is the mean value of the index. The points(circles) above/below the line are the outliers of each index. For crops ‘G’ and ‘P’, the mean values are relatively higher than other indices. For the crop ‘F’, ‘T’, and ‘R’, their mean values are around 0.2. The extracting band values are shown in Figure 15.

Figure 14

Index Distributions of Different Crops

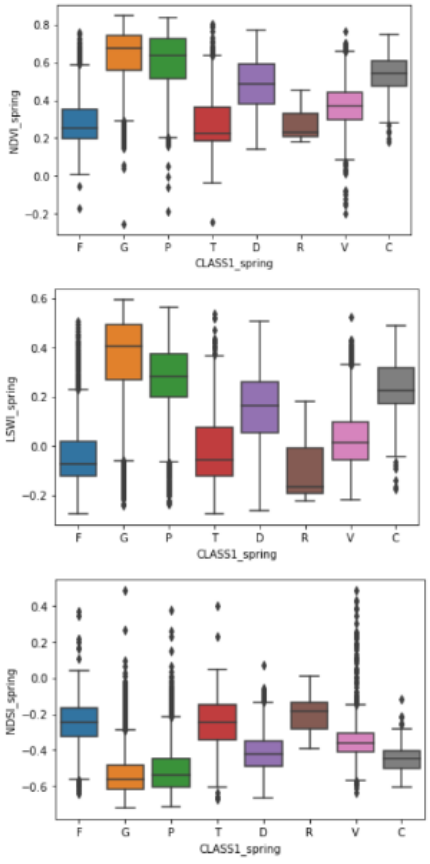


Figure 15

Extracting the Bands Using Python

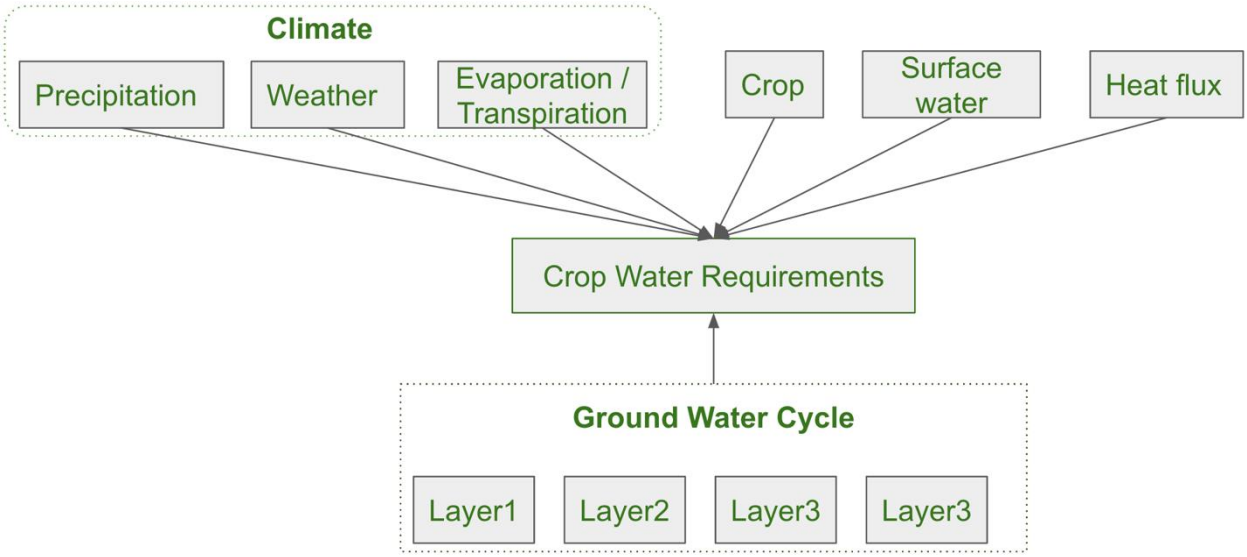
lex	B1_spring	B10_spring	B11_spring	B2_spring	B3_spring	B4_spring	B5_spring	B6_spring	B7_spring	B8_spring	B9_spring	CLASS1_spring
ab	0.174291	299.341827	296.964478	0.162792	0.153956	0.169576	0.260021	0.314058	0.257832	0.158310	0.007199	F
b5	0.157427	301.605499	298.911774	0.141596	0.125611	0.131975	0.207725	0.245464	0.201440	0.128418	0.007765	F
b8	0.168777	300.750458	298.205933	0.155936	0.145052	0.157183	0.242233	0.290206	0.236275	0.148758	0.007238	F
afd	0.133703	296.563812	294.972321	0.116205	0.118793	0.104201	0.387968	0.222333	0.136868	0.109767	0.003128	G
c2	0.129519	295.394073	293.930420	0.110509	0.112607	0.092614	0.407458	0.203739	0.116085	0.102388	0.003120	G
...	...	...	...	...	...	...	...	...	...	...	...	...
72	0.128363	294.559326	292.820374	0.107029	0.095681	0.081073	0.260530	0.175331	0.102809	0.087734	0.007360	V
16	0.142596	298.901337	296.929138	0.124923	0.122501	0.113922	0.337189	0.229040	0.165208	0.108735	0.006239	V
7a	0.132320	296.620697	294.638184	0.111956	0.108085	0.087173	0.339684	0.179832	0.102980	0.098184	0.006388	V
23	0.129731	293.359009	291.667206	0.108747	0.101903	0.076744	0.332543	0.177821	0.101857	0.090330	0.007383	V
24	0.129616	293.604187	291.715271	0.108449	0.099904	0.074216	0.374737	0.176227	0.097688	0.088046	0.007557	V

3.2.4 Crop Water Supply

The natural hydrologic system is a cyclic water environment of the earth, including the precipitation, evaporation, transpiration, streamflow, and underground water. In the irrigation section, we need to compare the crop water supply (CWS) against the crop water requirement (CWR), both consisting of several separated data sources. For the crop water supply (CWS) in figure 16, precipitation and underground water pumping are included. For the crop water requirement (CWR), evaporation and runoff are considered.

Figure 16

Crop Water Supply Structure

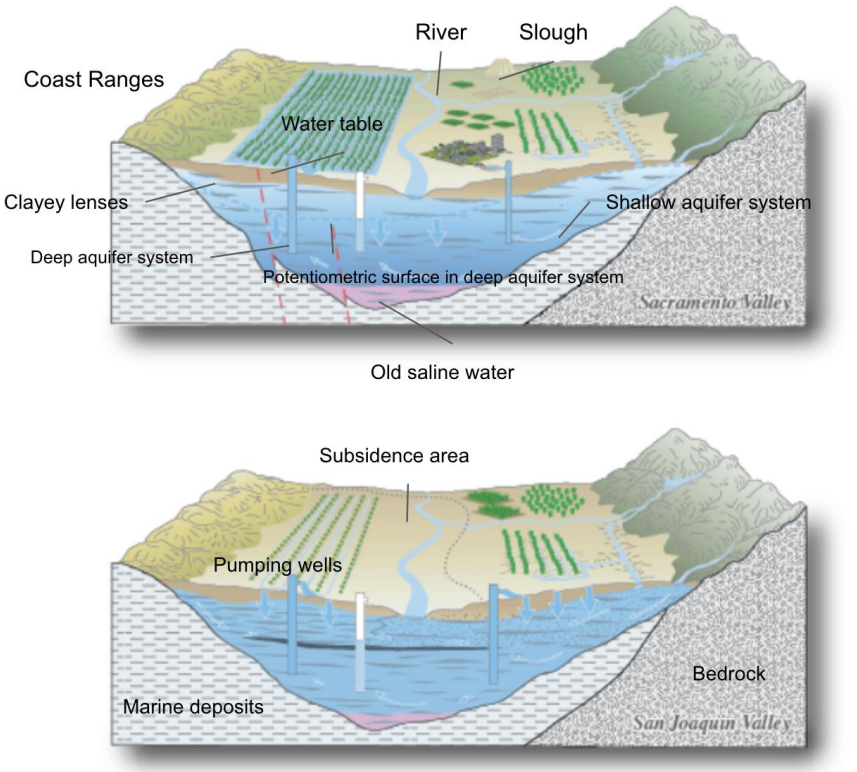


1. Precipitation and Snowfall

The definition of precipitation is the accumulated water amount that falls to the land surface. The amount of precipitation does not include the fog, dew, or other precipitation that has not fallen to the land ground before evaporating which is shown in figure 17. The unit of the precipitation measurement is the meter and is assumed to be spread evenly to the land. In our crop water supply system, we did not include snowfall in the precipitation parameter.

Figure 17

Hydrologic System - Precipitation Define



Post-development hydrogeology of the Sacramento and San Joaquin Valleys, California.

We extracted precipitation bands from ERA5-Land, which is a reanalysis dataset by combining the model data with the observation data. The precipitation parameter we used is the 'total_precipitation' band, which is the sum of evaporation from bare soil, evaporation from open water surfaces, evaporation from the canopy, evaporation from vegetation transpiration, and snow evaporation (Table 14).

Table 14

Description of Precipitation Bands from ERA5-Land

Description of Bands	Term of Use	Citation
evaporation_from_bare_soil	The amount of evaporation from land surface soil.	m of water equivalent
evaporation_from_open_water_surfaces_excluding_oceans	The amount of evaporation from the water surface, such as lakes but excluding oceans.	m of water equivalent

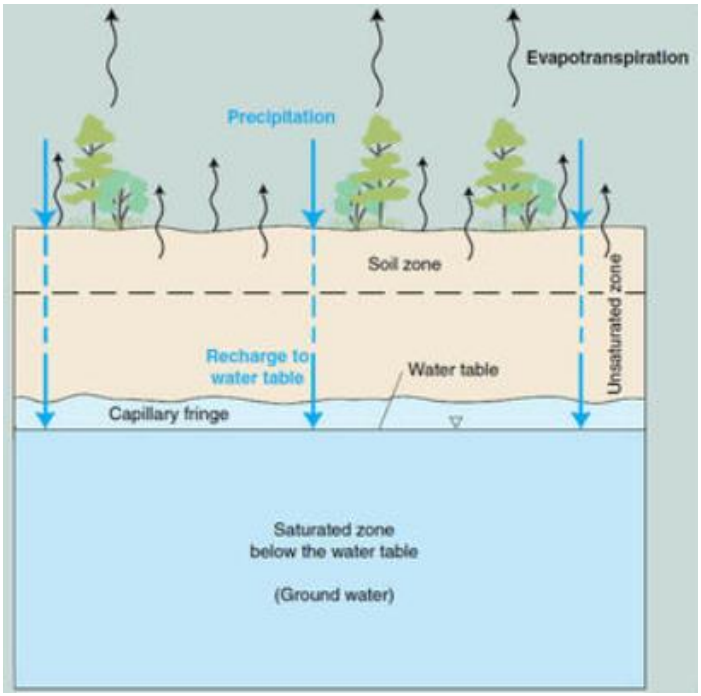
evaporation_from_the_top_of_canopy	The amount of evaporation from the canopy interception reservoir.	m of water equivalent
evaporation_from_vegetation_transpiration	Amount of evaporation from root extraction i.e., the amount of water extracted from the different soil layers.	m of water equivalent
snow_evaporation	Evaporation from snow averaged over the grid box.	m of water equivalent

2. Underground Water

Underground water is the water held by soil or rock under the land surface. Water is pulled through soil layers because of gravity, first reaching the upper layer' zone of saturation or called 'water table', then going down to the deep layer 'aquifer' beneath the water table. Water in the aquifer moves slowly and recharges into land surfaces such as rivers, lakes, and oceans which is shown in figure 18.

Figure 18

Underground Water Moving Cycle



The underground water can be pumped out when the aquifer is shallow, and the water is able to move through the aquifer layer. The height of the water table is greatly influenced by the precipitation patterns, climate, land surface changes, and human activities. We separated the soil above the groundwater into four layers, the shallowest layer is layer one and the deepest layer is layer four (Table 15). When layer one contains more water than layer four, the water table tends to become higher, meaning the excessive water is going to charge the water table and vice versa (Figure 19).

Figure 19

Process of Four Soil Layers (Earle, 2015)

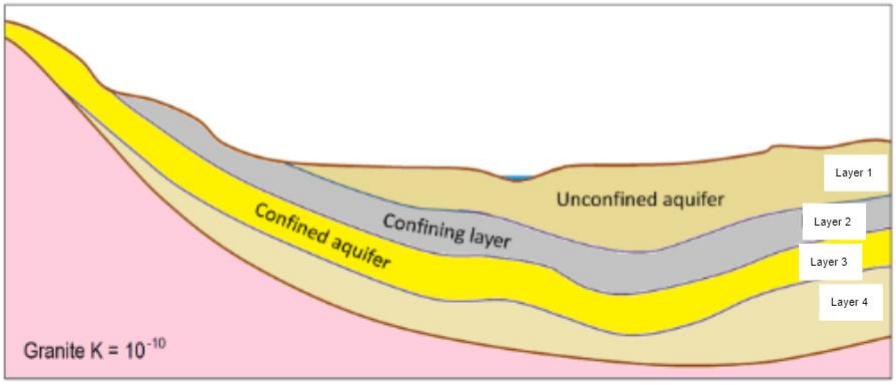


Table 15

Descriptions of Soil Layers

Layers	Citation	Term of Use
layer_1	m^3/m^3	Based on the ECMWF Integrated Forecasting System Volume of water in soil layer 1 (0 - 7 cm)
layer_2	m^3/m^3	Based on the ECMWF Integrated Forecasting System Volume of water in soil layer 2 (7 -28 cm)
layer_3	m^3/m^3	Based on the ECMWF Integrated Forecasting System Volume of water in soil layer 3 (28-100 cm)
layer_4	m^3/m^3	Based on the ECMWF Integrated Forecasting System

		Volume of water in soil layer 4 (100-289 cm)

3.2.5 Crop Water Requirement

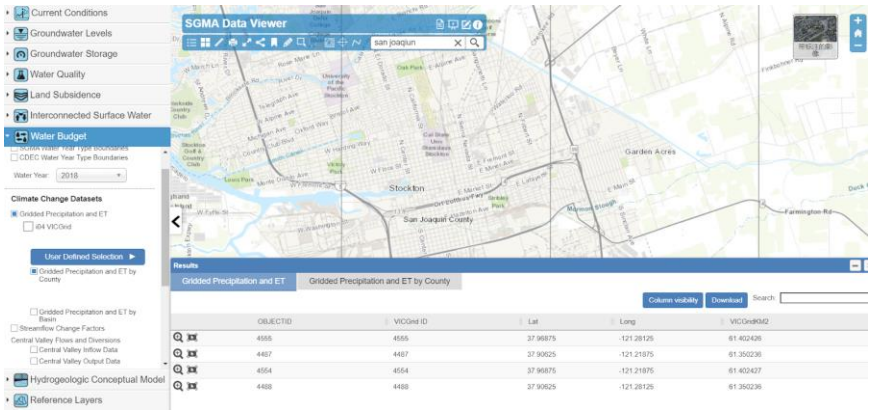
1. Land surface evapotranspiration (ET)

The evapotranspiration system includes two parts, one is the soil evaporation into the air, and another is the vegetation transpiration into the atmosphere. Monitoring and quantifying the water demand for the farmland is crucial to balance the hydrologic water and plan the precision irrigation system.

The source of the land surface ET data is from SGMA data viewer (Figure 20), which is an interactive tool to provide data related to groundwater in California, including current conditions, groundwater levels, groundwater storage, water quality, land subsidence, interconnected surface water, water budget, hydrogeologic conceptual model, and reference layers. And ET data belongs to the climate change dataset under the 'Water Budget' section.

Figure 20

SGMA Data Viewer



We selected the gridded precipitation and ET categorized by county, the time range is from 1915 to 2011, and the collecting frequency is once a month.

Figure 21

Description of Evaporation and Precipitation Dataset

```
df_all= pd.read_csv('sanj_climate_soil_data_v3.csv')
df_all
```

total_evaporation	total_precipitation	runoff	volumetric_soil_water_layer_1	volumetric_soil_water_layer_2	volumetric_soil_water_layer_3	volumet
-0.002283	0.00169	0.000232	0.340073	0.34613	0.36882	
-0.002518	0.001521	0.000191	0.30275	0.310059	0.336517	
-0.002518	0.001521	0.000191	0.30275	0.310059	0.336517	
-0.002482	0.001606	0.000412	0.258408	0.26799	0.296051	
-0.002482	0.001606	0.000412	0.258408	0.26799	0.296051	
...	...	...	...	...	...	
-0.002354	0.002512	0.00215	0.432325	0.440839	0.436185	
38.24690289390949]	[-121.0889626289887	38.23562216551106]	38.24838770201417]	[-121.08895370926558	38.24895499800474]	
-0.002478	0.002058	0.000442	0.443161	0.453354	0.470551	
121.28370041232371	38.07724225643089]	[-121.29471000481306	[-121.28327235560718	38.07750547723604]	[-121.28288891045293	
121.24820596415267	38.12657427178944]	[-121.25006549344954	[-121.24852708066182	38.12717407330584]	[-121.2486429688364	

2. Runoff

Runoff is a measurement of the ability to hold water by the soil, it is an effective indicator of flood or drought in the land (Table 16). The drainage of water from the land surface is called the surface runoff, while the drainage of water from the underground level is called the subsurface runoff. Both of these two subtype runoffs fall into the runoff category. The measurement of the unit is the depth in meters, which is assumed to be spread evenly across the grid area.

Table 16

Description of Runoff

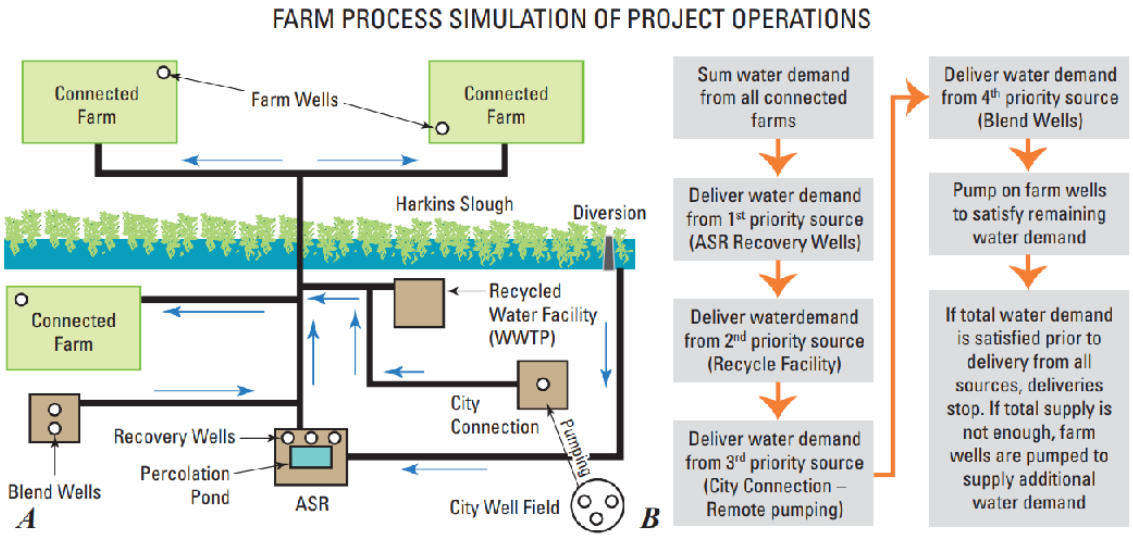
Description of Runoff	Term of Use	Citation
sub_surface_runoff	The water stored in the soil drains away from the sub land surface.	meter
surface_runoff	The water stored in the soil drains away from the land surface.	meter

3. Heat flux

Heat flux is the thermal energy flow over the heat transfer surface area (Figure 22). The parameter is measured by J/m^2, and the downward fluxes are positive.

Figure 22

Water Inflows and Outflows from a Farm



3.2.6 Fertilizer Recommendation Requirement

The Nitrogen prediction training data is taken directly from a research paper titled "Total nitrogen estimation in agricultural soils via aerial multispectral imaging and LIBS sensor by Hossen et al.(2021). The data set includes data taken from a drone mounted sensor providing sensor data that includes features such as Red (Red band pixel value), NIR (Near Infra-Red Pixel Value), Green (Green Band Pixel Value), NDVI (Normalized Differential Vegetation Index), RH (Related Humidity), Air Temperature. The data set include label in terms of actual Nitrogen content in soil in ppm. The training data set is shown in Figure 23 below.

Figure 23

Soil Nitrogen Content Prediction Dataset - Training dataset

Data count	Red (DN)	NIR (DN)	Green (DN)	NDVI	RH (%)	Air temp (C)	TN (ppm) at 493.4 nm
0	47.88	328.05	54.90	0.74	33.8	23.2	1179.39
1	51.99	329.97	56.50	0.72	33.8	23.2	1136.64
2	52.37	325.90	56.14	0.72	33.8	23.2	1093.90
3	55.81	200.32	43.11	0.56	33.8	23.2	418.46
4	59.30	209.11	45.47	0.55	33.8	23.2	413.75
5	56.36	208.75	44.72	0.57	33.8	23.2	409.28
6	74.64	223.63	52.27	0.50	33.8	23.2	1139.47
7	78.95	229.52	54.50	0.49	33.8	23.2	1098.66
8	79.21	230.47	55.09	0.49	33.8	23.2	1059.94
9	85.53	226.03	55.86	0.44	33.8	23.2	760.72
10	98.47	238.19	61.73	0.41	33.8	23.2	738.83
11	102.15	247.70	64.07	0.41	33.8	23.2	718.12
12	44.84	287.68	49.64	0.73	33.8	23.2	927.98

3.3 Data Pre-processing

3.3.1 Vegetation Index

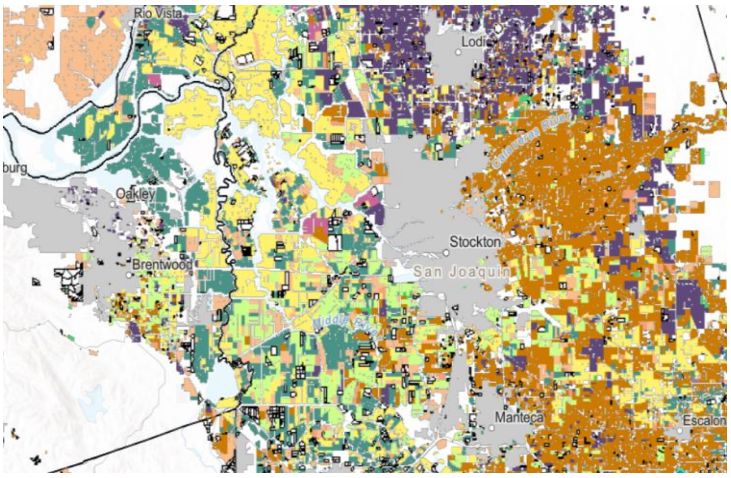
The polygons delineated in the land use map came from the digital images of the USDA National Agricultural Imagery Program (NAIP) in one-meter resolution. The boundaries were digitized by using ArcGIS with NAIP imagery as the base map. CDWR staff visit all the agricultural land on the map and enter the land use into their Surface Pro tablet computers. Furthermore, they confirm their locations by checking their points in the global positioning system (GPS). The survey map quality was double-checked by the DWR office, with a 95% accuracy of the land use attribute. And the max digital line error is six meters.

In order to compute the features within one polygon, we stack two input layers on top of each other, one is the crop features separated by polygon, and another is the remote sensing features (Figure 24). Then we are able to count and average the vegetation indices based on features that fall in the input polygon layer. Compared to the previous vegetation index calculation methods, such as pixel-by-pixel calculation or pixel-value average in one grid, the pixels of one single polygon can be summarized based on the boundary layer (Figure 25). We

average the NDVI in the line-up area, by using the Google Earth Engine map() function and reduceRegion() function.

Figure 24

Polygons Boundary to Separate San Joaquin County Region & Crop Identification Polygons



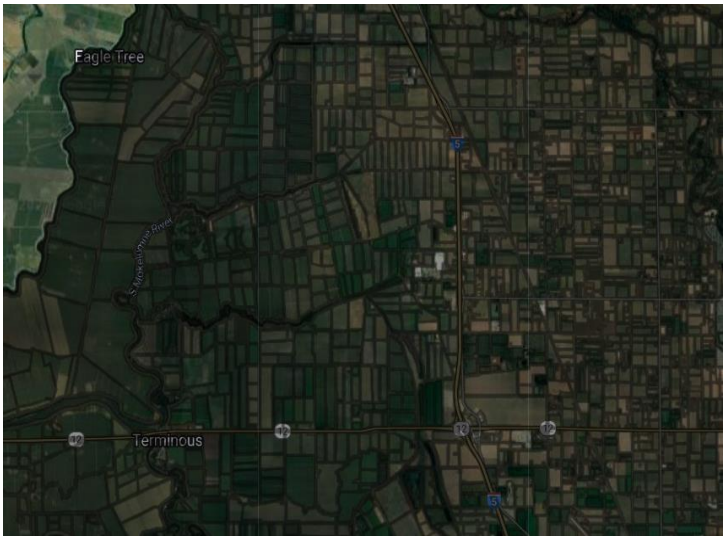
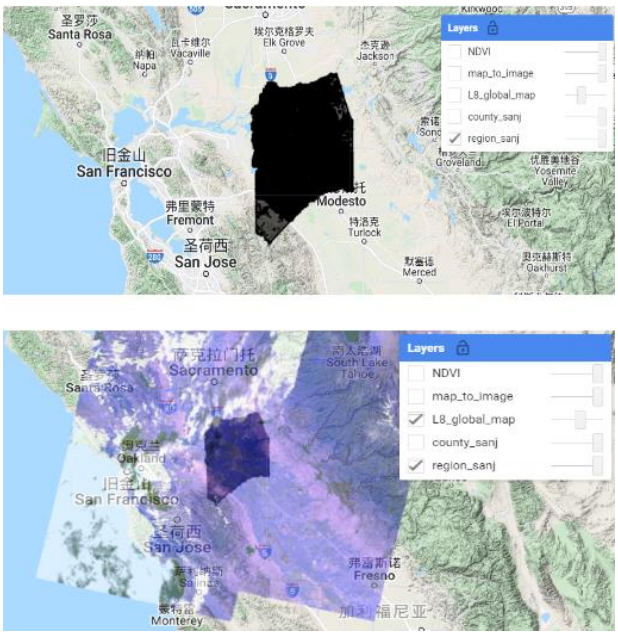


Figure 25

Image Layers Presenting in Google Earth Engine



We calculated the vegetation indices based on satellite bands by using Python Pandas. Each index has three sub-categories which are the spring index, summer index, and winter index, since the reflection amounts and absorbed amounts of vegetation are different in each season. The data extracting date of spring is April 1st to April 15th, the data extracting date of summer is August 1st to August 15th, and the data extracting date of winter is November 1st to November 15th.

According to the data descriptions, we found that the mean value of vegetation indices from summer is the highest, and the winter mean values are the lowest. Figure 26 shows the reason is that vegetation in summer is more intense, which reflects a more near-infrared spectrum back to the remote sensor.

Figure 26

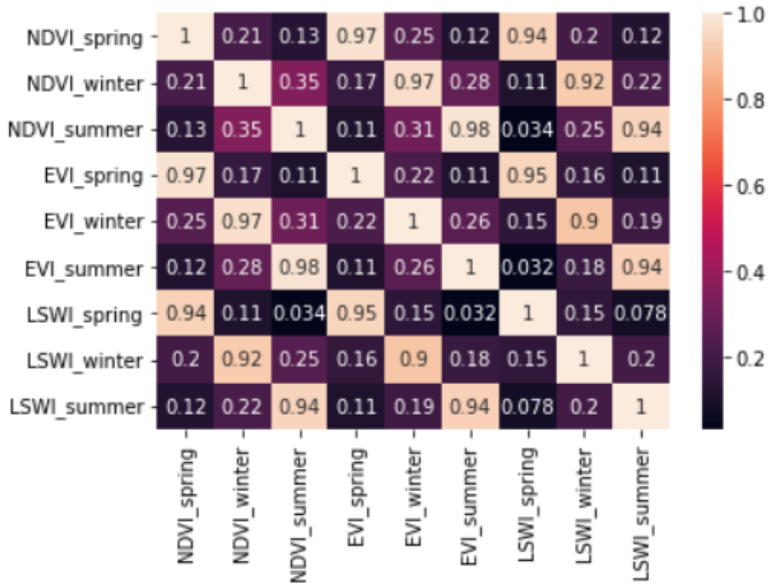
Vegetation Indices Based on Different Seasons

	NDVI_spring	NDVI_winter	NDVI_summer	EVI_spring	EVI_winter	EVI_summer	LSWI_spring	LSWI_winter	LSWI_summer
count	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000
mean	0.473924	0.377495	0.486682	0.489394	0.325543	0.517507	0.151813	0.063543	0.172386
std	0.168689	0.148413	0.149200	0.219290	0.160264	0.194918	0.182018	0.144258	0.154604
min	0.010267	-0.126886	0.009280	0.005326	-0.039447	0.005614	-0.276644	-0.370577	-0.250734
25%	0.344299	0.247619	0.382187	0.319100	0.202773	0.378850	0.003067	-0.049286	0.058901
50%	0.467552	0.389071	0.498343	0.470352	0.316161	0.512680	0.147883	0.063193	0.163660
75%	0.609944	0.485057	0.600469	0.633806	0.418246	0.645493	0.287116	0.170088	0.281541
max	0.846786	0.815658	0.814815	1.168279	1.096733	1.187912	0.592300	0.720568	0.570146

Figure 27 shows the relationships between different indices from one season, such as NDVI_spring and EVI_spring, which are highly correlated, such as the coefficient of NDVI_spring and EVI_spring is 0.97. While in the same seasons, the indices in the different categories are in low correlation, such as the coefficient of NDVI_spring and NDVI_winter is 0.21.

Figure 27

Coefficient of Vegetation Indices



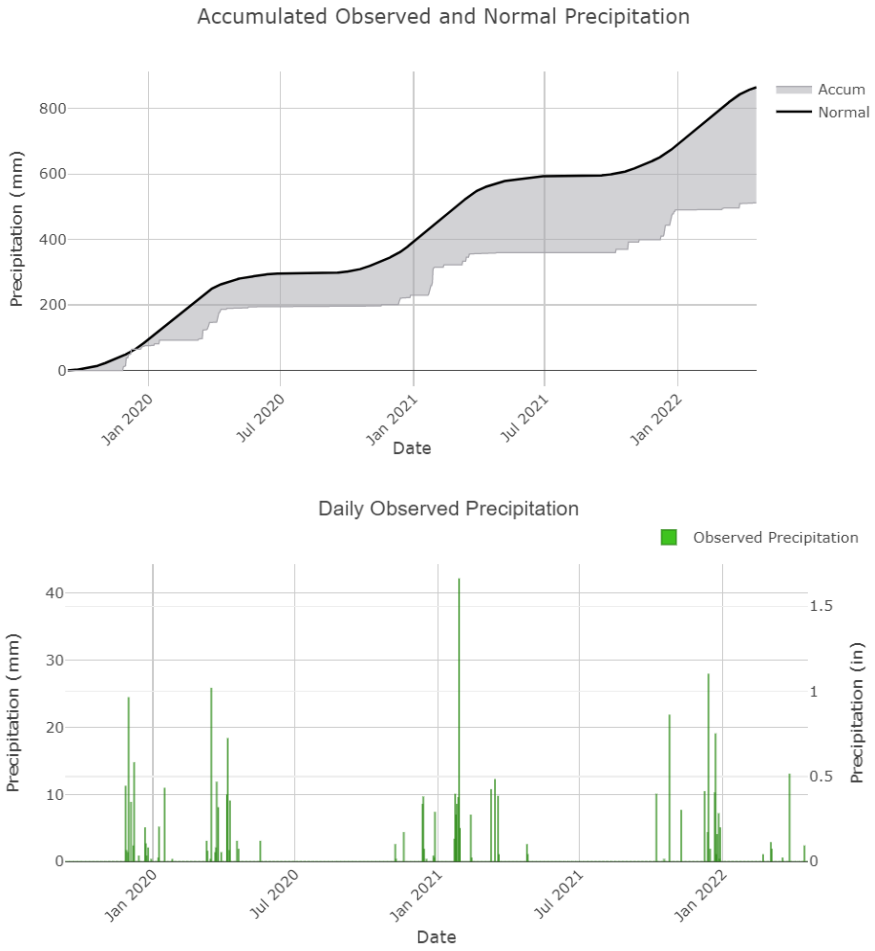
3.3.2 Crop Water Supply

1. Precipitation

The time series of precipitation in the following figures are collected from San Joaquin Valley, California, with around 1000 days from 2019 to 2022. For the missing values in the time series, we filled the blank by using the rolling average of the previous ten days. Moreover, we checked the data distribution in order to confirm that there are no outliers in the time series (Figure 28).

Figure 28

Precipitation Is Not Distributed Evenly During the Whole Year



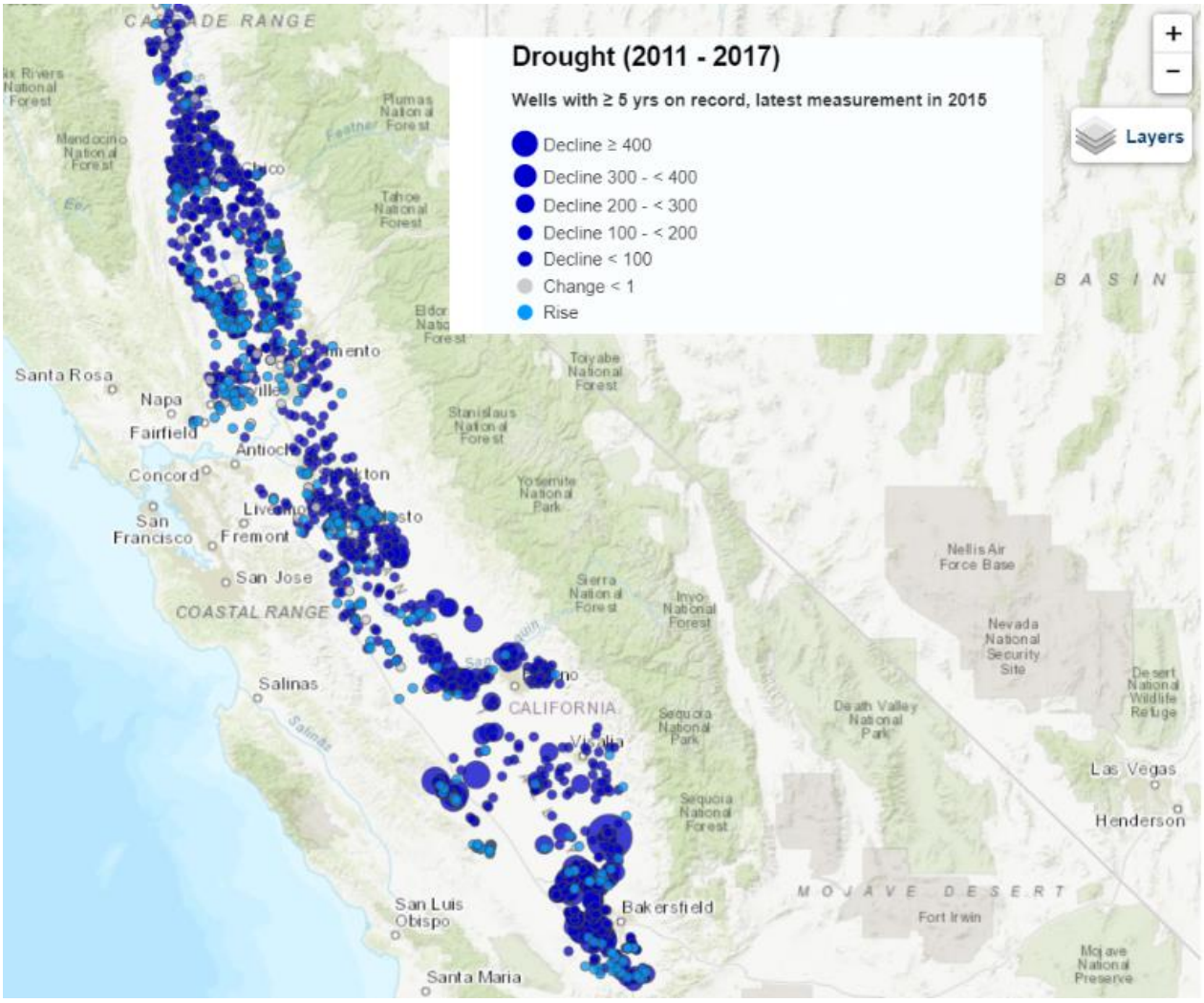
3. Underground Water

For the measurement of underground water, we utilized the “soil_water_layer” and level of wells to present the drought situation in our AOI. The “soil_water_layer” dataset covers the

whole American continent, other regions other than San Joaquin County can be used as the test dataset. The latest measurement date of a good dataset is 2015, with wells used for more than 5 years. The time range for the good dataset is from 1953 to 2015, and the data source is DWR (Figure 29).

Figure 29

Drought Indicators from 2011-2017 (California Water Science Center, n.d.)



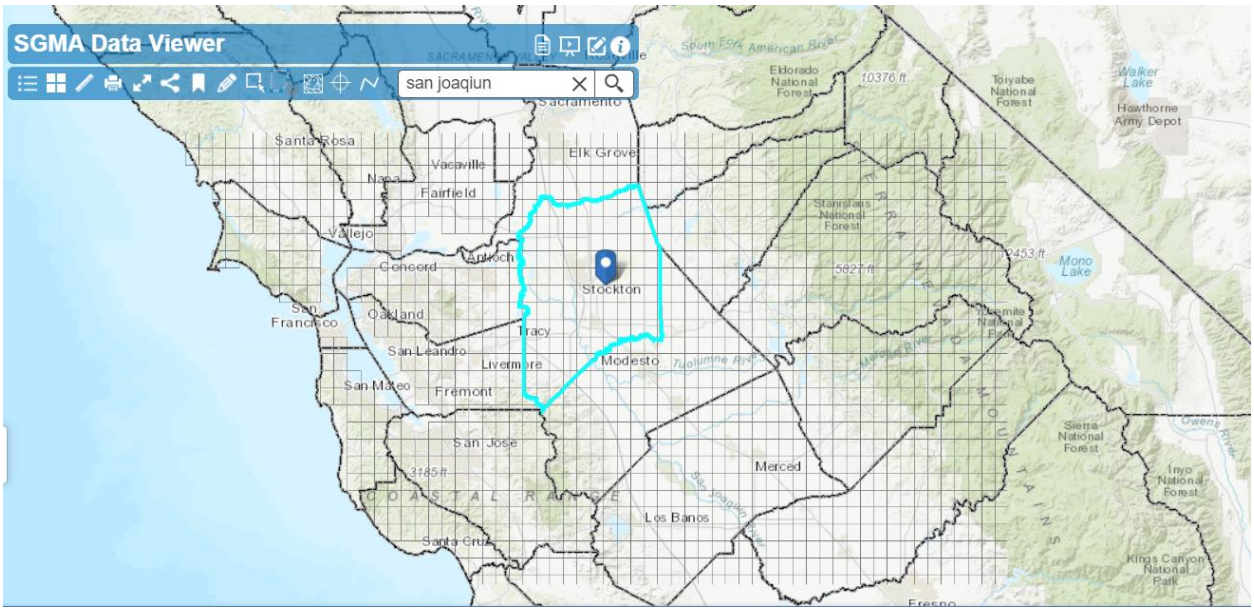
3.3.3 Crop Water Demand: Land Surface Evapotranspiration (ET)

Precipitation data is calculated based on the Variable infiltration capacity (VIC) model, and the output of the VIC model is the evapotranspiration data. The whole dataset is acquired

from the grid covering San Joaquin Valley at a 1/16-degree resolution. In order to get the ET value of San Joaquin County, we averaged all the ET values of each grid within the boundary of San Joaquin County (Figure 30).

Figure 30

Grid Mapping to Extract Precipitation and Evaporation Data



ET is dependent on the energy balance of the land surface which is calculated from satellite images. The land surface energy consumed by evapotranspiration is the difference between net spectral radiation (both short-wave and long-wave features) and heat flux. The heat flux number is the sum of the absorbed amount into the ground and the converted amount into the air.

$$LE = Rn - G - H$$

where:

LE is the energy consumed by evapotranspiration

Rn is the incoming radiation minus outgoing radiation

G is the heat flux absorbed into the ground

H is the heat flux converted into the air

The ET is the multiple of LE divided by ρw(density of water) and lambda(heat of vaporization).

$$ET_{inst} = 3{,}600\ \frac{LE}{\lambda p_w}$$

3.3.3 Fertilizer Recommendation

For the fertilizer recommendation, we used the features available in data set including 'RED (DN)','NIR (DN)','Green (DN)', 'NDVI'. The label in the data set 'Nitrogen (ppm)' indicates the actual Nitrogen estimation in the soli in parts per million (ppm). The label in the data set is a continuous value with range in between 0 ppm and 2000 ppm. The data set is at two frequency label 668nm and 621 nm. For the purpose of this project, we chose to go with 668nm frequency as this frequency matches with the Satellite sensor data frequency.

3.4 Data Transformation

3.4.1 Vegetation Index

1. NDVI

NDVI is an agriculture index to estimate the density of green on land. It is a popular measurement method since the inputs for NDVI come from the public remote sensing data, most are spectral bands of satellites. The logic behind the NDVI is based on the differences in plant reflectance. The pigment in plant leaves absorbs red light (red band in satellite image) but reflects near-infrared light. It means the greener segments on the land, the more difference between red waves and near-infrared waves. Moreover, NDVI is one way to detect the health of vegetation. Healthy vegetation absorbs abundant amounts of visible light and reflects a big portion of the near-infrared light, while unhealthy or sparse land absorbs less visible light and reflects less near-infrared light, making the light difference smaller.

Band 3 of the satellite image represents the visible light which is also taken as the red band, and Band 4 is called for the near-infrared band. The formula of NDVI is near-infrared band minus visible band divided by the near-infrared band plus visible band.

$$NDVI = (NIR - VIS)/(NIR + VIS)$$

$$= (Band4 - Band3)/(Band4 + Band3)$$

2. EVI

Enhanced vegetation index was adopted from two MODIS satellites for the purpose of eliminating the environment and background noise shown in NDVI. The formula of EVI is similar to the NDVI's, but with a correction to the mistook reflected light which is caused by land ground and air particles.

$$EVI = G \cdot \frac{(NIR - Red)}{(NIR + C_1 \cdot Red - C_2 \cdot Blue + L)}$$

Where:

NIR / Blue / Red are surface reflectance extracted from satellite images

L is the background adjustment, here we assume L equals 1

C is the coefficient to correct the red band reflectance

3. LSWI

LSWI presents the difference in short-wave reflectance (NIR and SWIR), both are sensitive to the soil humidity and vegetation moisture. This index is used as the moisture indicator of the soil (Chandrasekar et.al, 2010).

$$LSWI = (NIR - SWIR)/(NIR + SWIR)$$

$$= (Band4 - Band5)/(Band4 + Band5)$$

By applying the calculation of NDVI, EVI, and LSWI, it is possible to calculate the vegetation indices based on the given bands from the dataset which is shown in Figure 31.

Figure 31

Vegetation Indices Calculated Based on Bands Dataset Before Scaling

	NDVI_spring	NDVI_winter	NDVI_summer	EVI_spring	EVI_winter	EVI_summer	LSWI_spring	LSWI_winter	LSWI_summer
count	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000	21895.000000
mean	0.473924	0.377495	0.486682	0.489394	0.325543	0.517507	0.151813	0.063543	0.172386
std	0.168689	0.148413	0.149200	0.219290	0.160264	0.194918	0.182018	0.144258	0.154604
min	0.010267	-0.126886	0.009280	0.005326	-0.039447	0.005614	-0.276644	-0.370577	-0.250734
25%	0.344299	0.247619	0.382187	0.319100	0.202773	0.378850	0.003067	-0.049286	0.058901
50%	0.467552	0.389071	0.498343	0.470352	0.316161	0.512680	0.147883	0.063193	0.163660
75%	0.609944	0.485057	0.600469	0.633806	0.418246	0.645493	0.287116	0.170088	0.281541
max	0.846786	0.815658	0.814815	1.168279	1.096733	1.187912	0.592300	0.720568	0.570146

3.4.2 Crop Water Requirement (CWR)

For the crop water requirement (CWR), we combined three water output parameters as the final result, one is the evaporation value, another is the total runoff value and the underground water amount. The negative value for evaporation means the water flows out of the land surface, and the positive value for runoff represents the water running away from the soil layer, "ET_runoff" is the sum-up of evaporation and runoff. Then we added up the underground water value to the "ET_runoff" to get the total crop water requirement (Figure 32).

$$CWR = Evaporation - Runoff + Underground\ Water$$

$$= Evaporation - Runoff + Soil\ water\ layer\ 4 - Soil\ Water\ Layer\ 1$$

Figure 32

Calculating Crop Water Requirement

```
# crop water requirement
# first part --- ET and runoff
df_clean['ET_runoff'] = df_clean['total_evaporation'] - df_clean['runoff']

# second part --- underground water
df_clean['underground_water'] = df_clean['volumetric_soil_water_layer_4'] - df_clean['volumetric_soil_water_layer_1']
df_clean
```

```
# calculate crop water requirement

df_clean['CWR'] = df_clean['ET_runoff'] + df_clean['underground_water']
```

Figure 33 shows the processed dataset contains 3206 rows with eight input features. The range of CWS and ET_runoff is small and around zero, while the deviations of underground water and CWR are more apparent with some outliers far away from the mean values. So we decided to delete the outliers for underground water and CWR in order to have a better fitting model. We have made the boxplot for Crop Water Requirements as shown in figure 34.

Figure 33

Dataset for Crop Water Requirement

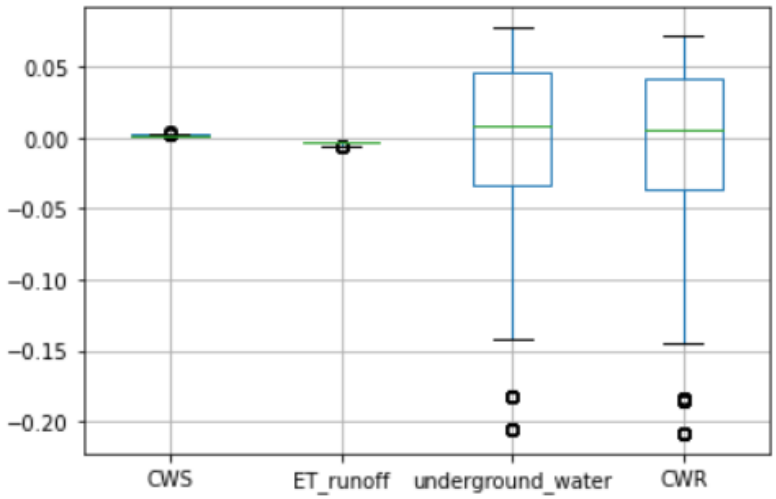
	CWS	ET_runoff	underground_water	CWR	surface_sensible_heat_flux	skin_temperature	u_component_of_wind_10m	NDVI
0	0.001690	-0.002515	0.004623	0.002108	-1900802.000	285.825851	0.928651	0.197296
1	0.001521	-0.002709	-0.025055	-0.027764	-922280.000	285.876633	1.283631	0.155782
2	0.001521	-0.002709	-0.025055	-0.027764	-922280.000	285.876633	1.283631	0.153965
3	0.001606	-0.002894	0.005706	0.002812	-1048284.000	285.620773	1.314393	0.146493
4	0.001606	-0.002894	0.005706	0.002812	-1048284.000	285.620773	1.314393	0.165946
...	...	...	...	...	...	...	...	...
3770	0.002084	-0.002741	-0.119888	-0.122629	-1500316.000	285.970383	1.067323	0.257687
3771	0.002084	-0.002741	-0.119888	-0.122629	-1500316.000	285.970383	1.067323	0.169968
3772	0.002334	-0.003143	-0.101700	-0.104843	-1463576.000	285.845383	0.908631	0.210841
3773	0.002512	-0.004504	-0.002462	-0.006966	-1430555.667	285.619797	0.809266	0.283981
3775	0.002058	-0.002920	-0.044182	-0.047102	-1480625.000	285.799484	1.022157	0.268732

3206 rows × 8 columns

Figure 34

Boxplot for Crop Water Requirements

```
1 boxplot = df_clean_v2.boxplot(column=['CWS','ET_runoff','underground_water','CWR'])
```



3.4.3 Fertilizer recommendation

As mentioned in the data preprocessing section, the label available in the training data set is a continuous value and hence is not suited for classification models. Hence as part of data transformation, we converted the continuous variable to discrete value. The discrete values were grouped in 3 'Tier' ranges. Low – 1 (0-500 ppm), Medium-2 (501 – 1000 ppm) and High-3 (>=1001 ppm). The processed dataset is shown the figure 35.

Figure 35

Transformed data sample for fertilizer recommendation model

	RED (DN)	NIR (DN)	Green (DN)	ndvi	RH (%)	Temp (C)	Nitrogen (ppm)	tier
0	47.88	328.05	54.90	0.74	33.8	23.2	1156.89	3
1	51.99	329.97	56.50	0.72	33.8	23.2	1118.38	3
2	52.37	325.90	56.14	0.72	33.8	23.2	1079.95	3
3	55.81	200.32	43.11	0.56	33.8	23.2	466.82	1
4	59.30	209.11	45.47	0.55	33.8	23.2	462.82	1
5	56.36	208.75	44.72	0.57	33.8	23.2	459.04	1
6	74.64	223.63	52.27	0.50	33.8	23.2	1110.15	3
7	78.95	229.52	54.50	0.49	33.8	23.2	1073.99	3
8	79.21	230.47	55.09	0.49	33.8	23.2	1039.71	3
9	85.53	226.03	55.86	0.44	33.8	23.2	996.15	2
10	98.47	238.19	61.73	0.41	33.8	23.2	965.72	2
11	102.15	247.70	64.07	0.41	33.8	23.2	936.81	2

3.5 Data Preparation

1. Crop Identification

We divided the crop identification dataset into training and testing datasets in Figure 36 and Figure 37, with 80% for training and 20% for testing. Since the label output is abnormal distribution, we stratified the training and testing dataset according to the weights of crop amounts. For the training dataset, there are 17516 rows with nine input features. For the testing dataset, there are 17516 rows with one output feature.

Figure 36

Stratified Dataset - Training Dataset

```
# stratify=df_target

x_train, x_test, y_train, y_test = train_test_split(df_features, df_target, test_size = 0.2, random_state = 11, stratify=df_target)
```

```
x_train
```

system:index	NDVI_spring	NDVI_winter	NDVI_summer	EVI_spring	EVI_winter	EVI_summer	LSWI_spring	LSWI_winter	LSWI_summer
000100000000000038f8	0.340210	0.409049	0.527692	0.320762	0.319692	0.572391	-0.036452	-0.026925	0.135792
0000000000000000c57a	0.209074	0.174329	0.621845	0.136952	0.119284	0.671988	-0.173304	-0.164359	0.325930
7236	0.207719	0.224400	0.587498	0.205883	0.187562	0.698182	-0.071676	-0.153021	0.320080
00000000000000007ca5	0.283647	0.166021	0.166845	0.237607	0.108709	0.141699	0.055115	-0.006653	-0.031206
0000000000000000b7b1	0.758214	0.129737	0.184803	0.917240	0.082745	0.155191	0.532155	-0.076073	-0.051751
...	...	...	...	...	...	...	...	...	...
000000000000000982d	0.587153	0.595019	0.502679	0.586758	0.558305	0.504070	0.267359	0.318840	0.199169
0000000000000009a8b	0.777277	0.642079	0.786970	1.013206	0.641587	1.081008	0.421254	0.223229	0.429854
000100000000000012b7	0.319338	0.249350	0.297612	0.314204	0.225890	0.296985	-0.026453	-0.082892	-0.055930
0000000000000000054f6	0.568355	0.412680	0.571514	0.643521	0.396435	0.678827	0.205342	0.099959	0.264490
000100000000000040ba	0.250077	0.419120	0.520290	0.218261	0.286053	0.556070	-0.122995	0.053586	0.166952

17516 rows × 9 columns

Figure 37

Stratified Dataset - Testing Dataset

```
1 y_train
```

```
system:index
00010000000000038f8    V
0000000000000000c57a    F
7236                    T
00000000000000007ca5    T
0000000000000000b7b1    G
                      ..
00000000000000000982d    D
0000000000000000a8b    P
000100000000000012b7    V
0000000000000000054f6    P
000100000000000040ba    V
Name: CLASS1_spring, Length: 17516, dtype: object
```

2. Irrigation Recommendation

We divided the irrigation dataset into training and testing datasets in Figure 38 and Figure 39, with 80% for training and 20% for testing. As the output feature is not skewness apparently, we did not use the stratify function. For the training dataset, there are 2030 rows with six input features. For the testing dataset, there are 2030 rows with one output feature.

Figure 38

Preprocessing Irrigation Dataset - Training Dataset

```
1 x_train=x_train.drop(['const'],axis=1)
2 x_train
```

	ET_runoff	underground_water	CWR	surface_sensible_heat_flux	skin_temperature	u_component_of_wind_10m
654	-0.390413	0.469797	0.469887	0.932158	-0.502209	-1.293970
2002	0.736316	-1.086674	-1.091059	1.710754	0.775520	0.871377
1993	0.899230	-1.659147	-1.671477	1.538746	0.382373	0.835446
2001	0.899230	-1.659147	-1.671477	1.538746	0.382373	0.835446
821	0.541871	-0.293509	-0.286102	1.228408	-0.512037	0.990519
...	...	...	...	...	...	...
1638	-0.274797	0.736154	0.744735	-0.520890	-0.394093	-0.955457
1095	0.888719	-1.524190	-1.533948	0.733453	0.313572	0.338078
1130	1.001182	-0.891122	-0.884765	-0.991369	0.480659	0.061972
1294	0.731061	-0.250217	-0.237161	-0.190991	0.490488	-1.031102
860	-0.315788	1.487238	1.510570	-0.599717	1.768216	0.442090

2030 rows × 6 columns

Figure 39

Preprocessing Irrigation Dataset - Testing Dataset

```
1 y_train

654      0.564817
2002    -1.081527
1993    -0.890157
2001    -0.890157
821     -0.842314
           ...
1638     0.857500
1095    -0.898600
1130    -0.878900
1294    -0.225991
860     -0.656573
Name: CWS, Length: 2030, dtype: float64
```

3. Fertilizer Recommendation

We divided the fertilizer transformed dataset into training and testing datasets with 80% for training and 20% for testing. As the output feature is not skewed apparently, we did not use the stratify function. For the training dataset, there are 43 rows with 4 input features. For the testing dataset, there are 11 rows with one output feature. The data of the first model is combined with the crop identification model data and subjected to a MLP model. The size of data is 17620 rows which is then split into training and testing dataset in the ratio of 80:20, thus training data size is 14096 while the testing data size is 3524.

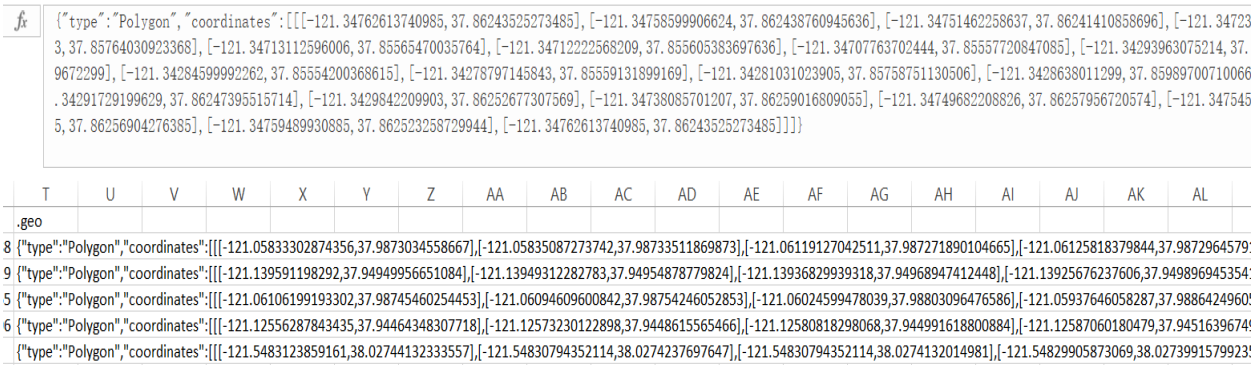
3.6 Data Statistics

3.6.1 Crop Identification

The raw data for the crop category comes from CADWR for the San Joaquin region. The whole dataset is downloaded into Shapefile format with the geographic location information. Each cropland is delineated by line and surrounded as a polygon on the map. There are 74797 polygons in the region. We used Google Earth Engine to open the shapefile and named that map layer as “Region_Sanj”. Then we added the bands layer from the Landsat 8 remote sensor, called “L8_global_map”. All the bands’ data are extracted based on the polygon boundary on the layer “Region_Sanj” as shown in Figure 40.

Figure 40

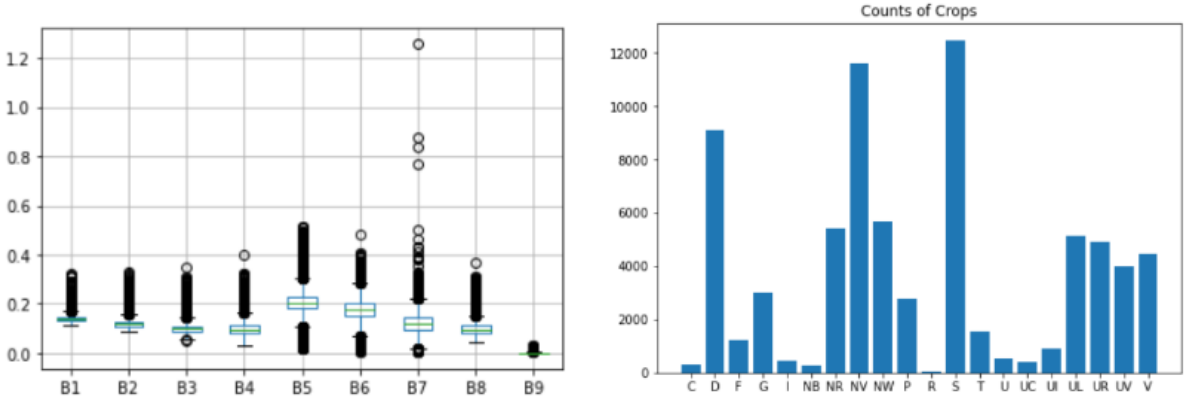
Coordinates For the Polygons in San Joaquin County



	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
	.geo																		
8	{"type":"Polygon","coordinates":[[[-121.05833302874356,37.9873034558667],[-121.05835087273742,37.98733511869873],[-121.06119127042511,37.987271890104665],[-121.06125818379844,37.9872964579																		
9	{"type":"Polygon","coordinates":[[[-121.139591198292,37.94949956651084],[-121.13949312282783,37.94954878779824],[-121.13936829939318,37.94968947412448],[-121.13925676237606,37.949896945354																		
5	{"type":"Polygon","coordinates":[[[-121.06106199193302,37.98745460254453],[-121.06094609600842,37.98754246052853],[-121.06024599478039,37.98803096476586],[-121.05937646058287,37.9886424960																		
6	{"type":"Polygon","coordinates":[[[-121.12556287843435,37.94464348307718],[-121.12573230122898,37.9448615565466],[-121.12580818298068,37.944991618800884],[-121.12587060180479,37.9451639674																		
	{"type":"Polygon","coordinates":[[[-121.5483123859161,38.02744132333557],[-121.54830794352114,38.0274237697647],[-121.54830794352114,38.0274132014981],[-121.54829905873069,38.0273991579923																		

Figure 41

Spectral Bands Boxplots & Crop Type Counts

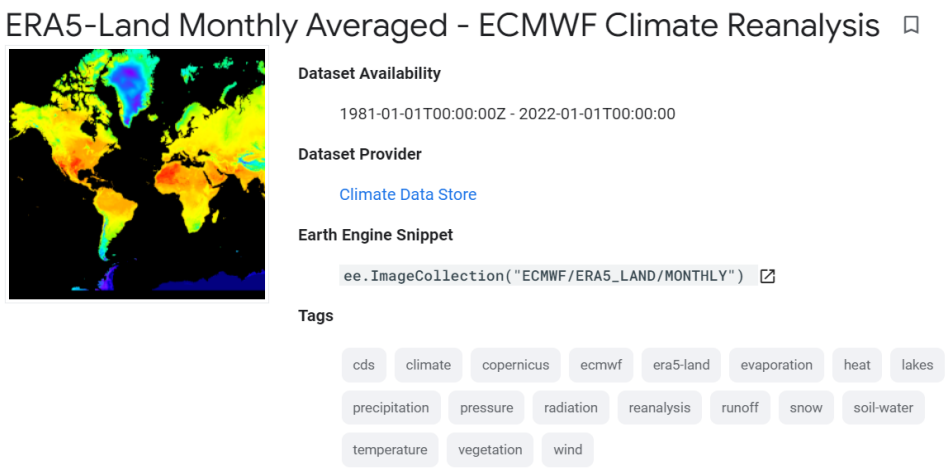


3.6.2 Irrigation System

The raw data for the climate and soil is found in the ERA5 dataset (Figure 42), which is a reanalysis dataset based on the satellite data. The time slot of the dataset is from 1981 to 2022, and we picked the analysis frequency as monthly averaged data. The data are included: evaporation data, precipitation data, pressure data, temperature data, soil water data, wind data, etc.

Figure 42

Description of ERA5 Dataset



For the pre-processed data, we downloaded the ERA5 data and saved it in CSV format displayed in figure 43. As the previous section mentioned, there are more than 70,000 polygons in San Joaquin County, meaning there should be more than 70,000 rows of data. But for climate data found in the ERA5 dataset, there are 3424 useful rows. So, we concatenate these two datasets by using the same key ‘index’ and ‘geo’. Finally, we matched 3424 polygons with effective climate data.

Figure 43

Pre-process Data Downloaded from the ERA5 Dataset

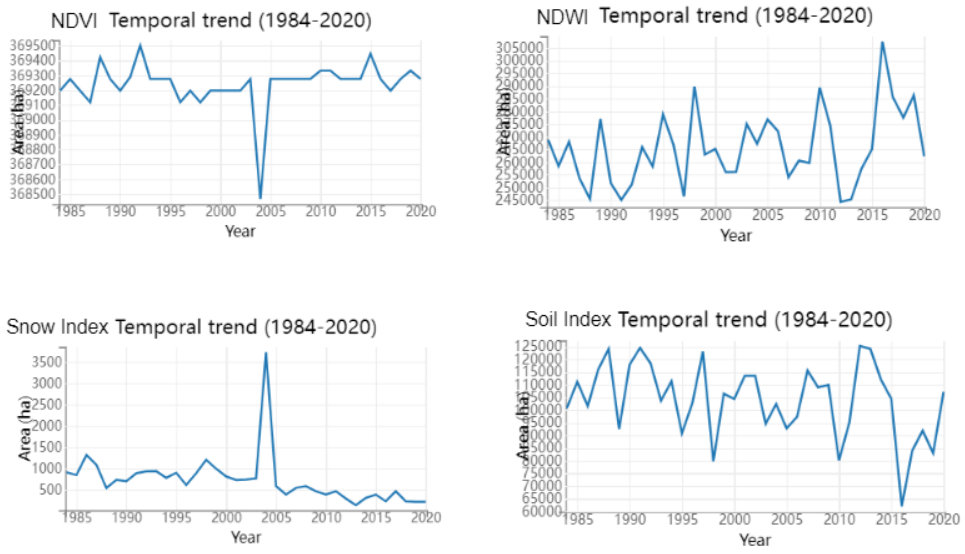
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	system:index	CLASS1	dewpoi	evapora	evapora	evapora	evapora	forecas	lake_bc	lake_ice	lake_ice	lake_m	lake_m	lake_sh	lake_to	leaf_ar	leaf_ar	potenti	runoff
7	9	38.028856	[-121.544	38.028863	[-121.544	38.028870	[-121.544	38.028874	[-121.544	38.028884	[-121.544	38.028891	[-121.544	38.028909	[-121.544	38.028923	[-121.544	38.028934	[-121.544
8	0000000000000000a7ab	F	281.4578	-0.00171	-3.09E-04	-2.64E-04	-1.77E-08	0.179144	282.8601	-273.15	-2.35E-04	16.42773	284.1809	-272.443	10.56802	1.914551	2.782104	-0.00502	2.32E-04
14	000000000000000005cdf	G	281.3797	-0.00193	-3.03E-04	-2.81E-04	-2.71E-07	0.188131	283.7864	-273.15	-2.35E-04	4.857422	285.6116	-272.467	12.02896	3.294922	2.01062	-0.00693	1.91E-04
15	5.00E+17	G	281.3797	-0.00193	-3.03E-04	-2.81E-04	-2.71E-07	0.188131	283.7864	-273.15	-2.35E-04	4.857422	285.6116	-272.467	12.02896	3.294922	2.01062	-0.00693	1.91E-04
16	000000000000000053a6	G	281.2371	-0.0018	-3.81E-04	-2.76E-04	-2.99E-05	0.184545	283.3269	-273.15	-2.35E-04	5.491211	285.4338	-272.455	11.73989	3.681885	1.775024	-0.00651	4.12E-04
19	5899	G	281.2371	-0.0018	-3.81E-04	-2.76E-04	-2.99E-05	0.184545	283.3269	-273.15	-2.35E-04	5.491211	285.4338	-272.455	11.73989	3.681885	1.775024	-0.00651	4.12E-04
21	0000000000000006c4c	G	281.325	-9.91E-04	-0.00117	-1.73E-04	-3.95E-05	0.171941	283.7917	-273.15	-2.35E-04	3.698242	286.3323	-272.458	12.36587	2.445801	1.223999	-0.00397	0.001332
22	0000000000000006c51	G	281.325	-9.91E-04	-0.00117	-1.73E-04	-3.95E-05	0.171941	283.7917	-273.15	-2.35E-04	3.698242	286.3323	-272.458	12.36587	2.445801	1.223999	-0.00397	0.001332
34	0000000000000000bf82	G	281.3425	-0.00193	-3.24E-04	-2.57E-04	-2.07E-05	0.174749	283.4031	-273.15	-2.35E-04	5.838867	285.593	-272.447	11.80825	2.265869	2.145508	-0.00462	3.59E-04
35	0000000000000000bf85	G	281.3425	-0.00193	-3.24E-04	-2.57E-04	-2.07E-05	0.174749	283.4031	-273.15	-2.35E-04	5.838867	285.593	-272.447	11.80825	2.265869	2.145508	-0.00462	3.59E-04
36	0000000000000000bf86	G	281.3425	-0.00193	-3.24E-04	-2.57E-04	-2.07E-05	0.174749	283.4031	-273.15	-2.35E-04	5.838867	285.593	-272.447	11.80825	2.265869	2.145508	-0.00462	3.59E-04
38	0000000000000000bf89	G	281.3425	-0.00193	-3.24E-04	-2.57E-04	-2.07E-05	0.174749	283.4031	-273.15	-2.35E-04	5.838867	285.593	-272.447	11.80825	2.265869	2.145508	-0.00462	3.59E-04
39	4343	G	281.0672	-0.00179	-3.53E-04	-2.72E-04	0	0.188375	285.0491	-273.15	-2.35E-04	3.805664	285.8342	-272.467	12.29751	2.033936	2.263916	-0.00453	5.05E-05
44	0000000000000000c10f	G	281.3425	-0.00193	-3.24E-04	-2.57E-04	-2.07E-05	0.174749	283.4031	-273.15	-2.35E-04	5.838867	285.593	-272.447	11.80825	2.265869	2.145508	-0.00462	3.59E-04
48	0000000000000005ca1	G	281.3797	-0.00193	-3.03E-04	-2.81E-04	-2.71E-07	0.188131	283.7864	-273.15	-2.35E-04	4.857422	285.6116	-272.467	12.02896	3.294922	2.01062	-0.00693	1.91E-04
56	0000000000000000065da	G	281.4187	-0.00189	-2.58E-04	-2.46E-04	-7.51E-07	0.185201	284.3123	-273.15	-2.35E-04	4.654297	285.7883	-272.478	12.28188	1.88916	2.45874	-0.00485	8.46E-05
59	6.00E+67	G	281.3425	-0.00193	-3.24E-04	-2.57E-04	-2.07E-05	0.174749	283.4031	-273.15	-2.35E-04	5.838867	285.593	-272.447	11.80825	2.265869	2.145508	-0.00462	3.59E-04
63	0000000000000000004a76	G	281.0672	-0.00179	-3.53E-04	-2.72E-04	0	0.188375	285.0491	-273.15	-2.35E-04	3.805664	285.8342	-272.467	12.29751	2.033936	2.263916	-0.00453	5.05E-05
67	0000000000000000004ff8	G	281.3797	-0.00193	-3.03E-04	-2.81E-04	-2.71E-07	0.188131	283.7864	-273.15	-2.35E-04	4.857422	285.6116	-272.467	12.02896	3.294922	2.01062	-0.00693	1.91E-04
76	0000000000000000588d	G	281.2371	-0.0018	-3.81E-04	-2.76E-04	-2.99E-05	0.184545	283.3269	-273.15	-2.35E-04	5.491211	285.4338	-272.455	11.73989	3.681885	1.775024	-0.00651	4.12E-04
79	3704	G	281.2273	-0.00183	-4.07E-04	-2.59E-04	0	0.193456	285.2517	-273.15	-2.35E-04	3.848633	286.0969	-272.467	12.55728	2.187988	2.14563	-0.00445	5.44E-05

NDVI index and snow index have fluctuated steadily for the past 40 years, except for

2005 with a great change. The NDWI index has an upward trend with more volatility, while the soil index has a downward trend (Figure 44). As the huge change of NDVI and snow index in 2005, we need further analysis to find out the cause.

Figure 44

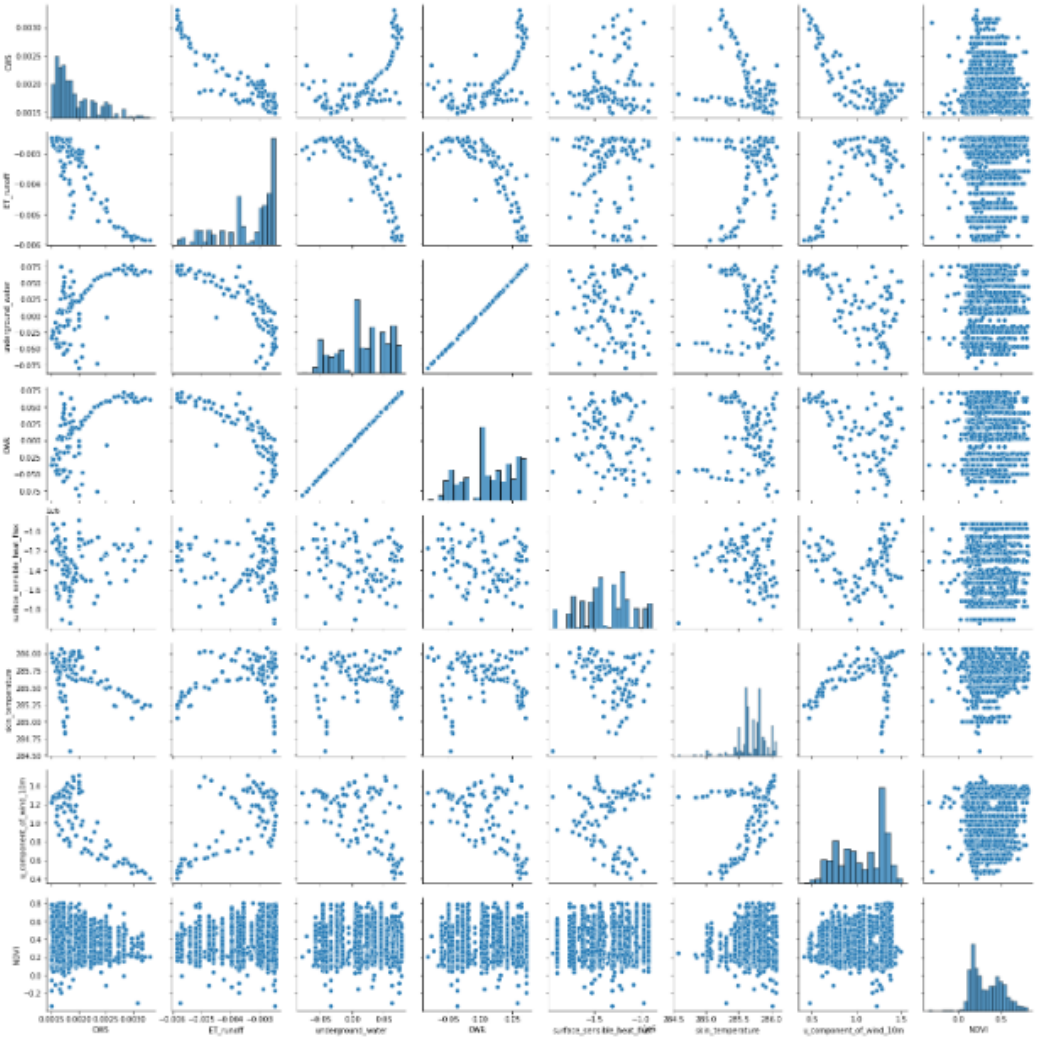
Time Series of Remote Sensor Index



In the prepared irrigation dataset, we select 8 input variables: CWS, ET_runoff, underground water, CWR, heat flux, skin temperature, wind, and NDVI. It is noted from Figure 45 that CWS has a positive relationship with underground water and CWR, and a negative relationship with ET_runoff and skin temperature.

Figure 45

Correlation Among Input Features of Irrigation Dataset



3.6.2 Fertilizer Recommendation System

The data utilized in the fertilizer recommendation system drone sensor data that includes features such as Red (DN), NIR (DN), Green (DN), NDVI, RH, Temp and the Label Nitrogen (PPM). The data description/ statistics is shown in the Figure 46 below.

Figure 46

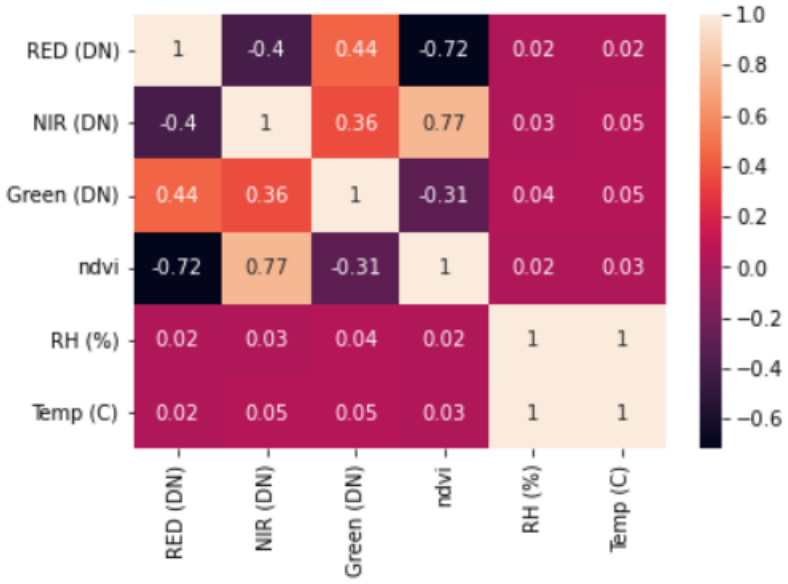
Processed data for the Fertilizer Recommendation System

	RED (DN)	NIR (DN)	Green (DN)	ndvi	RH (%)	Temp (C)	Nitrogen (ppm)
count	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000	54.000000
mean	63.206481	293.290926	56.277037	0.631111	45.866667	24.033333	748.785926
std	19.950117	50.433379	6.264775	0.126695	9.810430	0.661045	174.655508
min	39.490000	200.320000	43.110000	0.410000	33.800000	23.200000	451.550000
25%	48.112500	256.822500	52.332500	0.492500	33.800000	23.200000	603.790000
50%	56.385000	297.755000	56.315000	0.675000	46.200000	24.100000	707.040000
75%	75.802500	325.945000	61.655000	0.740000	57.600000	24.800000	921.347500
max	104.660000	400.780000	67.010000	0.790000	57.600000	24.800000	1086.190000

We further sought to see the correlation between data set features. The correlation between features is shown in the Figure 47.

Figure 47

Correlation Among Input Features of Fertilizer Recommendation (N) Dataset



Correlation can take a value between 1 and –1. A value close to 1 shows a positive correlation while –1 shows a negative correlation. Here as per the correlation matrix we see a high negative correlation between Red and NIR while a high positive correlation is seen between

NIR and NDVI. Other correlations between features are close to 0 and shows no correlation at all.

3.7 Data Analytics Results

Using diverse big data visualization formats is crucial and efficient to analyze and make decisions. It helps us recognize patterns in the dataset quickly and get better insights into breakdown data. Below are some of the visualization techniques we used in the project: graphs, histograms, fever charts, heatmap, map-based analytics, image analytics……

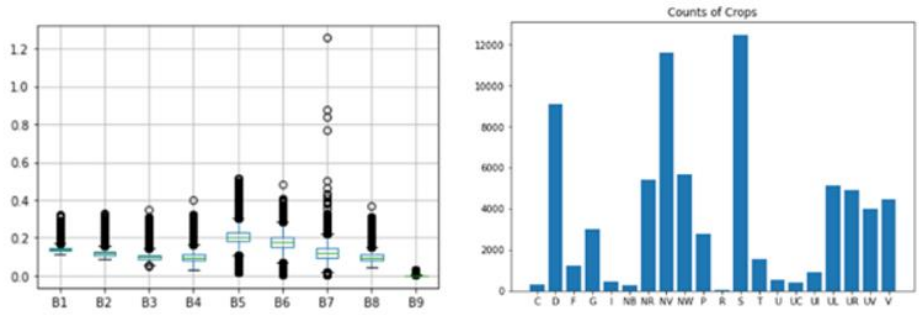
Histogram: We use histogram to compare similarities and organize data in a more interpretable way. It is similar to bar graph; both are able to present the features and amounts of numbers.

Heatmap: It is a graph using numerical data highlighted in light or dark colors to present the correlation among data. We use heatmap to measure the correlation among different vegetation indices. Correlation can take a value between 1 and –1. A value close to 1 shows a positive correlation while –1 shows a negative correlation.

Map-based analytics images: a map can easily show the features in the coordinate system. The farmland information from our project has to be related to map-based analytics, as the longitude and latitude can decide the crop type and surrounding environment.

Figure 48

Example for big data visualization format: boxplot and bar chart



Chapter 4 Model Development

4.1 Model Proposals

The main purpose of the project work is to identify crop types in farmland and then analyze and predict the irrigation and fertilizer circumstance in that area. To achieve the purpose of crop identification, classification machine learning models are used to group the features for different crop categories. Linear machine learning models are deployed in irrigation cycle recommendation system to get the demand and supply of usage water. Deep learning models are designed to predict the fertilizer usage for the third part, which allows to determine the performance of the former two models. Subsections explain the details of models applied in the following three systems:

- Crop identification
- Irrigation cycle
- Fertilizer recommendation

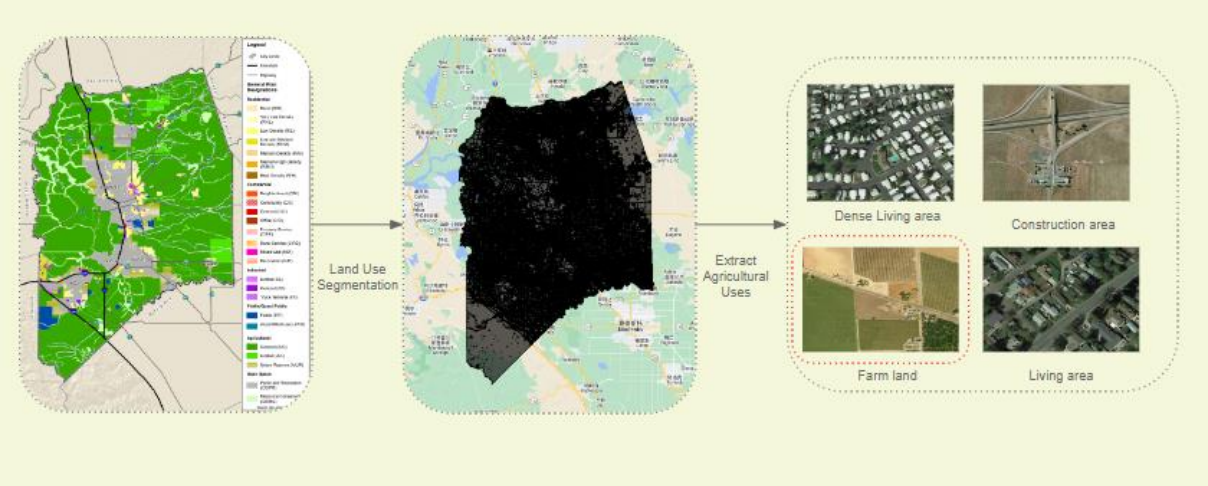
4.1.1 Crop Identification Model

Our crop identification model is formed with two stages, tone is the land cover segmentation, another is the crop identification. Before feeding spectral data into the machine learning models, we classify the land cover categories based on polygon boundary. The boundary delineation is the vectorization process of the spectral raster. The spectral data is used as the pixel attribute and converted into vectors. Each small polygon has similar vectors grouped into one polygon cluster. Any small polygons with less than pixel data are merged with neighboring polygon. Then the polygon layer is imposing to the crop classification layer imported from Google Earth Engine. For example, at the San Joaquin area that we worked on, first we found the land cover labels, and then combined our boundary layer which separate the

whole area into more than 4000 polygons with the land cover labels. After the calculation of vegetation index, the polygons representing farmland area can be selected for the further analysis. The second step is to deploy machine learning models to classify the crop types in each polygon area that we delineated. There are four machine learning models have been used in crop classification system: random forest tree, support vector machine, logistic model, and ensemble voting model based on forest tree (Figure 49).

Figure 49

Crop Classification Model

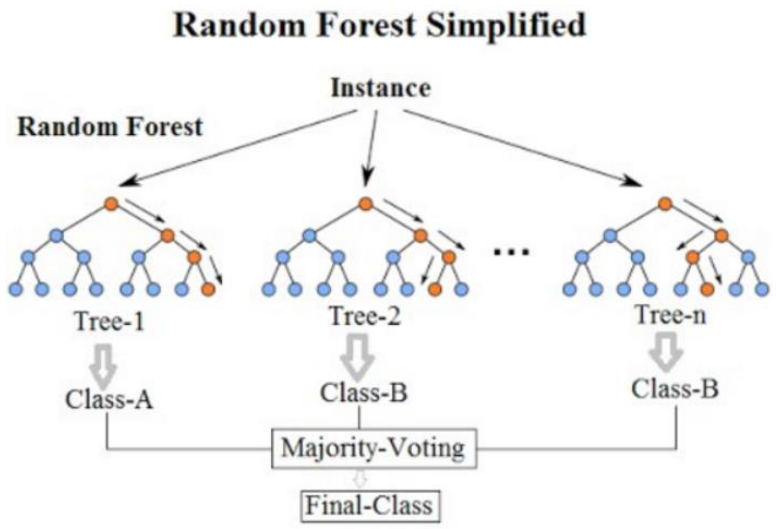


1. Random Forest Tree

Random forest tree is an ensemble method to classify individual sub-trees in the whole group (Figure 50) It is mostly used for classification and regression problems. Each sub-tree in the forest splits out multiple classes and the most votes classification becomes this subtree result.

Figure 50

Structure of Random Forest Tree

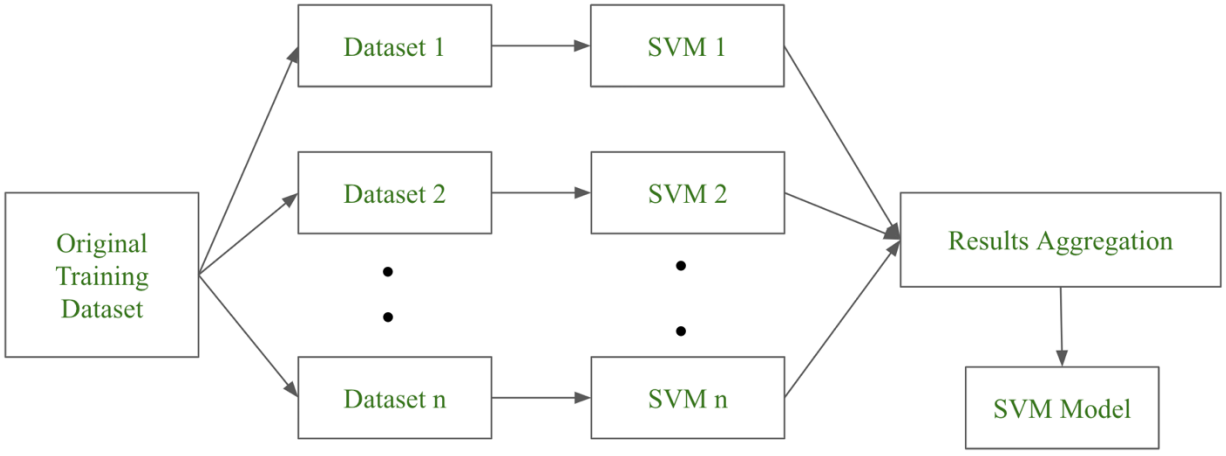


2. Support Vector Machine

Support vector machine is focused to find a hyperplane to maximize the distances between data points from both sides. By using the SVM model shown in figure 51, we are able to classify the new data with more accuracy. The dimensions of the hyperplane depend on the number of features, for example, the hyperplane is a line with two input features, but a two-dimensional plane with three input features.

Figure 51

Structure of SVM Model



The loss function helps to maximize the distance between two classes.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \qquad c(x, y, f(x)) = (1 - y * f(x))_+$$

Hinge loss function (function on left can be represented as a function on the right)

3. Innovative Model: Ensemble Voting Model

We combined four machine learning models to improve the accuracy of the model. The voting ensemble model can be separated into a hard voting classifier and a soft voting classifier. For the hard voting classifier, each individual machine learning model makes its own prediction, and then the ensemble model picks the result with the majority of votes. For the soft voting classifier, the ensemble model makes the prediction based on the average of each component model.

Table 17

Example of Hard Voting Classifier

Observation	Model A Prediction	Model B Prediction	Model C Prediction	Sum of "1" ClassVotes	Hard Voting Classifier Prediction
1	0	1	0	1	0
2	1	1	1	3	1
3	0	0	1	1	0
4	1	1	0	2	1
5	0	0	0	0	0

Table 18

Example of Soft Voting Classifier

Observation	Model A Prediction	Model B Prediction	Model C Prediction	Sum of "1" ClassVotes	Soft Voting Classifier Prediction
1	.44	.75	.48	.56	1
2	.65	.72	.54	.64	1
3	.26	.41	.64	.44	0
4	.54	.61	.38	.51	1

5	.41	.32	.42	.38	0

We pick the Logistic regression, support vector machine, Naive Bayes, Random Forest as the individual machine learning models, presenting as ‘clf1’, ‘clf2’, ‘clf3’, ‘clf4’. The limit set for the SVM is C equal to 1 and the kernel is linear. An ensemble classifier in the ‘mlxtend’ package is used to ensemble these four models. With the cross-validation frequency of 5 times, we are able to assess the ensemble model score mean and standard deviation.

Figure 52

Code for Soft Voting Classifier with 3 Individual Machine Learning Model

```
from mlxtend.classifier import EnsembleVoteClassifier

eclf = EnsembleVoteClassifier(clfs=[clf1, clf2, clf4], voting='soft', weights=[1,1,15])

labels = ['Logistic Regression', 'Support Vector Machine', 'Random Forest', 'Ensemble']
for clf, label in zip([clf1, clf2, clf4, eclf], labels):

    scores = model_selection.cross_val_score(clf, x_train,y_train,
                                             cv=5,
                                             scoring='accuracy')
    print("Accuracy: %0.2f (+/- %0.2f) [%s]"
          % (scores.mean(), scores.std(), label))
```

4.1.2 Irrigation Cycle Model

Using data-driven to analyze soil will help people who are interested in agriculture in general and especially farmers improve their efficiency in agriculture. The irrigated cycle is one of the most popular methods to reduce water runoff and harvested water. In general, the irrigation cycle is caused by surface water runoff that exceeds the soil infiltration rate. There are four types of irrigation: surface, sprinkler, drip/trickle, and subsurface (“Northeast Region Certified crop adviser”, n.d.). Figure 53 will show how the irrigation cycle works with the fertilizer cycle and the crop cycle. The root of vegetation extracts water from infiltration, and the left water in the rootzone will leach to the water table. We can see that for a plant to grow, it depends on many factors to analyze the soil condition in order to choose the corresponding crop.

As a result, we want to propose two main facts which are: crop water requirements (CWR) and crop water supply (CWS).

In our research, we focus on the irrigation cycle for the surface. Based on the survey, we can use the features such as snow evaporation, total evaporation, total precipitation, runoff, volumetric soil water layer, evaporation from bare soil, evaporation from open water surfaces, and evaporation from vegetation transpiration. These features can help us determine whether that soil part has enough water from the fourth layer to go through the first layer. Using crop water requirements (CWR) to compare to the crop water supply (CWS) which CWS has irrigation and rainfall (Abuzar et.al, 2013). CWS is the irrigation system that supplies water to the farm/field level, and it estimates the rainfall. To calculate CWR, we use:

$$CWR = Evaporation - Runoff + Underground\ Water$$

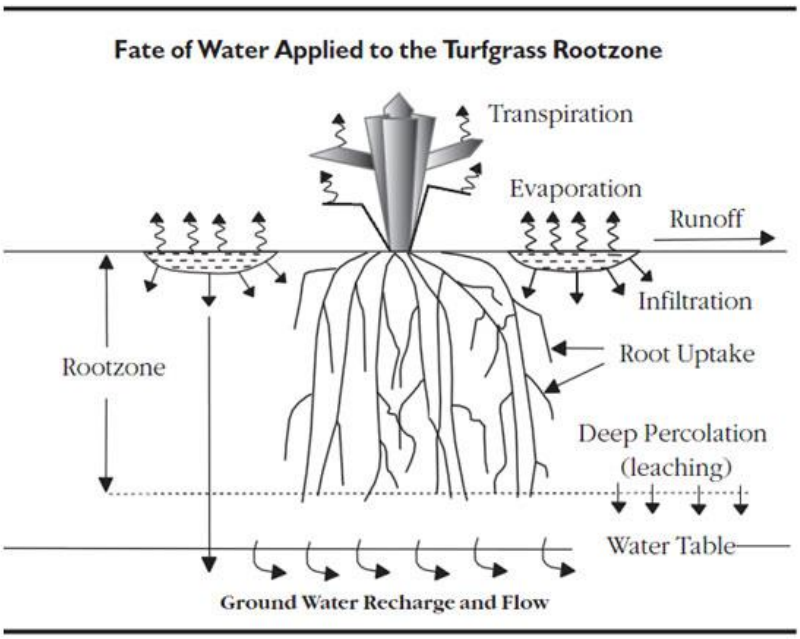
$$= KeEvaporation - Kr \cdot Runoff + KsSoil\ water\ layer\ 4 - Soil\ water\ layer\ 1$$

where:

Ke is the evaporation coefficient, Kr is the runoff coefficient, Ks is the soil water layer coefficient

Figure 53

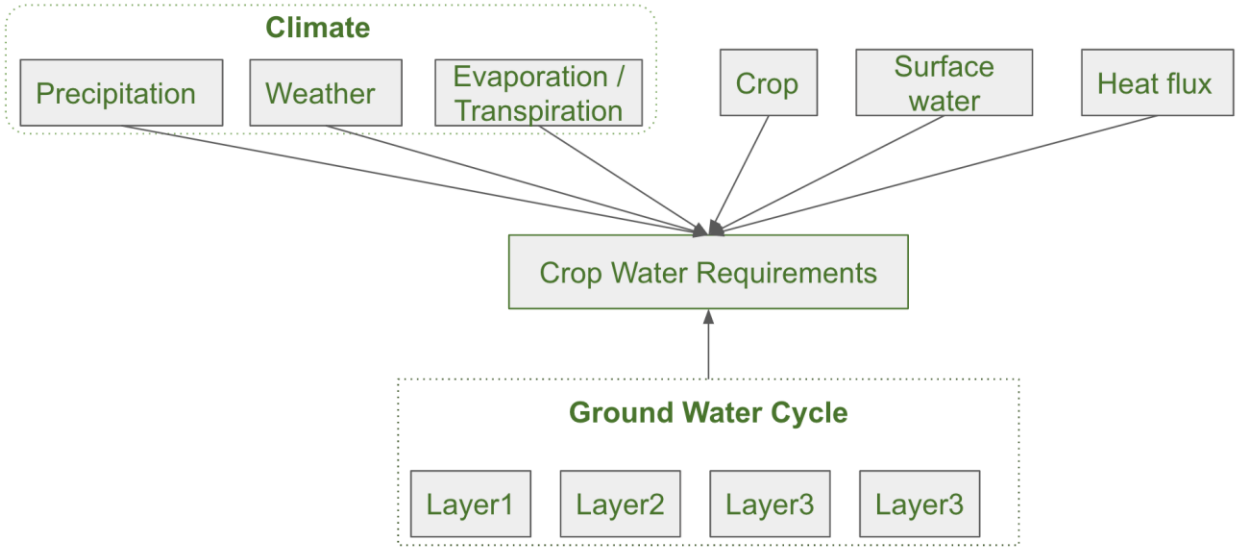
Fate of Water to Rootzone ("Maximizing Irrigation Efficiency and Water Conservation", 2016)



Irrigation Cycle dataset is formed in a CSV file format which provides the data of precipitation, water supply, water demand, evaporation from bare soil, vegetation transpiration, and volume of different soil layers which is a total of four layers. From the given dataset, we calculated the crop water requirement based on evaporation, runoff, and transpiration. The results came out with negative and positive numbers. In addition, we also calculate groundwater storage to estimate the amount of water in the deep ground. We calculate the underground water level based on soil water in different layers. The negative results show that water support from layer four which is the last layer is lacking and needs to take the water from the first layer. Groundwater storage can be described as a storehouse of water which means if they give a positive number then the storage capacity is enough for the crop. Otherwise, if a negative number is given, the groundwater storage needs to precipitate more water. We want to propose an irrigation model using the CWR and CWS to see if the natural inflows are enough for the crop since CWR will propose the minimum water required to maintain the growth in every condition.

Figure 54

Irrigation Cycle Model



1. Original Linear Regression Model

A simple regression model is formed with one input (X) to estimate the output (Y). In the higher dimension linear regression model, there are multiple inputs (X) in the equation and the line is called a hyper-plane.

$$y = B_0 + B_1x + B_2x + B_3x$$

We put different combinations of inputs to evaluate the result of y. By using statistical approaches such as standard deviations, correlations, or ordinary least squares, we seek the optimized accuracy.

2. Linear Regression Model with Lasso Regularization

When the coefficient changes, the influence of the input variables changes. If the coefficient drops to zero, the input is in-relevant to the output. If the coefficient is a big number no matter in the positive or negative way, this input has a high relationship with the output. Lasso regression is to minimize the absolute sum of the coefficients, which is an effective way to reduce inputs(X).

3. Polynomial Regression Model

Polynomial regression model is a specific linear regression model by adding some polynomial terms to the linear regression. We convert the one-degree input variable into polynomial terms with multiple degrees.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

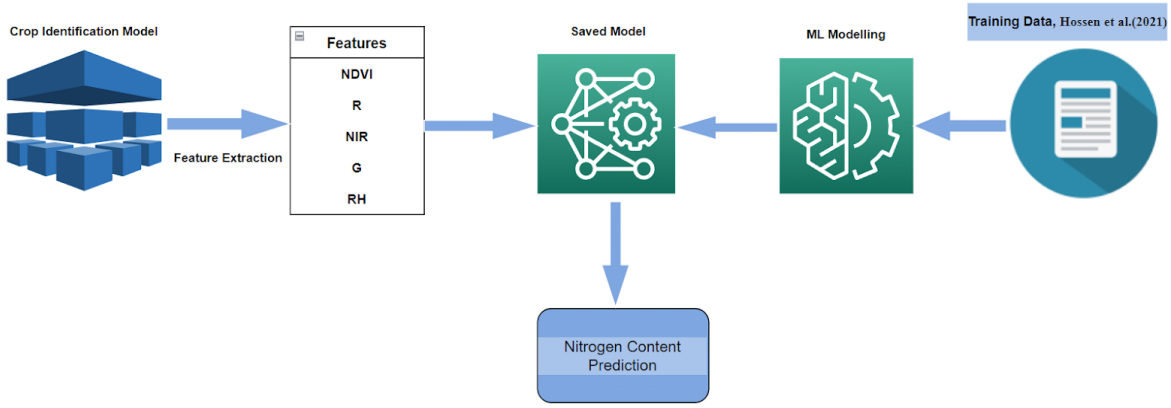
4.1.3 Fertilizer Recommendation Model

Nitrogen is one of the most important soil minerals vital for plant growth and development. Nitrogen deficiency in soil because of poor fertilizer can lead to poor crop productivity and vice versa, over usage of nitrate fertilizers can lead to nitrate toxicity in soil. As per the study conducted by Hossen et al. (2021), the nitrogen content of the soil can be estimated by using the Spectral sensor data (NIR, R, G band), along with LIBS data (Laser-Induced Breakdown Spectroscopy for Soil Measurements: Recent Progress and Potential, 2020) calibrated against the NIST LIBS database (NIST LIBS, 2022). They also published the training data with the label total nitrogen content in part per million (ppm) that can be used by similar related research works to estimate the Nitrogen content. The training features available in this model are red band Pixel Values (R), Near Infrared Band pixel value (NIR), Green Band Pixel Value (G), Normalized Differential Vegetation Index (NDVI), Relative Humidity (RH), and Air Temperature (C). The training data obtained thus will be used to build both regression and classification model to predict nitrogen content in the soil in part per million ranges (for regression model). We will also be grouping the continuous nitrogen data to discrete label, ie categorizing into bins, and this training data will be used with classification algorithm to predict the nitrogen content into labels (Low, Medium, High). The real-world data will come through the

'Crop Identification Model', which will be fed into a saved model to predict soil nitrogen content as shown in Figure 55.

Figure 55

Fertilizer (Nitrogen) Recommendation Model



1. Random Forest Classification

Random forest is a supervised method in machine learning which is used for both regression and classification problems. It uses the ensemble method for regression and classification.

2. Ada Boost Classification

Ada boost also name as Adaptive Boosting is another type of Ensemble model in machine learning. It makes use of multiple classifiers to increase the accuracy of classifiers.

3. Support Vector Machine Classification

Support vector machine is a powerful algorithm which is widely used in supervised machine learning use cases for both classification and Regression kind of problems. Support vector machine works by creating a hyperplane between data features and known to work well with data sets having less dimensionality and limited data sets.

4. MLP (Multilayer Perceptron) Classification

MLP is the simplest form of neural network where each layer has defined number if identical units. The first layer is called input layer, last layer is called output layer and in-between layer are called hidden layers. Each unit in the layer is fully connected with all the other units in preceding and following layers as in the Figure 56.

Figure 56

MLP Model

```
model=Sequential()
model.add(Dense(40,input_dim=5,activation='relu'))
model.add(Dense(40,activation='relu'))
model.add(Dense(40,activation='relu'))
model.add(Dropout(0.1))
model.add(Dense(5,activation='softmax'))
model.compile(loss='categorical_crossentropy',
              optimizer='adam',metrics=['accuracy'])

model.summary()
history = model.fit(X_train,y_train,
                    validation_data=(X_test,y_test)
                    ,batch_size=400,epochs=100,verbose=1)
```

4.2 Model Supports

Figure 57 illustrate three models involved in the report, the first is the crop identification, the second is the irrigation recommendation system, the third is the fertilizer recommendation system.

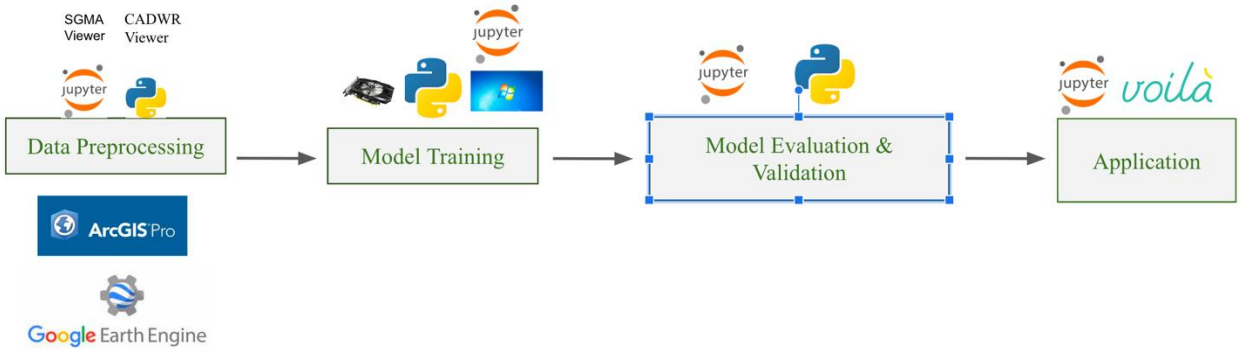
1. The platforms involved in the crop identification model:

- The user interface of vegetation analysis based on machine learning tools. The administrators are access to the interactive widgets and the back datasets. The user can only interact with the widgets in the web page.
- Voila from Jupyter notebook serves as the web stack hosting.
- Satellite datasets from Google Earth Engine are the dataset storage.

- CADWR Land Use Viewer - land use that allows access to statewide and existing DWR county land use survey datasets (Cadwr, n.d.)

Figure 57

Support Platforms for Crop Identification



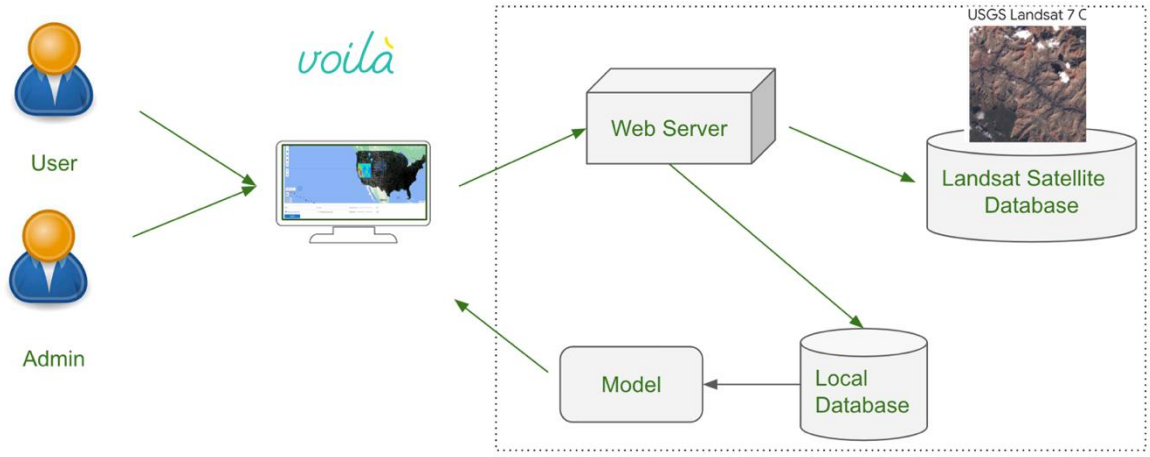
The platforms involved in the irrigation system recommendation:

- Google Earth Engine (GEE) - multi-petabyte satellite imaginary datasets and geospatial datasets with planetary-scale analysis capabilities. GEE lets users store satellite imagery to the cloud (Google, 2022).
- ArcGIS Pro - web-based mapping software, server software, and provide geographic information system (“About ArcGIS pro”, 2022).
- Python - is our main language to develop all of the models. Python provides libraries such as sklearn, logistic regression, geopandas, etc.
- Jupyter Notebook - a web-based interactive computing platform where we can share our live codes and modify them together.
- CADWR Land Use Viewer - land use that allows access to statewide and existing DWR county land use survey datasets (Cadwr, n.d.)

- SGMA Viewer - interactive tool shows groundwater level dataset of California area. The depth below the ground surface, groundwater elevation, and groundwater change in elevation.
- Voila - convert the Jupyter notebook into a standalone web application with a dashboard (“Using voila, n.d.).

Figure 58

Support Platforms for Irrigation System Recommendation

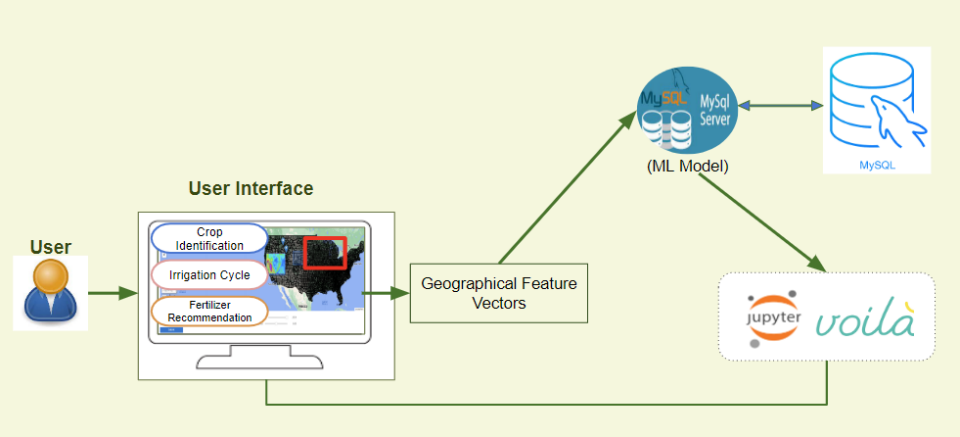


The platform involved in the fertilizer recommendation system:

- Python - is our main language to develop all the models. Python provides libraries such as sklearn, logistic regression, geopandas, etc.
- Jupyter Notebook - a web-based interactive computing platform where we can share our live codes and modify them together.

Figure 59

Support Platforms for Fertilizer System Recommendation



4.3 Model Comparison and Justification

The machine learning problem to solve at our end is a classification problem for the 'Crop Identification Model'. Given a set of input features from Landsat Sensor data, the task at hand is to identify a Crop based on input features. As the types of crops fall into various categories, it's a multi-class classification problem. The Fertilizer recommendation model is a multi-class classification problem, where given the set of features the task is to classify the Nitrogen content in the soil in defined ranges. The classification machine learning algorithms that are known to do well in multiclass classification are 1) Decision Tree, 2) Random Forest, 3) K Nearest Neighbor, 4) Logistic Regression, and 5) Naive Bayes Classifiers. For regression we have evaluated Random Forest Regression and Support Vector regression methods.

1. Decision Tree

A decision tree is a supervised learning method that is used to make decisions. A decision tree can be used for categorical variables and continuous variables without scaling the data ("*Decision tree in machine learning", 2021)*. A decision tree is the best method to use with the time-series data. One of the best characteristics of the decision tree is that it is very efficient and

fast and does not require data preprocessing. We are using a decision tree to identify the crop of the crop identification model (K,D, 2020).

2. KNN

KNN is a supervised method in machine learning where the algorithm decides on the classification based on the nearest data points (neighbors). The main advantage of using the KNN Classification method is that it works well in low dimension data space and this method does not require any training, and new data can be added at any time to the model seamlessly without impacting the accuracy of the algorithm. The problem associated with the KNN is that it is not suitable for large data sets and datasets with high dimensionality. As for the current use case of the 'Crop Identification Model' and 'Fertilizer Recommendation' model, the size of the data set could be high hence KNN application in this use case would have limitations in its application.

3. Naive Bayes Classification

Naive Bayes classification algorithms are known to do well in multi-class prediction use cases. For Naive Bayes classification to work flawlessly, input features have to be independent of each other. Another limitation in using Naive Bayes algorithm is that it works better with categorical input variables than numeric variables. In current use cases of Crop Recommendation and Fertilizer Recommendation mostly consist of numerical input variables.

4. Random Forest

Random Forest is one of the most widely used Classification algorithms which uses the Ensemble Learning Technique (Singh, 2020). One of the main advantages of Random Forest is that it creates many decision trees on the input subset data and combines all the outputs into a

final prediction. Random Forest methods are less prone to overfitting problems and tend to have better accuracy than other classification methods.

5. Multilayer perceptron (MLP)

Multilayer perceptron (MLP) is a kind of feed-forward neural network consisting of Input layer, output layer and Hidden Layer. Each layer fully connected with other layers. MLP are widely used to solve supervised machine learning problems. The weights and bias during the training can be adjusted to minimize the classification error.

Table 19

Crop Identification Results

Model		**Improved Accuracy of Selected Crop Type**	**Accuracy of Selected Crop Type**	**Accuracy of All Crop Type**
Logistic Regression		65.83%	72%	38%
Support Vector Machine	Kernel = linear	67.82%	71.87%	41.31%
	Kernel = rbf	72.23%	73.78%	45.31%
	Kernel = poly	71.35%	72.57%	43.73%
Random Forest		86.58%	78%	69%
Ensemble		86.6%	78%	52%

Table 20

Irrigation Cycle Results

Model	**RMSE**	**MSE**	**R2**	**Training Score**	**Testing Score**

Linear Regression		0.256	0.06554	0.939	0.93936	0.935395
Lasso Regularization	Alpha= 0.1	0.292	0.08535	0.917	0.92048	0.917162
	Alpha= 0.01	0.260	0.06759	0.934	0.93768	0.934405
	Alpha= 0.001	0.256	0.06554	0.936	0.939358	0.936394
Polynomial Regression		0.244	0.05983	0.939	0.939458	0.936400

Table 21

Fertilizer Recommendation Results

Model	Accuracy	Parameters
Random Forest	82%	estimators = 10
Ada Boost	82%	estimators = 10
Gaussian NB	64%	
Support Vector Machine	73%	
MLP (ReLU Activation)	93.3%	Trainable params: 1024, activation='relu', loss='categorical_crossentropy',optimizer='adam'

4.4 Model Evaluation Method

4.4.1 Crop Identification

The crop identification system is built to recognize the crop type in one specific farmland, accuracy, sensitivity, and specificity are effective indicators to compare the models' performances.

- Accuracy score meets the requirement to measure the classification performance and the ratio of predicted observations to the total observations. The total observation is the sum of the components in the confusion matrix (Kumar, 2022).

$$Accuracy = \frac{True\ Positive\ +\ True\ Negative}{True\ Positive + True\ Negative + False\ Positive +\ False\ Negative}$$

- Sensitivity is the true positive rate which shows the model is correctly predicted.

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}$$

- Specificity is the true negative rate which measures how many times does the classifier was precise.

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

- Feature Importance

4.4.2 Irrigation Recommendation

The irrigation recommendation system is deployed with linear regression models to quantify the relationship between supply water need and other variables. Two metrics we used to quantify how well our models fit the dataset are MSE & RMSE and R squared.

1. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

MSE can be used to measure how the data points get fitted to the line. MSE measures the amount of error for the models using the average squared between the real and predicted values (Frost, 2021).

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

y: i-th value, $\hat{y}$: predicted value, n: number of values

RMSE is used to measure the differences between the values using the standard deviation of the prediction errors. This measures the error while we are predicting the models.

$$RMSE=\frac{\sqrt{\sum_{i=1}^{N}(x_i-\widehat{x_i})^2}}{N}$$

Where:

ŷi: predicted value, yi: real value, n: numbers of values

2. R-squared

R squared is used to explain how well the independent variables explains the variability of dependent variables. The values range from 0 to 1, which is interpreted as the percentages. It is one way to measure the correlation while not the accuracy of model, so we used R squared approach with MSE to check whether high R squared makes sense.

4.4.3 Nitrogen Recommendation Model

In the nitrogen recommendation system, Multi-Layer Perceptron is used to predict the nitrogen amount. One effective way to evaluate the MLP model is F1 score and ROC curve.

1. F1 score

F1-score is calculated based on the sensitivity and specificity of the classifier. F1 score can show the performance of the two classifiers by comparing them. The results show which one has better performance (“What is F1-score?”, 2022)

$$F1=2\cdot\frac{\text{Sensitivity}*\text{Specificity}}{\text{Sensitivity}+\text{Specificity}}$$

2. ROC curve

ROC Curve shows the performance of the classification models using Sensitivity and Specificity at different thresholds using a graph. If the threshold is lowering meaning that it classifies more items as positive and if it is increasing meaning that true positives and false

positives rates are at one decision threshold. The Area Under ROC (AUC) shows the aggregate measures of all classification thresholds (“Classification”, n.d.)

3. Mean Squared Error (MSE)

MSE can be used to measure how the data points get fitted to the line. MSE measures the amount of error for the models using the average squared between the real and predicted values (Frost, 2021).

$$MSE=\frac{1}{N}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

4.5 Model Validation and Evaluation Results

4.5.1 Crop Identification System

The weighted average accuracy score is 77% for the crop identification model, with crop ’R’ as 100%, crop ‘D’ as 77%, crop ‘G’ as 78%, crop ‘P’ as 86%, crop ‘V’ as 79%. But for the crop ‘F’ and crop ‘C’, their precisions are just 57% and 50% that we need to take a further look at them. When the type of crops increased from 8 to 21, the accuracy went down to 70%.

Table 22

Evaluation Report for Crop Identification Models

RF with Selected Crop Types	**RF with all Crop Types**

	precision	recall	f1-score	support
C	0.50	0.19	0.27	59
D	0.77	0.90	0.83	1812
F	0.57	0.54	0.55	220
G	0.78	0.79	0.78	582
P	0.86	0.67	0.75	545
R	1.00	0.70	0.82	10
T	0.68	0.49	0.57	295
V	0.79	0.76	0.77	856
accuracy			0.77	4379
macro avg	0.74	0.63	0.67	4379
weighted avg	0.77	0.77	0.76	4379

	precision	recall	f1-score	support
C	0.92	1.00	0.96	500
D	0.55	0.60	0.57	522
F	0.87	0.94	0.90	502
G	0.78	0.76	0.77	524
I	0.83	1.00	0.91	494
NB	0.96	1.00	0.98	503
NR	0.42	0.41	0.41	490
NV	0.26	0.19	0.22	499
NW	0.44	0.28	0.34	512
P	0.77	0.73	0.75	491
R	1.00	1.00	1.00	455
S	0.25	0.25	0.25	495
T	0.85	0.90	0.88	461
U	0.92	1.00	0.96	460
UC	0.91	1.00	0.95	478
UI	0.88	0.97	0.92	506
UL	0.50	0.55	0.53	502
UR	0.42	0.51	0.46	506
UV	0.37	0.20	0.26	520
V	0.69	0.73	0.71	525
accuracy			0.70	9945
macro avg	0.68	0.70	0.69	9945
weighted avg	0.67	0.70	0.68	9945

As seen in Figure 60, the test accuracy fluctuated between Epoch 10 to 18 by reaching around 70% accuracy score, compared to train accuracy as 99%. The test accuracy is always lower than the train accuracy, and it keep flat after 15 depths, so there is not apparent overfitting in the testing data.

Figure 60

Training & Testing Accuracy for Random Forest Model with Different Depths

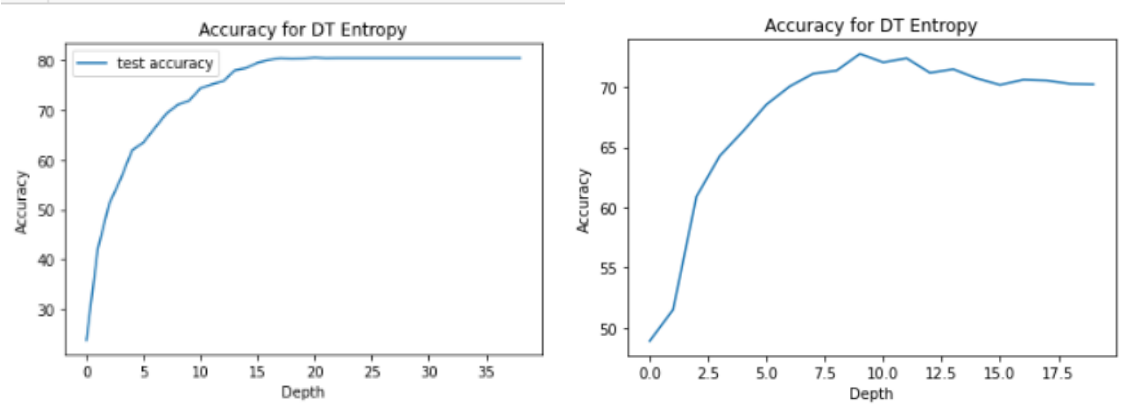


Table 23

Crop Identification Results

Model		Improved Accuracy of Selected Crop Type	Accuracy of Selected Crop Type	Accuracy of All Crop Type
Logistic Regression		65.83%	72%	38%
Support Vector Machine	Kernel = linear	67.82%	71.87%	41.31%
	Kernel = rbf	72.23%	73.78%	45.31%
	Kernel = poly	71.35%	72.57%	43.73%
Random Forest		86.58%	78%	69%
Ensemble		86.6%	78%	52%

4.5.2 Irrigation Recommendation System

We tuned the irrigation model by adjusting the hyperparameters. The optimized irrigation model has the accuracy at 76.8669, with the parameter as the number of trees in the forest is 200, the max number of features for splitting a node is one, the max number of levels is 10.

- n_estimators = number of trees in the forest = 200
- max_features = max number of features considered for splitting a node = [1,'sqrt','log2']
- max_depth = max number of levels in each decision tree = [None,5,10,15]
- min_samples_leaf = min number of data points allowed in a leaf node = [None,2,5,10]

Figure 61

Evaluation Report for Irrigation System Model

	feature_list	depth_list	accuracy_list
0	1	NaN	0.762503
1	1	NaN	0.440740
2	1	NaN	0.564284
3	1	NaN	0.644896
4	1	5.0	0.664535
5	1	5.0	0.440740
6	1	5.0	0.564284
7	1	5.0	0.641471
8	1	10.0	0.722083
9	1	10.0	0.440740
10	1	10.0	0.564284
11	1	10.0	0.644896
12	1	15.0	0.752227
13	1	15.0	0.440740
14	1	15.0	0.564284
15	1	15.0	0.644896
16	sqrt	NaN	0.766613
17	sqrt	NaN	0.493948
18	sqrt	NaN	0.603106
19	sqrt	NaN	0.647408
20	sqrt	5.0	0.676182
21	sqrt	5.0	0.493948
22	sqrt	5.0	0.603106
23	sqrt	5.0	0.647180
24	sqrt	10.0	0.738296
25	sqrt	10.0	0.493948
26	sqrt	10.0	0.603106
27	sqrt	10.0	0.647408
28	sqrt	15.0	0.768669
29	sqrt	15.0	0.493948
30	sqrt	15.0	0.603106
31	sqrt	15.0	0.647408
32	log2	NaN	0.766613
33	log2	NaN	0.493948
34	log2	NaN	0.603106
35	log2	NaN	0.647408
36	log2	5.0	0.676182
37	log2	5.0	0.493948
38	log2	5.0	0.603106
39	log2	5.0	0.647180
40	log2	10.0	0.738296
41	log2	10.0	0.493948
42	log2	10.0	0.603106
43	log2	10.0	0.647408
44	log2	15.0	0.768669
45	log2	15.0	0.493948
46	log2	15.0	0.603106
47	log2	15.0	0.647408

The R squared of CWS model is 0.918. The coefficient of ‘ET_runoff’ is –0.6817, followed by the coefficient of ‘wind_10m’ as –0.4137. It means that 68% of the variable ‘CWS’ can be explained by independent variable ‘ET_runoff’ and 41% can be explained by independent variable ‘wind_10’ (Figure 62).

Figure 62

Summary for Crop Identification Linear Model

Dep. Variable:	CWS	R-squared:	0.918
Model:	OLS	Adj. R-squared:	0.918
Method:	Least Squares	F-statistic:	4547.
Date:	Fri, 22 Apr 2022	Prob (F-statistic):	0.00
Time:	11:50:06	Log-Likelihood:	-341.14
No. Observations:	2030	AIC:	694.3
Df Residuals:	2024	BIC:	728.0
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0014	0.006	0.226	0.821	-0.011	0.014
ET_runoff	-0.6817	0.013	-52.601	0.000	-0.707	-0.656
underground_water	-0.0444	0.007	-6.461	0.000	-0.058	-0.031
CWR	-0.0624	0.007	-8.567	0.000	-0.077	-0.048
surface_sensible_heat_flux	-0.0061	0.008	-0.804	0.422	-0.021	0.009
skin_temperature	-0.1986	0.008	-24.835	0.000	-0.214	-0.183
u_component_of_wind_10m	-0.4137	0.009	-48.667	0.000	-0.430	-0.397

Omnibus:	930.796	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6676.407
Skew:	2.024	Prob(JB):	0.00
Kurtosis:	10.908	Cond. No.	2.02e+16

Table 24

Combined Evaluation Report for Irrigation Cycle

Model		RMSE	MSE	R2	Training Score	Testing Score
Linear Regression		0.256	0.06554	0.939	0.93936	0.935395
Lasso Regularization	Alpha= 0.1	0.292	0.08535	0.917	0.92048	0.917162
	Alpha= 0.01	0.260	0.06759	0.934	0.93768	0.934405
	Alpha= 0.001	0.256	0.06554	0.936	0.939358	0.936394
Polynomial Regression		0.244	0.05983	0.939	0.939458	0.936400

4.5.3 Fertilizer Recommendation System

We trained the fertilizer recommendation model on multiple knows classification algorithm models including Random Forest, Ada Boost, Gaussian NB, Logistic Regression, Support Vector Machine and MLP (Multi-Layer Perceptron) models. The accuracy of each model is shown in Table 21, Figure 63. Random Forest, Ada Boost, Logistic Regression classifiers were able to achieve validation accuracy of 82%, while the accuracy for Support Vector Machine, Gaussian NB and MLP Deep Learning model came to about 73%, 64% and 93.3% respectively.

Figure 63

Evaluation Report for Fertilizer – Nitrogen Recommendation Model (Random Forest, Ada Boost, SVM)

```
---------------------------
Train accuracy: 0.98
Test accuracy : 0.82
---------------------------
Random Forest Classification Report
              precision    recall  f1-score   support

           1       1.00      1.00      1.00         1
           2       0.80      1.00      0.89         8
           3       0.00      0.00      0.00         2

    accuracy                           0.82        11
   macro avg       0.60      0.67      0.63        11
weighted avg       0.67      0.82      0.74        11

---------------------------
Random Forest Model Accuracy 0.8181818181818182
---------------------------
```

```
---------------------------
Train accuracy: 0.74
Test accuracy : 0.82
---------------------------
AdaBoostCLF Classification Report
              precision    recall  f1-score   support

           1       1.00      1.00      1.00         1
           2       0.80      1.00      0.89         8
           3       0.00      0.00      0.00         2

    accuracy                           0.82        11
   macro avg       0.60      0.67      0.63        11
weighted avg       0.67      0.82      0.74        11

---------------------------
AdaBoostCLF Model Accuracy 0.8181818181818182
---------------------------
```

```
---------------------------
Train accuracy: 0.81
Test accuracy : 0.73
---------------------------
SVM Classification Report
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         1
           2       0.73      1.00      0.84         8
           3       0.00      0.00      0.00         2

    accuracy                           0.73        11
   macro avg       0.24      0.33      0.28        11
weighted avg       0.53      0.73      0.61        11

---------------------------
SVM Model Accuracy 0.7272727272727273
---------------------------
```

Figure 64

Evaluation Report for Fertilizer – Nitrogen Recommendation Model (MLP)

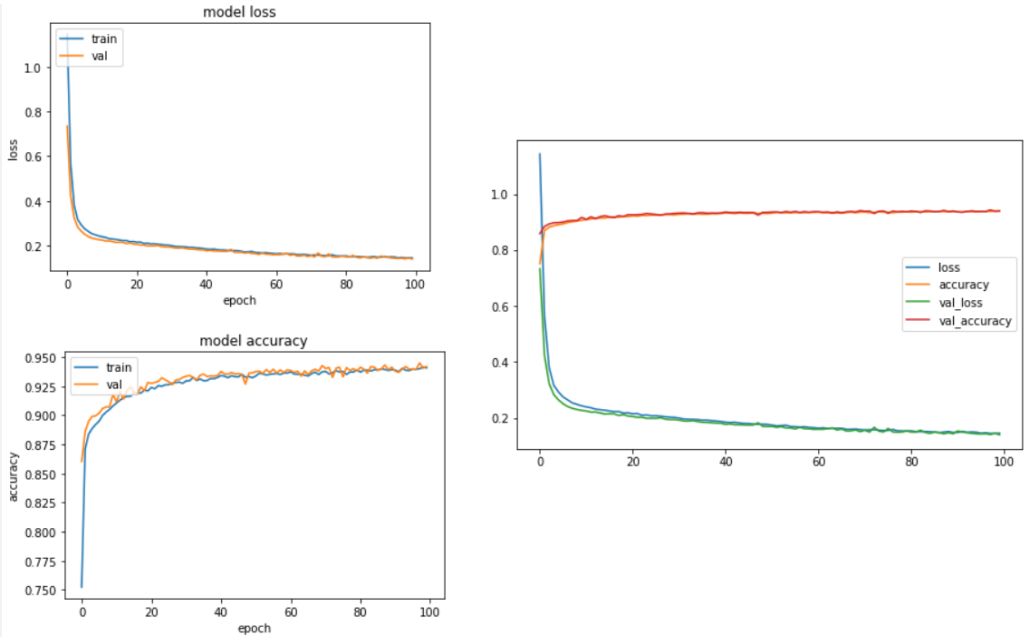


Table 25

Combined Evaluation Report for Fertilizer – Nitrogen Recommendation Model

Model	Accuracy	Parameters
Random Forest	82%	estimators = 10
Ada Boost	82%	estimators = 10
Support Vector Machine	73%	
MLP (ReLU Activation)	93.3%	Trainable params: 1024, activation='relu', loss='categorical_crossentropy',optimizer='adam'

4.5.4 Solution to Target Problem based on Validated Results

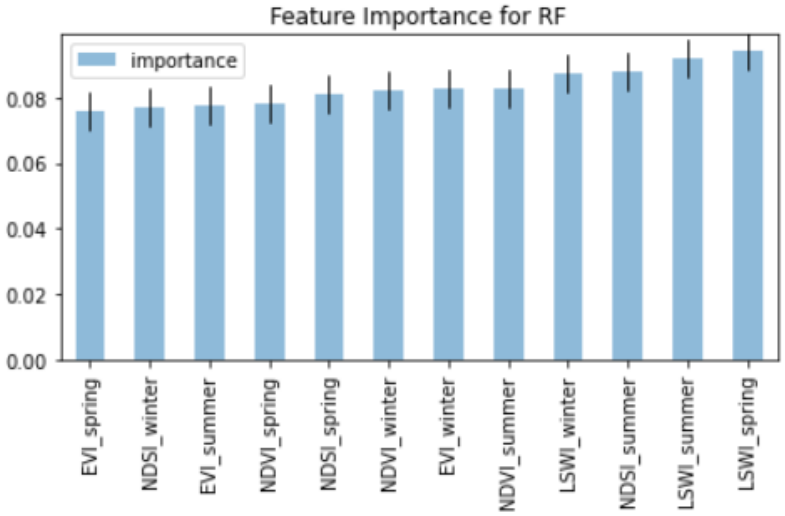
1. Crop Identification System Improvement

Apart from knowing the what the classification is, we also wonder which features are most valuable to determine the forecast. In the Random Forest model, we can see the

LSWI_spring is the most important feature followed by LSWI_summer and NDSI_summer. We will put more weights on the top 5 features.

Figure 65

Feature Importance for Random Forest



2. Irrigation Recommendation System

In Table 26 of irrigation model evaluation, the testing score decreased as the alpha value getting higher from 0.001 to 0.1. The regularization function from Lasso balance off the overfitting and training cost. From the figure 66 of coefficient magnitude, it shows that the coefficient magnitude fluctuated the most when alpha equals to 0.01, followed by alpha equals to 0.00001.

Table 26

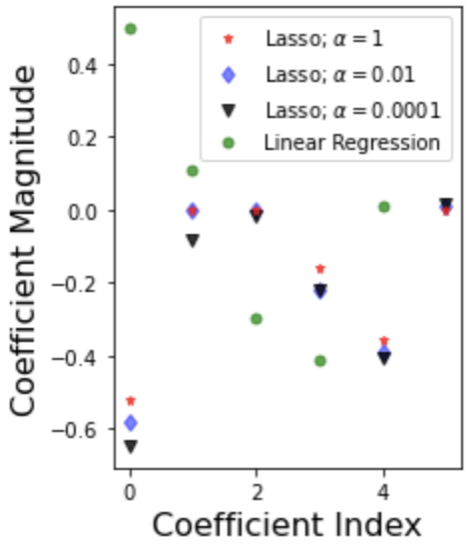
Irrigation Cycle Model Evaluation

	MSE	**Training Score**	**Testing Score**
Linear Regression	0.06554	0.93936	0.935395
Lasso			
• **Alpha = 0.1**	0.08535	0.92048	0.917162
• **Alpha = 0.01**	0.06759	0.93768	0.934405
• **Alpha = 0.001**	0.06554	0.939358	0.936394

Polynomial Regression	0.05983	0.939458	0.936400

Figure 66

Coefficient Magnitude for different Linear Regularizations



Chapter 5 Data Analytics System

5.1 System Requirements Analysis

5.1.1 System Boundary and Use Cases

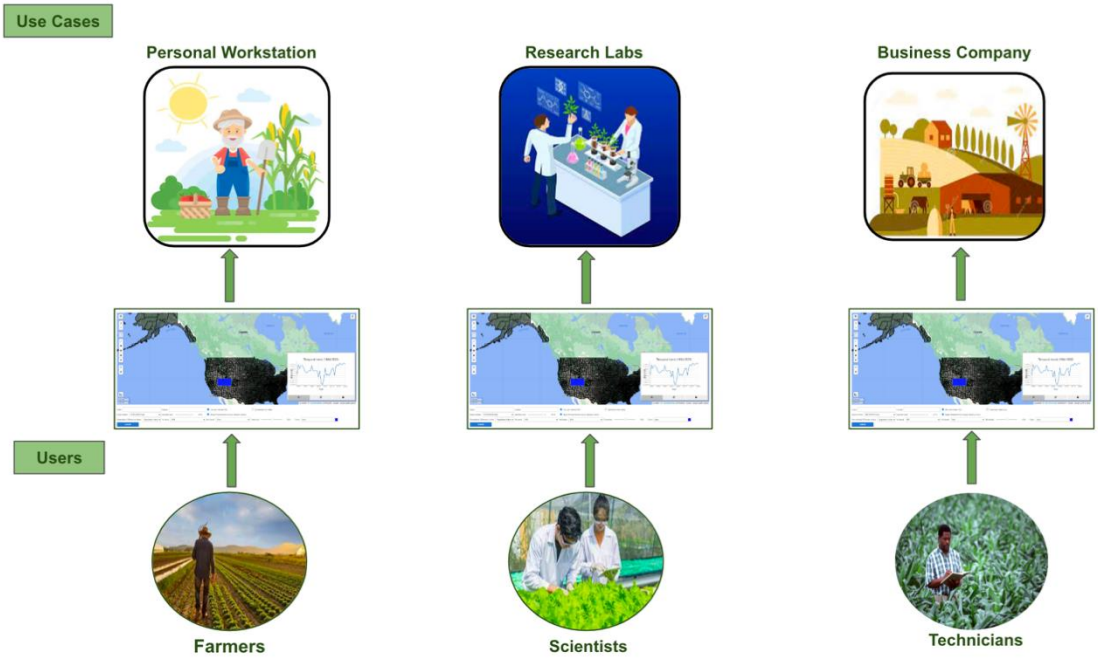
Our use cases mainly happen in three environments: personal workstation, research labs and agriculture business company. The corresponding users of our system are farmers who plant and study in their personal workstation, scientists of research labs and technicians from agriculture business company. Figure 67 illustrates three different kinds of scenarios of each user and the use case. The scenario can be summarized as follow:

A farm worker (e.g., farmer, scientist, or technician) performs the assessment of the interested farmland using this system before determining the cultivation plan. The system returns the analysis report listing irrigation cycle, crop identification and fertilizer recommendation

given the selected farmland. The system is a web application and would be run in the farm worker's workspace (e.g., personal workstation, research lab or inside agriculture business company). The detailed analysis of farmlands is essential and vital for every cultivation since different farmlands should be taken care differently, and it leads to a more productive, efficient, and ecological cultivation.

Figure 67

System Boundary and Use Cases



System High-Level Data Analytics and Machine Learning Requirements

This project will deliver a recommendation system and soil analytic tool to provide a detailed and organized information to the farmland users. There are three separate platforms in the system supporting crop identification, irrigation recommendation, and fertilizer application. The datasets come from the satellite data of Google Earth Engine, soil properties data of electromagnetic sensors, thermal data from thermal cameras.

The high-level data analytics of the project is made up of three tiers: reporting, insights,

and prediction. In the reporting level, we construct a crop identification platform, which able to recognize different types of crops based on the sensors' bands. Each crop has its special vegetation indices features, we trained the machine learning models to learn the general disciplines among diversified crops and try to tell the type of crop with an accuracy around 93%.

In the insights level, we created our own formula to analyze irrigation situations based on the on the surface water and underground water data. The calculation of surface water is dealing with the precipitation, irrigation, and temperature of the land. The calculation of the underground water includes the different layers of water level (four layers).

In the prediction level, we forecast the fertilizer usage of one specific farmland for the next season combined with the results from the first crop identification part and the second irrigation recommendation part. Additionally, the nutrition data about the soil (NPK), collected from the sensors, is critical to the prediction process. With the respect to the use of the data, deep learning models will be deployed to recommend fertilizer usage.

The machine learning requirements of the project include the linear regression model, classification machine learning models, and deep learning models. The inputs are numerical data, some of which are transformed from categorical data and text content. Normalization is used in order to minimize the effect of noise data. The hyperparameters are tuned based on the evaluation process for each model. Three hidden layers are constructed in our deep learning model.

5.2 System Design

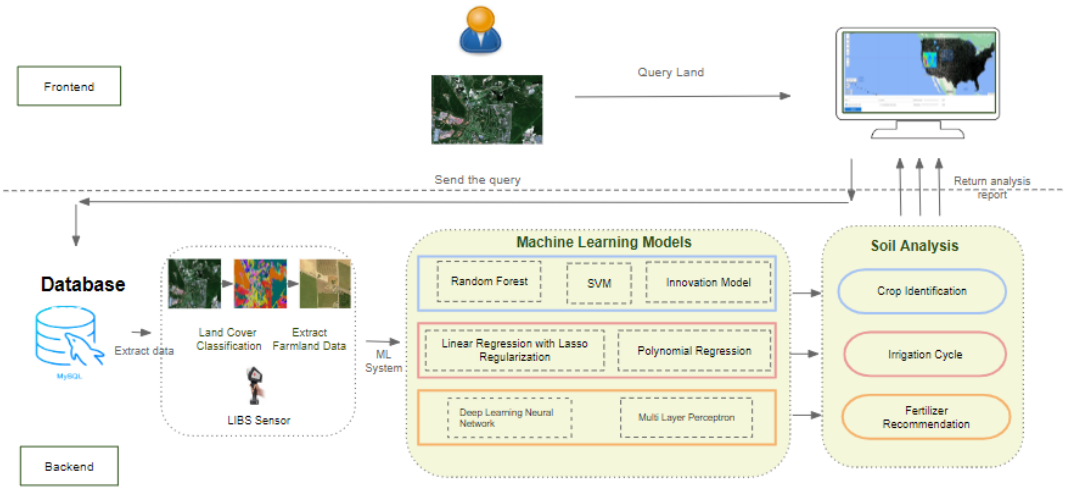
5.2.1 System Architecture and Infrastructure

Figure 68 shows our system architecture (frontend and backend) of this project. The scenario begins with a user (e.g., farmer, scientist, or technician) and the interested farmland.

The user can interact with our system through web application. Through the user interface, the user can specify the farmland information by entering geographical location or selecting the area of interest from the visualized map. After our frontend receives the query of the farmland, the frontend then sends the query to the backend. According to the information, the backend extracts the corresponding data including both satellite images and LIBS sensor data (McMillan, 2018) and sends them to our trained machine learning model. Our machine learning model consists of three modules: crop identification, irrigation cycle prediction and fertilizer recommendation. In the first module, random forest, SVM and innovation model are deployed in identifying the crop. The irrigation cycle prediction module is performed by using linear regression with lasso regularization and polynomial regression. The fertilizer recommendation is mainly conducted by the deep learning methods including MLP. After three modules complete our comprehensive soil analysis, the report of the soil analysis is returned to the frontend and is displayed on the user interface.

Figure 68

Full Stack System Architecture



5.2.2 System Data Management Solution and Data Repository Design

- System Data Management

We are going to use Google Cloud SQL (MySQL database service) to store and utilize raw and processed data. The data coming from sources (satellite, libs) are CSV data set which would be uploaded to a shared google drive directory. An ETL batch process will pick this data and store / archive in MySQL database tables with an archival timestamp. A sequential process for transforming raw data written in python will transform data using pandas and the transformed data will be stored in separate tables in MySQL database. The transformed data can directly be accessed by google collab for machine learning/deep learning training modules. The whole process would be automated using a python ETL framework that includes a logging mechanism, job scheduling mechanism and an interface to run the model on demand. Various components of data management solution being applied in this project execution is represented in the Figure 69 below.

Figure 69

Data Management Solution Architecture

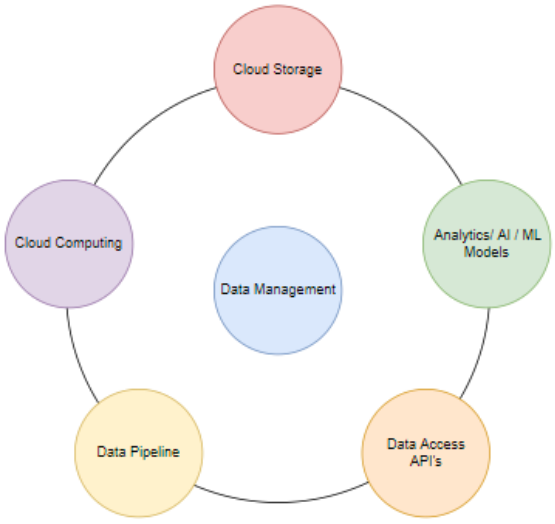
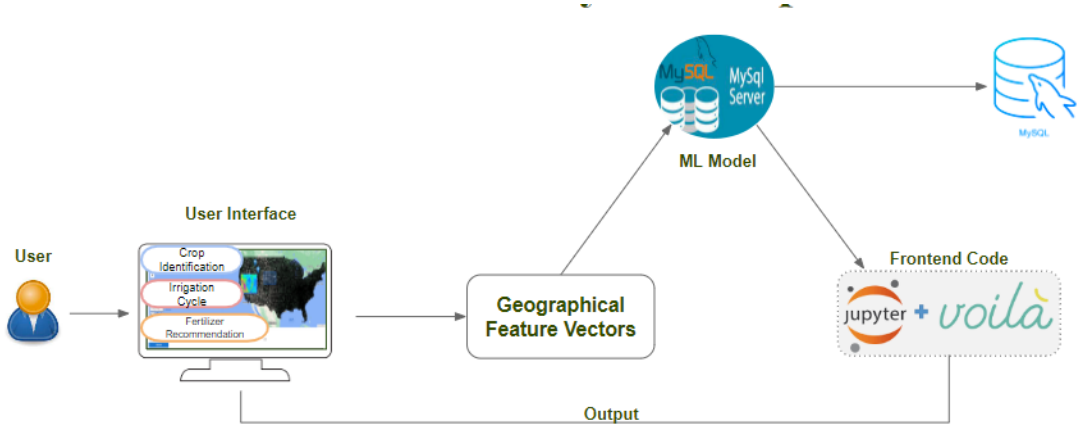


Figure 70 shows the data flow of our system. First, the user is required to enter the geographical location or selecting the area of interest. The frontend user interface then consolidates the information into a geographical feature vector and send to the MySQL server. After receiving the geographical feature vectors, the server first queries the MySQL database to extract the corresponding CSV files of satellite images and LIBS sensor data. Then the server utilizes the stored ML model to perform the soil analysis output and passes to the soil analysis system for result processing. Finally, the output result is displayed on the user interface.

Figure 70

Data Flow



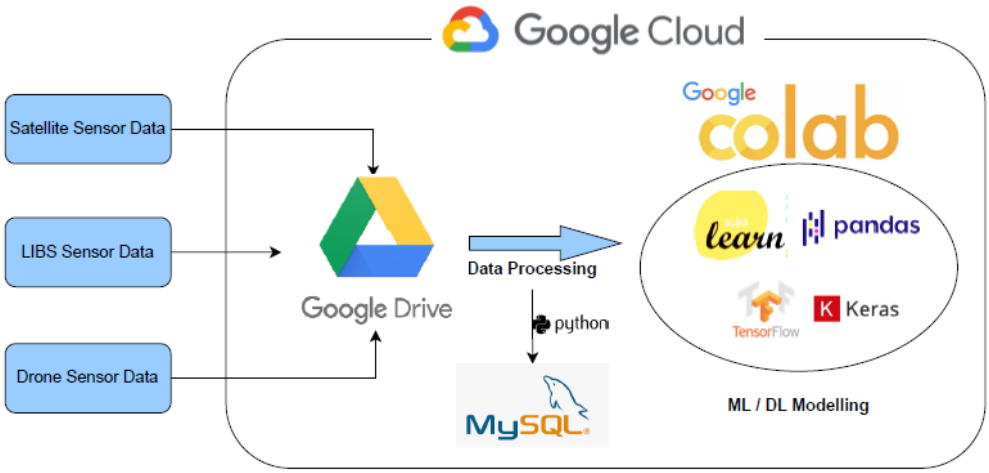
- Data Repository Design

For this project, we aim to come up with a comprehensive machine learning model that can perform crop identification, irrigation cycle recommendation and soil quality analysis in term of nitrogen estimation in soil and based on that recommend fertilizer application. The raw data that we are going to get is coming from satellite spectral sensor data, spectral sensor data from drone mounted cameras and LIBS sensor data from handheld LIBS equipment. The data is sourced from respective devices and uploaded to a common storage space in google drive in a google cloud space. Within the same google cloud space, we are going to make use of google

collab for data processing and transformation using pandas. The transformed data would be saved into google drive and will be further used of machine learning modelling using python scikit-learn package and deep learning modelling involving TensorFlow and Keras package in google collab. For saving the transformed/processed data, we are going to use google cloud SQL (MySQL service). Saved model (ML/DL) would reside on shared google drive itself and any application that would require to use saved model would be given access to saved model on demand basis for any downstream application. The cloud framework and design are depicted in the flowchart Figure 71 below.

Figure 71

Supporting Platform, Cloud Framework / Environments.

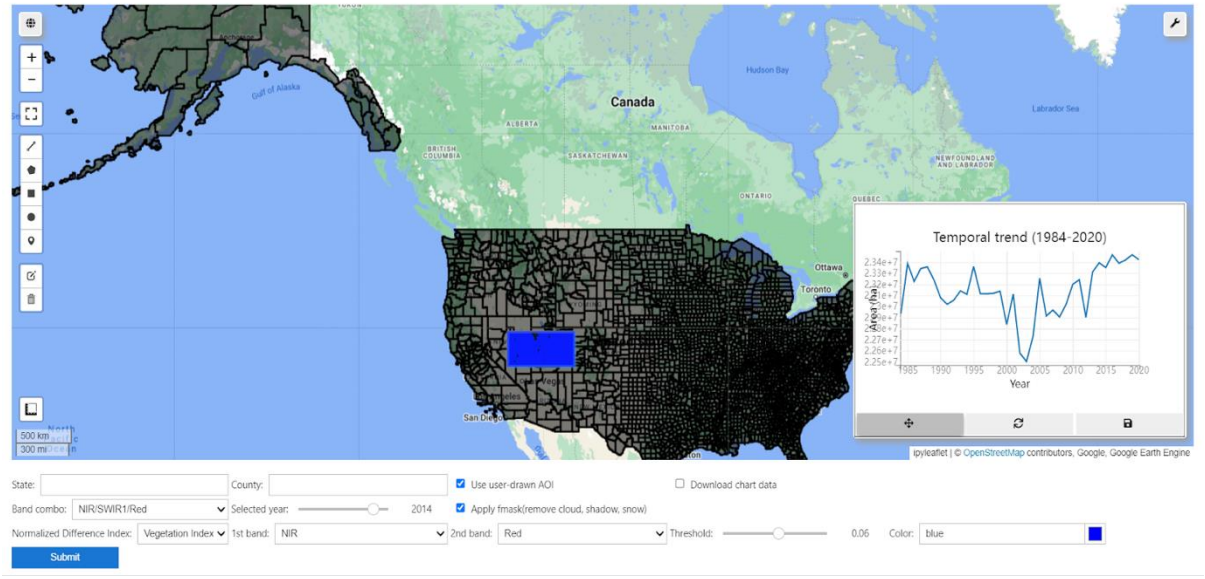


5.2.3 Design User Interface and Data Visualization

We designed the user interface (UI) to allow the users to interactive with crop and irrigation information (Figure 72). For the crop identification model, the UI is designed as an interactive map for users to check different vegetation indices, such as NDVI, NDWI, in certain space of the US. Moreover, the user can choose his own interest area as the AOI and download the chosen time series data into csv file.

Figure 72

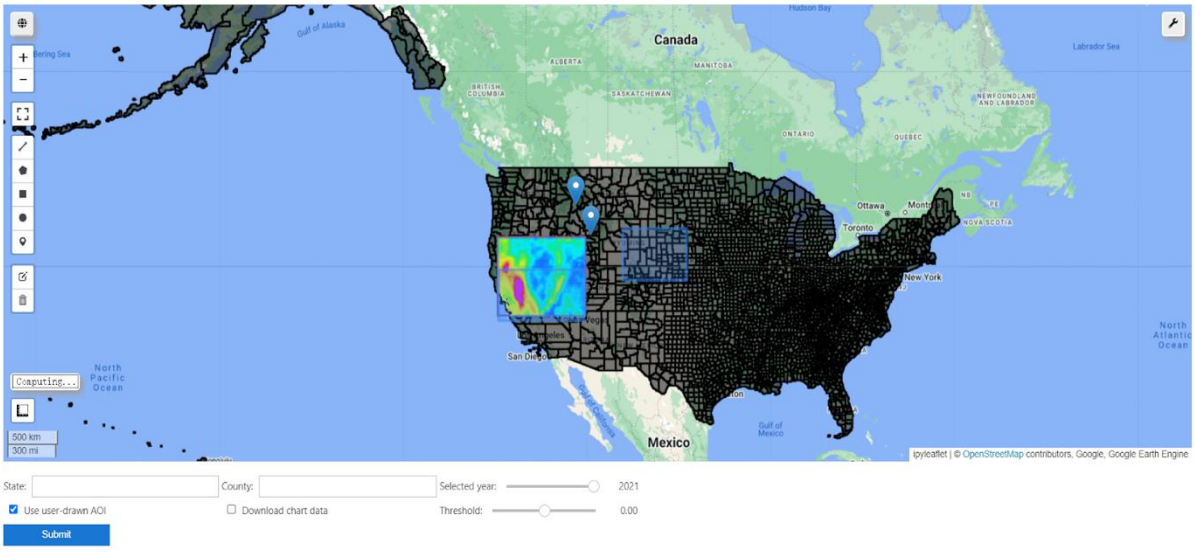
UI of the Vegetation Index



There are six widgets in the irrigation system UI: state button, county button, selected year slide bar, AOI check button, download check button and threshold slide bar. We can select any county and state in the U.S. Alternatively, draw user's own interest area is allowed. Selected year ranges from 1985 to 2022. Threshold value ranges from 0.0 to 1.0.

Figure 73

UI of the Irrigation System



5.3 Intelligent Solution

5.3.1 AI and Machine Learning Models

The system is built up with three sub platforms to analyze the surface and underground farmland situations. For the crop identification part, classification models from machine learning, random forest tree is deployed finally to predict categories. For the irrigation analysis part, regression models with lasso regularization is used to calculate the demand of water. For the fertilizer recommendation part, we deployed MLP from deep learning to integrate the geolocation information into farmland analysis.

1. Random Forest tree selected for crop identification system

Random Forest tree is an ensemble machine learning model, which gives the highest accuracy score for our prediction result among all the machine learning models. 89% crop types can be classified correctly under random forest tree, with no max depth limitation.

2. Linear regression model selected for irrigation recommendation system

A regression model is formed with multiple inputs to the equation. By changing the value of coefficient, we set one linear regression model with Lasso regulation and a polynomial

regression model. The output we analyzed depend on multiple crop water requirements, such as evaporation, runoff, transpiration, underground water, and surface heat. Thus, lasso regulation is added to control overfitting by minimizing the sum amount of coefficient weights. The Polynomial regression model allows us to transfer low degree inputs into high dimensional inputs, exploring other possible coefficient weights.

3. MLP selected for fertilizer recommendation system

A Multilayer perceptron model is a simple neural network model that is widely used in Classification kind of use cases. A MLP model can be tuned in terms of choosing activation function such as Sigmoid, or Relu and optimized using various optimizers and methods according to the data set and trained in a way to get desired result.

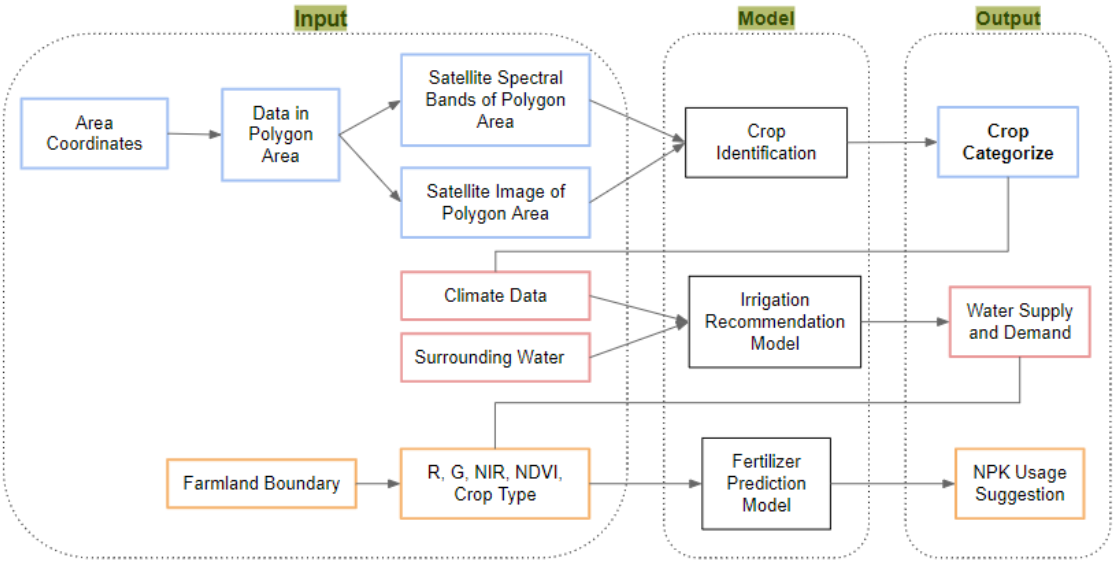
5.3.2 Inputs, Outputs, Supporting System Contexts

1. Input Datasets:

The data flow shown in figure 74 shows of the whole system includes four parts: user input, agriculture inputs, model, and module outputs, the outputs connect the three modules and support the results. At the first input part, users give the area of interest such as latitude and longitude coordinates to settle the specific location they want to focus on. Then the system will retrieve the selected area from the global satellite dataset in Google Earth Engine. Based on the algorithm to retrieve data, the system will pull the spectral bands and images from the dataset. The output from the crop identification model flows into the second module as the inputs, combined with climate data and surrounding water data. The chemical prediction model will be fed with soil moisture data from the second module and spectral bands data from the first module.

Figure 74

Inputs/Output System Requirements



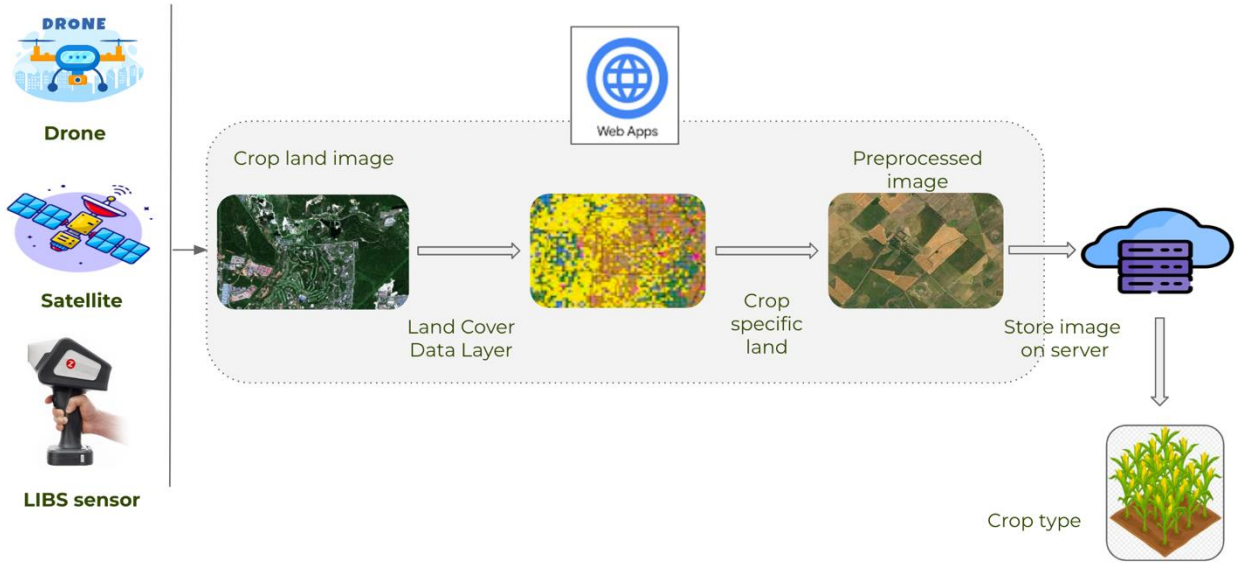
2. Expected Outputs and Supporting System:

Crop identification module: An initial land image collected from satellite and drone photography is not useful to the machine learning models. Through finding the reflectance values and temperature value from the land surface, we get the spectral bands to calculate the vegetation indices as one output. Moreover, the pixel value signatures are retrieved from each related band of the image, and the farmland boundary coordinates, as another output, are able to be obtained under the boundary detection algorithm.

The module is designed as the flow in the following picture. After the collection and preprocessing of sensor spectral data, image pixel and location data, the results are put in a GitHub repository and then backup in the local storage.

Figure 75

Supporting System for Crop Identification Module



The sensor data include temperature, moisture, humidity, and water lever, they are sent to the server through IoT, combined the data extracted from satellite images from module one. Output is the volumetric water needed per day for the farmland, which is calculated from the regression algorithms. The phone app is allowed to observe the real-time volatility of sensor data. The GUI of the system served as the numerical and graphical presentation to the client as shown in Figure 76.

Figure 76

Supporting System for Irrigation Module

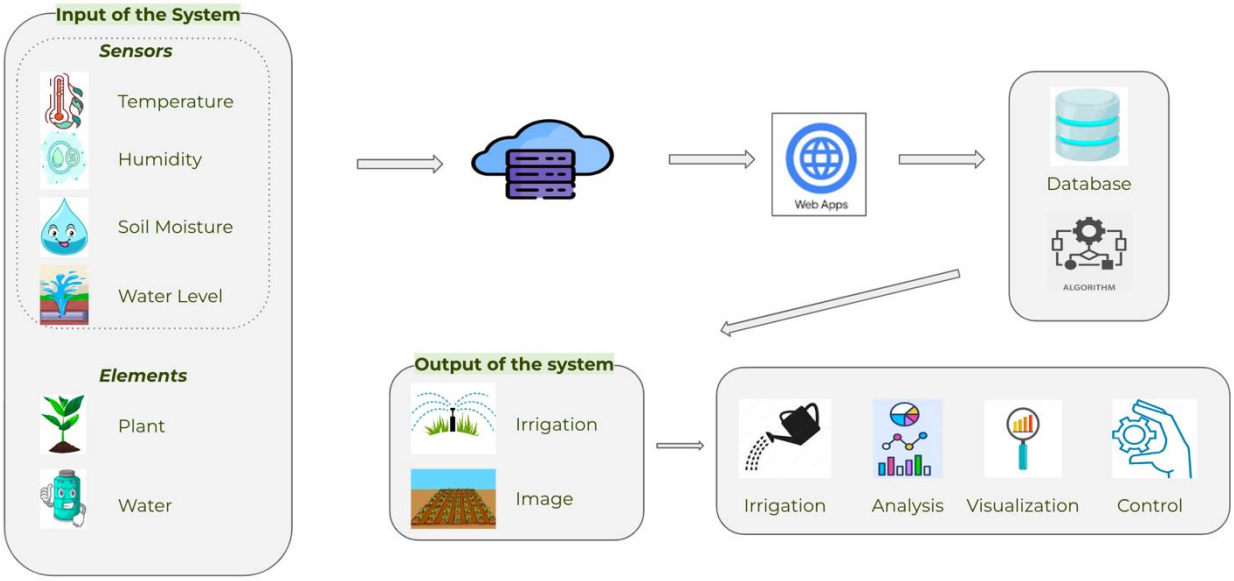
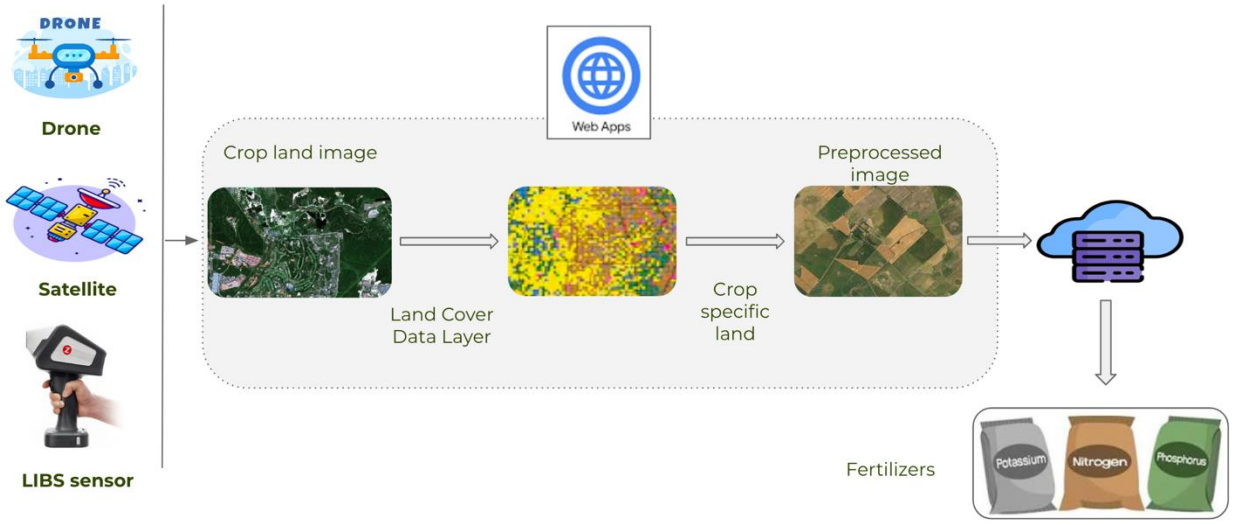


Figure 77

Supporting System for Fertilizer Module



5.3.2 Solution APIs

a. Google earth engine APIs support multiple functions for our project.

Opensource Python libraries (GitHub repo): formulate and translate computer codes into request to earth engine servers

REST API: give the access to Earth engine servers through HTTP.

b. ArcGIS REST APIs help us use the location services stored in the cloud.

Spatial analysis service: process spatial datasets

Elevation service: generate elevation views and profiles

The geometry service: create geographic shapes

Stream service: provide real-time and low latency data

Base map layer service: access satellite base map

c. Google vision APIs support the farmland image detection.

Label detection: detect the categories of different objects

Boundary detection: detect the boundary of the land

REST API: give the access to image servers

5.4 System Development and implementation

1. Interacting Subsystem

The whole system is composed of four interacting subsystems in a hierarchical manner. The feature of crop is the first subsystems should be extracted from clients' request. Then the client can check the irrigation situation of that farmland with the spectral features from first sub-function. When the use wants to know the nutrients constrained in the soil, he should customize the land area and time. Then the system will present the NPK recommendation for that specific farmland based on client's requirements.

2. Spatial Boundary

The analytical area can be defined using subjective boundary, one with the administration segmentation, another with the customized polygon area. After delineating the map to identify the study area, farmers make the specific land usage purposes. For example, farmers in Napa Valley may focus on the land in the north part of California. If drought happened this year, they

may put emphasis on the quantity of irrigation recommendation for next year or quantify the nutrient loads in their farmlands.

3. Time Dimensional System

Other than the traditional approach to use just one single image from one season as the input, we incorporate multiple seasons' image as inputs on the farmland segmentation task. The delineating result improved a lot in non-regular shape segmentations. Client can pick properties of the farmland on different time coverage. The dataset is linked to the satellite and sensor dataset which contains data covering four seasons of years.

4. Hierarchical System

Each level in the nested hierarchy has its own properties and supports to the larger system. The analytical field can be separated into one specific farmland, and then be aggregated into county, state, and country. The properties from the first level (crop identification) support the inputs of the second level (irrigation recommendation), and all the features should be added into the last level (nutrients suggestion).

5. Integration

The system includes three agriculture analytical platforms that combine soil, climate, surrounding water, air, and ecological data to simulate crop types, water usage, and outputs of nutrients and chemicals in the farmland. The nutrients and chemicals prediction model integrates the satellite spectral bands, geolocation data and time series to predict the usage of NPK in soil with high accuracy. The irrigation recommendation model takes the surrounding water data into account to make suggestion of irrigation amount as output.

Chapter 6 System Evaluation and Visualization

6.1 Analysis of Model Execution and Evaluation Results

By tuning all the models using machine learning methodology, we keep training our models to improve the accuracy of our system with selected crop types. The accuracy of selected crop types from 'G','R','F','P','T','D','C','V' is higher than the scores of all types. In order to keep an accurate classification result, eight crop types are selected in the application we deployed.

For crop identification model which shown in Table 27, we used Random Forest, Ensemble, Support Vector Machine (SVM), and Logistic Regression. Using the original Random Forest method gives us the accuracy is 78%. The Ensemble model shows the accuracy of 78% with the standard deviation is 0.1. While the other two models give lower accuracy level with SVM, and Logistic Regression are both 72%.

Table 27

Crop Identification Results

Models	**Accuracy of Selected Crop Type**	**Accuracy of All Crop Type**
Logistic Regression	72%	38%
Support Vector Machine	Kernel = linear 71.87% Kernel = rbf 73.78% Kernel = poly 72.57%	Kernel = linear 41.31% Kernel = rbf 45.31% Kernel = poly 43.73%
Random Forest	78%	69%

Ensemble	78%	52%

To classify irrigation cycle model, we used Linear Regression Model, Linear Regression Model with Lasso Regularization, and Polynomial Regression Model. The MSE score for Linear Regression Model is 0.08, R-squared score is 0.939. The accuracy level for using Linear Regression with “ET_runoff” and “underground_water” is around 93.6% while without these two parameters, the accuracy level is 82.8%, which is around 9% lower. For Linear Regression with Lasso Regularization, we have tested with three different alphas, which is 0.1, 0.01, and 0.001 and the MSE scores are 0.10, 0.08, and 0.08, respectively. The accuracy with alpha at 0.1 is 91.7% with three features used, 0.01 is 93.5% with four features used, and 0.001 is 93.6% with six features used. For Polynomial Regression, the accuracy level is 93.6%, RMSE is 0.2446, and R-squared score is 0.939.

Table 28

Irrigation Cycle Results

Models	**RMSE**	**MSE**	**R2**
Linear Regression	0.256	0.06554	0.939
Lasso Regularization	0.292	0.08535	0.917
Alpha = 0.1	0.260	0.06759	0.934
Alpha = 0.01	0.256	0.06554	0.936
Alpha = 0.001	0.244	0.05983	0.939

Polynomial Regression	0.256	0.06554	0.939

For Fertilizer Recommendation system, we used Random Forest, Ada Boost, Logistic Regression, Support Vector Machine, and MLP (Multilayer Perceptron Classification). Figure 78, Table 29 shows the individual accuracy report for classifiers algorithms.

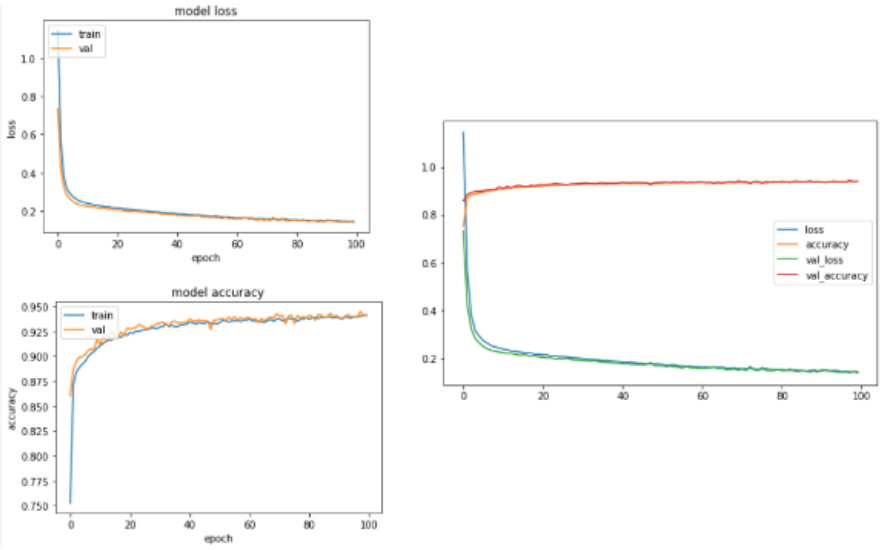
Table 29

Fertilizer – Nitrogen Recommendation Model Result

Model	Accuracy	Parameters
Random Forest	82%	estimators = 10
Ada Boost	82%	estimators = 10
Support Vector Machine	73%	
MLP (ReLU Activation)	94%	Trainable params: 3725, activation='relu', loss='categorical_crossentropy',optimizer='adam'

Figure 78

MLP (Multilayer Perceptron) using ReLU activation function (Accuracy, Loss, Model Properties)



In our results when we compare the accuracy of different models applied for fertilizer recommendation, we found that conventional classification methods like Random Forest seems to be performing well with source training data, while the MLP model works best with the combined training data of Crop Identification Model and LIBS sensor data.

6.2 Achievements and Constraints

The goal of this project is to come up with a comprehensive machine learning evaluation model for soil analysis. We first break down the comprehensive model into three tasks: (1) crop identification and (2) irrigation cycle and (3) fertilizer recommendation for the crop.

6.2.1 Achievements and Milestones

To solve these tasks, we list the necessary milestones that need to be achieved:

- Explore the relevant source and sensor data for soil analysis. Assessment of soil health involves knowing the performance of soil in physical, chemical and biological functions. The traditional laboratory analysis cannot provide country wide and high-resolution soil data. Therefore, instead of collecting laboratory data, we collect sensor-based data from satellite and soil sensors in a cost-effective way.

- Survey and investigate the related research fields and the machine learning methods.

The soil health analysis combines knowledge from agriculture, biology and geographic. We did academic research in crop growth factors, irrigation and environment system, fertilizer and chemical in agriculture, and then put all these information into one syncretic soil health analysis platform.

- Integrate the three models: crop identification model and irrigation cycle model with fertilizer recommendation model.

Three modules in our system are designed to integrate and support each other with data. At the first crop identification module, users give the area of interest such as latitude and longitude coordinates to settle the specific location they want to focus on. Then the system will retrieve the selected area from the global satellite dataset in Google Earth Engine. Based on the algorithm to retrieve data, the system will pull the spectral bands and images from the dataset. Based on the bands' inputs, crop types which will be counted as the input to second irrigation module will be calculated. The third module, fertilizer prediction, will be fed with soil moisture data from the second module and spectral bands data from the first module. The achievement from the model integration is the data flow and effective usage from each module.

- Fine-tune each model individually to improve the overall performance.

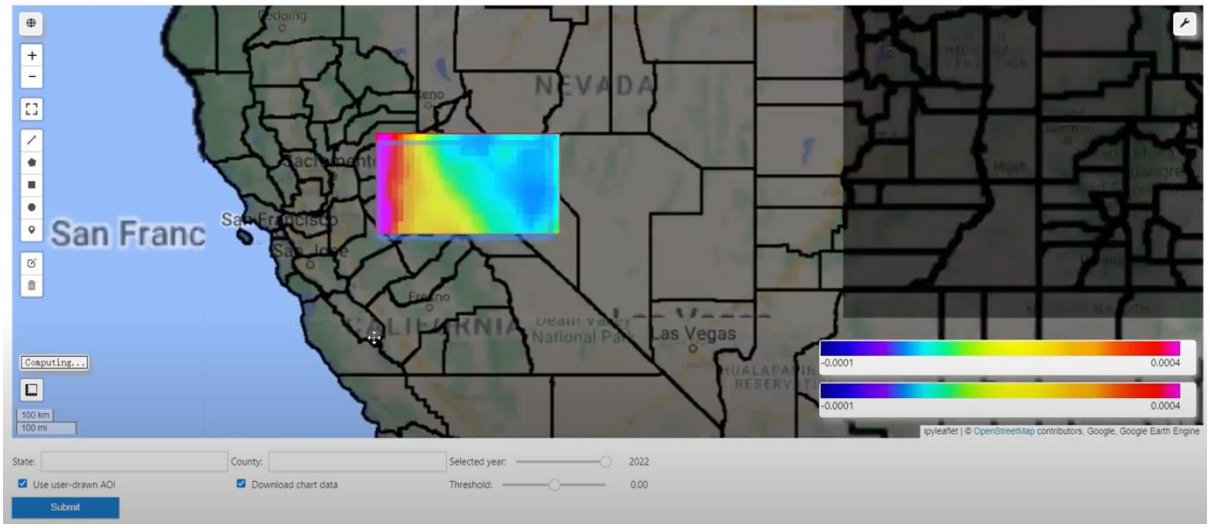
As the single models is hard to make up the best ensemble model, bagging multiple fairly weak machine learning models is a common way to improve the model performance. We tuned our forest tree model with XGboosting and bagging approaches and adding regularization to our linear regression models. In such a way, the improvement of the accuracy reached.

- Create a user-friendly interface to display the analysis result for users.

It is challenging to build an interface to display the data analysis result in a real-time map. We imported geemap function to create an interactive base map, and then add layer ’US counties and layer “US states” from GEE to segment the land area into state and county. The solution to capture user interaction in our map is to build a function called ‘handle_interaction’ to capture the click on the widgets we designed, and then build the ‘submit_click’ function to calculate the bands statistics based on the user inputs. The achievement we made in the interactive map is to link the GEE real-time database with the user’s request from our platform.

Figure 79

Sample Output from Integration



6.2.2 Constraints

The first and the most important constraint is the accessibility of the data. For example, we can identify the best crop of the field learning from the spectral bands. However, it is difficult for us to access the spectral bands data with specific location information. We add multiple geolocation layers about the state and county on the base GEE map, so that the bands dataset combined with location information such as longitude and latitude is downloaded. The advantage

of the solution is we are able to not only identify the crop type, but also to identify the farmland area at the meantime.

Next, we focus on the irrigation cycle for the surface since the other three types (i.e., sprinkler, drip/trickle and subsurface) are hard to acquire the relevant data for analysis. Through the fertilizer recommendation model, we have been able to predict the nitrogen content in soil using the spectral data and vegetation indices. The constraints in fertilizers recommendation model are availability of sufficient training data, including climate data such as precipitation, evaporation, surface water runoff, ground water amount.

Figure 80

Sample Code to Build the User Interface

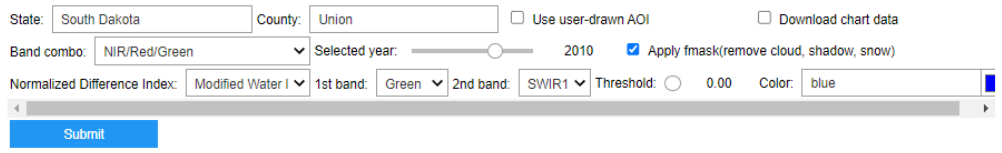
```
def nd_index_change(change):
    if nd_indices.value == 'Vegetation Index (NDVI)':
        first_band.value = 'NIR'
        second_band.value = 'Red'
    elif nd_indices.value == 'Water Index (NDWI)':
        first_band.value = 'NIR'
        second_band.value = 'SWIR1'
    elif nd_indices.value == 'Modified Water Index (MNDWI)':
        first_band.value = 'Green'
        second_band.value = 'SWIR1'
    elif nd_indices.value == 'Snow Index (NDSI)':
        first_band.value = 'Green'
        second_band.value = 'SWIR1'
    elif nd_indices.value == 'Soil Index (NDSI)':
        first_band.value = 'SWIR1'
        second_band.value = 'NIR'
    elif nd_indices.value == 'Burn Ratio (NBR)':
        first_band.value = 'NIR'
        second_band.value = 'SWIR2'
    elif nd_indices.value == 'Customized':
        first_band.value = None
        second_band.value = None


nd_indices.observe(nd_index_change, names='value')

submit = widgets.Button(
    description='Submit', button_style='primary', tooltip='Click me', style=style
)

full_widget = widgets.VBox(
    [
        widgets.HBox([admin1_widget, admin2_widget, aoi_widget, download_widget]),
        widgets.HBox([band_combo, year_widget, fmask_widget]),
        widgets.HBox([nd_indices, first_band, second_band, nd_threshold, nd_color]),
        submit,
    ]
)

full_widget
```



6.3 Quality Evaluation of Model Functions and Performance

The correctness of the model will be evaluated in each model classification probability. The run-time performance of meeting system will be evaluated by response time targets by using python time(). The response time differs from each device we used according to different specs.

The average runtime performance is calculated in seconds. The runtime is difference based on the devices and the processors that is used (Table 30):

- Crop Identification: the average runtime is 1.737 seconds
- Irrigation Cycle: the average runtime is 0.014 seconds
- Fertilizer Recommendation: the average runtime is under five minutes.

Table 30

Features Run-Time Performance Comparisons

Features	Devices	Average Run-time Performance (seconds)
Crop Identification	Razer Blade13 Processor: Intel i7-8565U CPU @ 1.80GHz Memory:16 GB GPU: GeForce MX150	1.737
Irrigation Cycle	MacBook Pro Processor: 2.8GHz Quad-Core Intel i7 Memory: 16GB 2133MHZ GPU: Intel HD Graphics 630	0.014
Fertilizer Recommendation	HP Omen PC Processor: Intel Core i7 Memory: 32 GB GPU: Nvidia 3080ti, 12 GB	< 5 Mins

Table 31

Model Performance Approach Comparisons

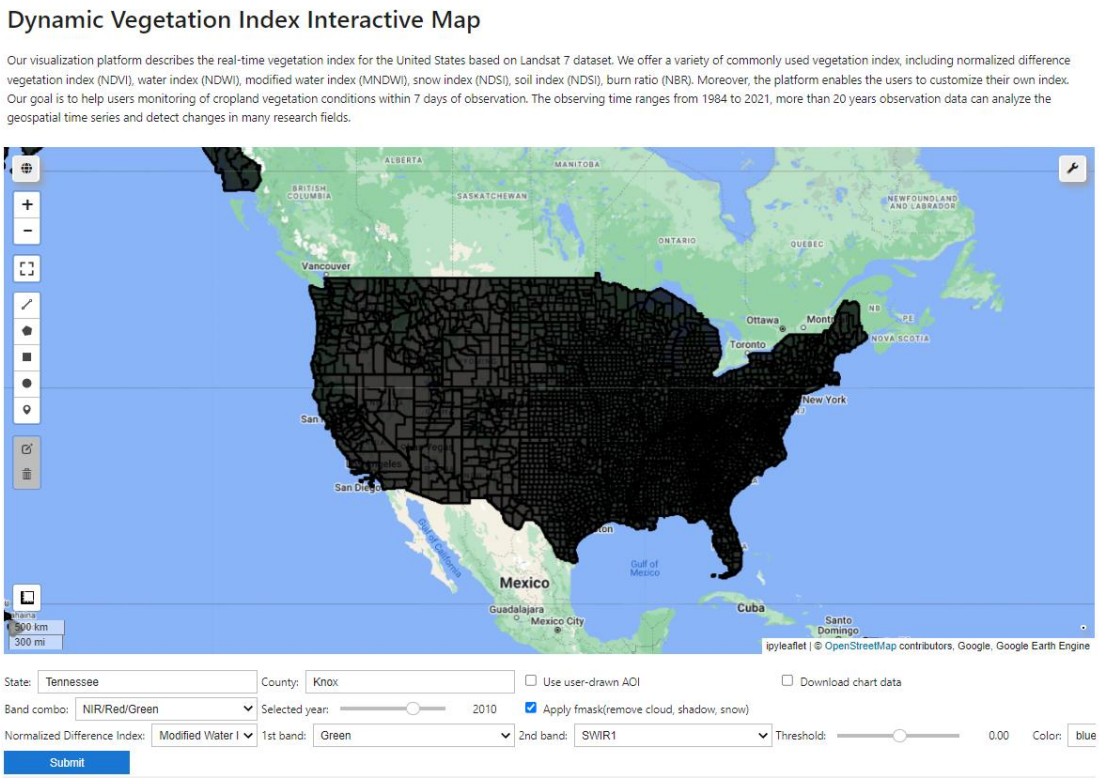
Features	Model Performance Approach	Purpose
Crop Identification	Accuracy Confusion Matrix Precision F1 score	Correct prediction frequency
Irrigation Cycle	MSE, RMSE R squared	Loss and cost
Fertilizer Recommendation	Loss Accuracy Confusion Matrix Precision	Distance from true value to prediction value

6.4 Project Information Visualization

Our visualization platform describes the real-time vegetation index for the United States based on Landsat 7 dataset. We offer a variety of commonly used vegetation index, including normalized difference vegetation index (NDVI), water index (NDWI), modified water index (MNDWI), snow index (NDSI), soil index (NDSI), burn ratio (NBR). Moreover, the platform enables the users to customize their own index. Our goal is to help users monitoring of cropland vegetation conditions within 7 days of observation. The observing time ranges from 1984 to 2021, more than 20 years observation data can analyze the geospatial time series and detect changes in many research fields.

Figure 81

Dynamic Vegetation Index Interactive Map UI



At the top left of the UI, there is a weigh to select location or data that the user is interested in. There are three ways to search the location, the first way is ‘name/address’ function searching by place name or address such as Paris, the second way is ‘lat-lon’ function searching by ‘lat-lon’ such as 40, -100, the third way is ‘data’ function searching by GEE data catalog by keywords such as elevation. The map can zoom in to the county-level and can click certain county in USA such as San Joaquin. Once the county is selected, the user can select whether download the chart data to check the statistics.

In Figure 82, the upper picture shows the selection widget of normalized difference index, which are commonly used in soil analysis. The lower picture gives the combinations of spectral bands.

Figure 82

Left: Location Selection Widget; Right: County-level Location Selection

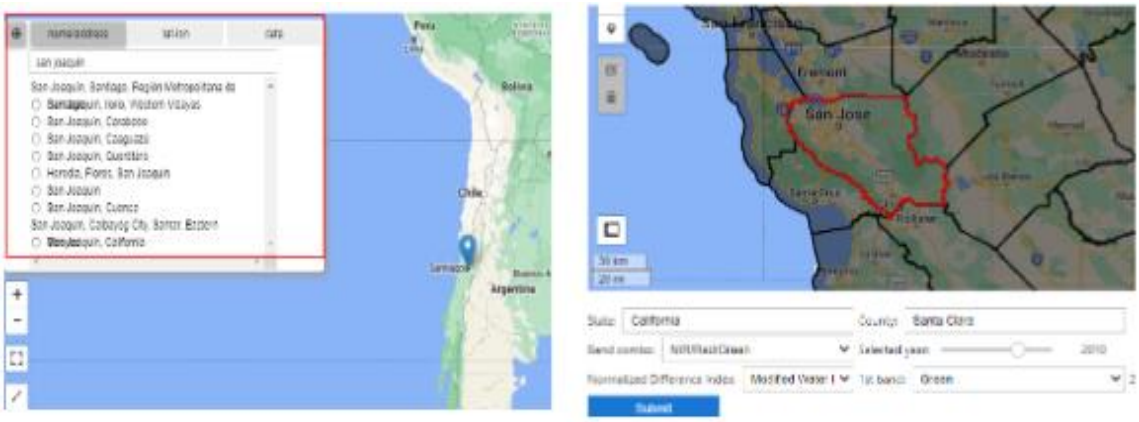
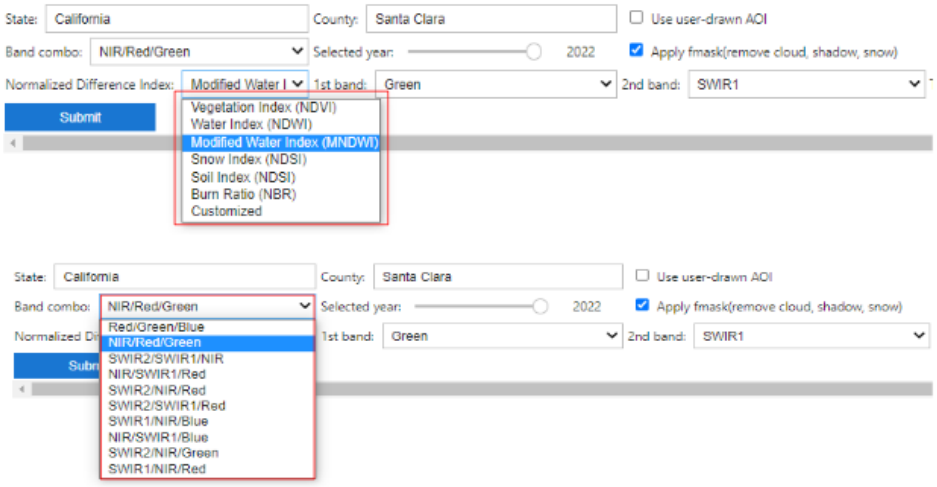


Figure 83

Upper: Selection Button for Normalized Difference Index

Lower: Spectral Bands Combinations

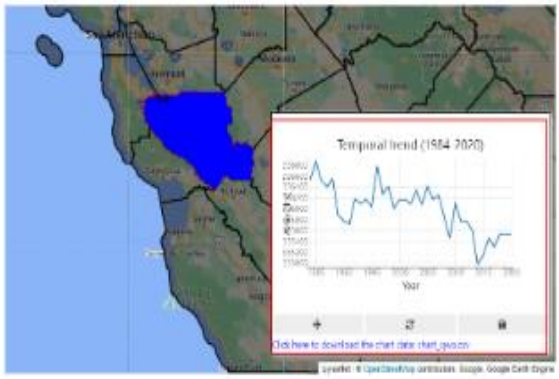


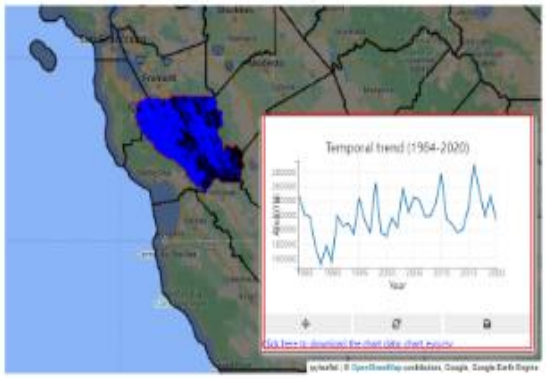
The statistical data is calculated based on the selected location and years. In the figure 84, we checked the NDVI and NDWI for the same location, Santa Clara, from 1985 to 2020. At the bottom of the panel are ‘panzoom’, ‘refresh’ and ‘save’ button. When the user clicks the ‘panzoom’ buttom, they are able to see the detailed line for certain time interval. Then the line chart can be saved into png file.

Figure 84

Left: Chart Data for NDVI from 1985 to 2020

Right: Chart Data for NDWI from 1985 to 2020





Chapter 7 Conclusion

7.1 Summary

7.1.1 Major Achievements and Major Findings

One of the main achievements of this report is to combine soil health analysis with crop type, land water and fertilizer usage. The precision of the crop type classification is dependent on the spectral bands data, then the crop type information will be infused into the irrigation prediction system. All these crop and land water data are the inputs to the fertilizer analysis. In such way, we minimize the number of parameters coming from outside resource and use our own dataset in a cyclic utilization.

The second main achievement is the land segmentation based on vegetation indices on pixel level. The land segmentation skill improved the performance of classification machine learning model result. And the segmentation is a simple way to replace the deep learning approach to delineate the land boundary.

The third main achievement is the irrigation calculation with high R squared. The theory behind the calculation comes from natural hydrologic system, which is a cyclic water environment of the earth including the precipitation, evaporation, transpiration, streamflow and underground water. We calculate the crop water supply (CWS) against and crop water

requirement (CWR). Both parameters consist of several separated data sources. For the crop water supply (CWS), precipitation and underground water pumping are included. For the crop water requirement (CWR), evaporation and runoff are considered.

The fourth main achievement is that the prediction of fertilizer is only based on small number of input parameters, spectral bands (R, NIR, Green, NDVI), and crop types. No IoT nor pH sensors are used to collect agriculture chemicals.

7.1.2 Key Points and Implications in Each Section

Chapter one is the introduction of the whole project with 5 sections: background and summary, project requirements, project deliverables, technology and solution survey, and literature survey. Our goal is to provide an expert system in both soil analysis and irrigation product recommendation for people involved in agriculture fields and is user-friendly without requiring prior agriculture knowledge. The system is a web-based application.

Chapter two is data and project management plan. We made plans about data management, project development, project organization, project resource requirements. The data we collected include satellite spectral bands, soil, and environment property data. We developed two model cycles, a neural network deep learning cycle and a machine learning cycle, to implement the project tasks.

Chapter three is the data engineering part with 7 sections: data process, data collection, data pre-processing, data transformation, data preparation, data statistics, data analytics results. We did ETL to the original datasets before building the data pipeline, and then analyze the data distribution based on data statistic results. We found that the vegetation indices have high correlation in one same season, and irrigation system is dependent on multiple hydrologic data.

Chapter four is the model development with 5 sections: model proposals, model supports, model comparison and justification, model evaluation, and validation results. To achieve the purpose of crop identification, classification machine learning models are used to group the features for different crop categories. Linear machine learning models are deployed in irrigation cycle recommendation system to get the demand and supply of usage water. Deep learning models are designed to predict the fertilizer usage for the third part.

Chapter five is the data analytics system with 4 sections: system requirement analysis, system design, intelligent solution, system development and implementation. Our use cases mainly happen in three environments: personal workstation, research labs and agriculture business company. The corresponding users of our system are farmers who plant and study in their personal workstation, scientists of research labs and technicians from agriculture business company.

Chapter six is the system evaluation and visualization with 5 sections: analysis of the model results, achievements and constraints, quality evaluation, and project information visualization. We integrated three research fields and create a user-friendly interface to display the analysis results.

7.2 Benefits and Shortcoming

1. Benefits

Agriculture is highly related to our daily life, especially the products such as rice, corn, or wheat play critical roles in supporting our health and diet. As the population worldwide is rising daily, the need for agriculture soars. Global climate change, on the other hand, poses a significant challenge for agriculture and hinders the development of agriculture since it impacts and alters the ecology. According to World Food Program, a global food crisis has already been

posted in 2022. It is about 828 million people starving every night. Therefore, it is desirable to develop a novel system to assist agricultural production. To meet the challenges, we proposed a soil analysis recommendation system for crop identification, irrigation cycle, and fertilizer. Our soil analysis recommendation system can help enthusiasts about agriculture, including farmers, agriculturists, and researchers, analyze and predict their future annual crop yield. The rationale behind that is that the revenue of agricultural products expands only if the soil is healthy. In addition, our soil analysis system incorporates satellite information. It takes the longitude and latitude of the field as inputs to assist the users in monitoring their interested field even if they work remotely, which benefits the remote research happening in companies or labs for soil analysis.

2.Shortcomings

Despite all the benefits that our application has brought to the users, like others existing products, our application also has a few shortcomings. First, our application currently only focuses on farms in the United States, which restricts the usage of applications. In the future, we want to expand our application and include more regions in our system to achieve the goal of global service. In this case, we can connect all the farmers, agriculturists, and researchers together and further create a global agriculture network. Besides, our current application analyzes the soil based on limited factors and elements. However, the soil itself is complicated, and its property should relate to broader aspects. For example, local climate, local animals, or even the surrounding industry can affect the soil conditions. By involving more elements, we can enhance our AI system's robustness and accuracy in the future**.**

7.3 Potential System and Model Applications

The soil health analysis is routinely challenged by the dynamic and complicated nature of environmental changes. A dynamics modelling system is able to demonstrate continuous value across the farmlands to help users understand and predict the dynamic environment and behavior in support of the decision making. We will try to build a data pipeline which can update data automatically.

We have assembled a set of attributes for a soil health detection system, based on three attributes: crop types, soil biological status, and soil nutrient status. These features provide data to be used in our final platform for the users to assess the soil health information. However, we narrowed our study range from 18 crop types into 8, to improve the accuracy and efficiency of the machine learning model.

7.4 Experience and Lessons Learned

7.4.1 Data Collection

Choosing data collection approaches and tools are critical to the performance of the project. Our project is highly dependent on real-time agriculture and environment data, it is not easy to collect such data without any advanced sensors or machines. Satellite sensor data is the most useful data for us, as the highly reliability and correctness of the dataset. Sensors installed on drones or AVE machines is also an efficient way to collect data from one specific area. The location is precise, but the collecting cost is higher, and the collecting process is complicated.

7.4.2 Sensor Data Polygons

Delineating the boundary for farmlands is not easy with the crop classification mission. The crop classification dataset is based on the 250-meter resolution of Landsat 8, but the polygons we want to draw are based on farmland bound area. It’s hard to combine these two

layers into one coordinate system. We put polygon lay based on county level which is downloaded from US agriculture website to the crop classification layer.

7.4.3 Web Application

Building a real-time web application is complex in congregating all the datasets and information in one platform. We do not need to build every part of the web application from scratch, using some web editor or text editor makes the job more efficiently. The frontend editor we picked voila build in python, which can read our python scripts easily. It makes our functionalities working and finalizing our web application quicker.

7.5 Recommendations for Future Work

7.5.1 Data from Sensors Installed on Drones

We will try to collaborate with drone vendors with installed thermal sensors, Lidar (light detection and ranging) sensor, and ground penetrating radar (GPR) sensor and antenna configuration. Lidar sensor can help us create pixel points of farmlands and ground surfaces. This can be used to classify crop more correctly and find the ground features ordinarily hard to identify. The ground penetrating radar (GPR) sensor can demonstrate the buried object underground and help us to understand the underground layers more quickly.

7.5.2 Model Ttuning

We have tried feature selection and subset model training for linear and machine learning models. We fitted the model to 80% of our dataset and evaluate the model performance on the rest dataset using cross-validation. In the future we will try to search the tuning parameter space or the grid search, which will deploy every combination of parameters in our grid.

7.5.3 Engineering Crops to Climate-resilient and Disease-resistant

To meet increasing global food demand, farmers and scientists both aim to increase the crop yield and quality of crop. Our system can identify the irrigation cycle and know the usage of fertilizer.

7.6 Contributions and Impacts on Society

The change in ecology brought by global climate change severely impacts agriculture production. Besides, the spread of disease and the rise of disasters worldwide aggravate the food shortage. For example, a pandemic happened in 2019 due to the coronavirus disease (COVID-19). Paused activities, including business, industry, and so on, significantly plummeted agricultural production. However, our daily diet still existed despite the fact that agricultural production was hindered. Besides, the rise in population further emphasizes the fact of the urgent demand for agricultural products. Using AI to assist soil analysis can not only improve farm usage efficiency and production but also provide security for farmers in their agriculture business. There are some examples of the benefits to society, such as:

National. The country can monitor and predict agriculture production. Meanwhile, using the prediction to make relevant policies in order to secure and maintain people's life.

Economic. Since agriculture is the fundamental industry supporting the whole industry, having a stable agriculture production and development using soil analysis can ensure good economic development.

Educational. The students can verify the textbook materials (e.g., the relation between the weather and the soil type) by interacting with AI-based soil analysis. Meanwhile, since the soil analysis system is interactable, it can enhance and enrich the teaching instead of merely relying on words in textbooks.

Agriculture Industry. Agriculturists and researchers can investigate more sophisticated problems using our AI system. At the same time, their feedback can help us improve our AI system for future development to achieve the goals such as faster production or zero-chemicals fertilization.

Farmers. Farmers can easily monitor their own farms, such as soil conditions, which can reduce their time and expense. They also can utilize this AI system to improve their skill and techniques.

References

A Global Food Crisis: World Food Programme. UN World Food Programme. (n.d.). Retrieved November 30, 2022, from https://www.wfp.org/global-hunger-crisis#:~:text=2022%3A%20a%20year%20of%20unprecedented%20hunger&text=As%20many%20as%20828%20million,on%20the%20edge%20of%20famine.

Abhigyan. "Understanding Polynomial Regression!!!" Medium, Analytics Vidhya, 2 Aug. 2020, https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18.

Aung, H. L., Uzkent, B., Burke, M., Lobell, D., & Ermon, S. (2020). Farm parcel delineation using spatio-temporal convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 76-77).

Abuzar, M., Whitfield, D., McAllister, A., Lamb, G., Sheffield, K., & O'Connell, M. (2013, July). *Satellite remote sensing of crop water use in an irrigation area of south-east Australia. In 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS (pp. 3269-3272). IEEE.*

Agarwal, S., Bhangale, N., Dhanure, K., Gavhane, S., Chakkarwar, V. A., & Nagori, M. B. (2018, July). Application of colorimetry to determine soil fertility through Naive Bayes classification algorithm. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

About arcgis pro. About ArcGIS Pro-ArcGIS Pro | Documentation. (n.d.). Retrieved April 21, 2022, from https://pro.arcgis.com/en/pro-app/2.8/get-started/get-started.htm

Barton, R. (2022). *Know Your Garden Soil: How to Make the Most of Your Soil Type*. Eartheasy Guides & Articles. https://learn.eartheasy.com/articles/know-your-garden-soil-how-to-make-the-most-of-your-soil-type/

Ball, J. E., Kari, S., & Younan, N. H. (2004, September). Hyperspectral pixel unmixing using singular value decomposition. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium* (Vol. 5, pp. 3253-3256). IEEE.

Bhosle, K., & Musande, V. (2019). Evaluation of deep learning CNN model for land use land cover classification and crop identification using hyperspectral remote sensing images. *Journal of the Indian Society of Remote Sensing*, *47*(11), 1949-1958.

Buckley, S. (2020). Laser-Induced Breakdown Spectroscopy for Soil Measurements: Recent Progress and Potential.

Cadwr_land_use_viewer_version_3. (n.d.). Retrieved April 21, 2022, from https://gis.water.ca.gov/app/CADWRLandUseViewer/?page=home

Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., & Xue, X. (2019). Research on soil moisture prediction model based on deep learning. *PloS one*, *14*(4), e0214508.

Caturegli, L., Gaetani, M., Volterrani, M., Magni, S., Minelli, A., Baldi, A., ... & Grossi, N. (2020). Normalized Difference Vegetation Index versus Dark Green Colour Index to estimate nitrogen status on bermudagrass hybrid and tall fescue. *International Journal of Remote Sensing*, *41*(2), 455-470.

California Water Science Center, U. S. G. S. (n.d.). *Central Valley: Drought indicators*. Central Valley Subsidence Data| USGS California Water Science Center. Retrieved April 22, 2022, from https://ca.water.usgs.gov/land_subsidence/central-valley-subsidence-data.html

Classification: Roc curve and AUC | machine learning crash course | Google developers. Google. Retrieved April 22, 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=An%20ROC%20curve%20(receiver%20operating,False%20Positive%20Rate

Chandrasekar, K., Sesha Sai, M. V. R., Roy, P. S., & Dwevedi, R. S. (2010). Land Surface Water Index (LSWI) response to rainfall and NDVI using the MODIS Vegetation Index product. *International Journal of Remote Sensing*, *31*(15), 3987-4005.

Crop monitoring: Satellite-based software for agricultural needs. (n.d.). Retrieved March 25, 2022, from https://crop-monitoring.eos.com/main-map/fields/all

Data On a Tangent - Jiji C. (2021, March 1). Evaluation metrics 101. Medium. Retrieved November 28, 2022, from https://medium.datadriveninvestor.com/evaluation-metrics-101-7c8b4c3421c2

Decision tree in machine learning: Types, advantages, disadvantages in 5 points. Jigsaw Academy. (2021, January 13). Retrieved April 22, 2022, from https://www.jigsawacademy.com/blogs/data-science/decision-tree-in-machine-learning/#:~:text=Advantages%20of%20decision%20tree%3A&text=It%20can%20handle%20both%20continuous,to%20credit%20the%20missing%20values.

Dr. Martin Luther King Jr. library. San Jose State University Library. (n.d.). Retrieved April 21, 2022, from https://ascelibrary-org.libaccess.sjlibrary.org/doi/full/10.1061/%28ASCE%290733-9437%282007%29133%3A4%28380%29

Earle, S. (2015, September 1). *14.1 groundwater and aquifers*. Physical Geology. Retrieved April 22, 2022, from https://opentextbc.ca/geology/chapter/14-1-groundwater-and-aquifers/

EARTH OBSERVING SYSTEM. (2022, March 4). *Vegetation indices as a satellite-based add-on for Agri Solutions*. Vegetation Indices to Drive Digital Agri Solutions. Retrieved March 21, 2022, from https://eos.com/blog/vegetation-indices/#:~:text=Red%2DEdge%20Chlorophyll%20Vegetation%20Index,activity%20of%20the%20canopy%20cover.

Escalante, H. J., Rodríguez-Sánchez, S., Jiménez-Lizárraga, M., Morales-Reyes, A., De La Calleja, J., & Vazquez, R. (2019). Barley yield and fertilizer analysis from UAV imagery: a deep learning approach. *International Journal of Remote Sensing*, *40*(7), 2493-2516.

Food And Agriculture Organization. (2022). *Food and Agriculture Organization of the United Nations*. Www.Fao.Org. https://www.fao.org/fileadmin/user_upload/soils-2015/docs/EN/EN_Print_IYS_food.pdf

Frost, J. (2021, November 14). *Mean squared error (MSE)*. Statistics By Jim. Retrieved April 22, 2022, from https://statisticsbyjim.com/regression/mean-squared-error-mse/

Gandhi, R. (2018, July 5). Support Vector Machine - introduction to machine learning algorithms. Medium. Retrieved November 27, 2022, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gandhi, R. (2018, May 28). Introduction to machine learning algorithms: Linear regression. Medium. Retrieved November 27, 2022, from

https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

Giasson, E., Clarke, R. T., Inda Junior, A. V., Merten, G. H., & Tornquist, C. G. (2006). Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. Scientia Agricola, 63, 262-268.

Gewali, U. B., Monteiro, S. T., & Saber, E. (2018). Machine learning based hyperspectral image analysis: a survey. *arXiv preprint arXiv:1802.08701*.

Google. (n.d.). Google Earth engine. Retrieved April 21, 2022, from https://earthengine.google.com/

Google. (n.d.). *Classification: Roc curve and AUC | machine learning crash course | google developers*. Google. Retrieved April 22, 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=An%20ROC%20curve%20(receiver%20operating,False%20Positive%20Rate

Goyal, C. (2021, May 31). *Artificial Neural Network | Beginners Guide to ANN*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/beginners-guide-to-artificial-neural-network/

Gulati, M. (2020, October 11). *Choosing Evaluation Metrics for Classification Model*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/

GISGeography. (2021, October 29). *What is NDVI (normalized difference vegetation index)?* GIS Geography. Retrieved March 20, 2022, from https://gisgeography.com/ndvi-normalized-difference-vegetation-index/

Hossen, M. A., Diwakar, P. K., & Ragi, S. (2021). Total nitrogen estimation in agricultural soils via aerial multispectral imaging and LIBS. *Scientific Reports, 11*(1). https://doi.org/10.1038/s41598-021-90624-6

Jian, J. (2020, January 13). *A database for global soil health assessment*. Nature. https://www.nature.com/articles/s41597-020-0356-3?error=cookies_not_supported&code=54b2f0c4-94a8-4c19-acc1-ce6e9aa909d4

Jiang, Z., Liu, C., Ganapathysubramanian, B., Hayes, D. J., & Sarkar, S. (2020). Predicting county-scale maize yields with publicly available data. *Scientific Reports, 10*(1), 1-12.

Jha, K., Doshi, A., Patel, P., & Shah, M. (2019). A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture, 2*, 1-12.

K, D. (2020, December 26). *Top 5 advantages and disadvantages of Decision Tree Algorithm*. Medium. Retrieved April 23, 2022, from https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

Kumar, A. (2022, April 8). *Accuracy, precision, Recall & F1-Score - Python examples*. Data Analytics. Retrieved April 22, 2022, from https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

Landsat Mission. (n.d.). *Normalized difference moisture index*. Normalized Difference Moisture Index | U.S. Geological Survey. Retrieved March 21, 2022, from https://www.usgs.gov/landsat-missions/normalized-difference-moisture-index#:~:text=Normalized%20Difference%20Moisture%20Index%20(NDMI,SWIR%20values%20in%20traditional%20fashion.

Liu, L., Wang, B., Zhang, L., & Zhang, J. Q. (2007, July). Decomposition of mixed pixels using Bayesian Self-Organizing Map (BSOM) neural networks. In *2007 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2014-2017). IEEE.

Liu, N., Zhao, R., Qiao, L., Zhang, Y., Li, M., Sun, H., ... & Wang, X. (2020). Growth stages classification of potato crop based on analysis of spectral response and variables optimization. *Sensors*, *20*(14), 3995.

Maximizing Irrigation Efficiency and Water Conservation. Center for Agriculture, Food, and the Environment. (2016, November 14). Retrieved April 21, 2022, from https://ag.umass.edu/turf/fact-sheets/maximizing-irrigation-efficiency-water-conservation

Mahmud, I., & Nafi, N. A. (2020, December). An approach of cost-effective automatic irrigation and soil testing system. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) (pp. 1-5). IEEE.

Matsushita, B., Yang, W., Chen, J., Onda, Y., & Qiu, G. (2007). Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: a case study in high-density cypress forest. *Sensors*, *7*(11), 2636-2651.

Mandal, M. (2021, May 1). *CNN for deep learning: Convolutional Neural Networks*. Analytics Vidhya. Retrieved March 22, 2022, from https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/

Madhumathi, R., Arumuganathan, T., Shruthi, R., & Iyer, R. S. (2020, October). Soil Nutrient Analysis using Colorimetry Method. In *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)* (pp. 252-256). IEEE.

Memon, R., Memon, M., Malioto, N., & Raza, M. O. (2021, October). Identification of growth stages of crops using mobile phone images and machine learning. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)* (pp. 1-6). IEEE.

MicaSense. (2021, January 27). *What is Ndre?* Retrieved March 21, 2022, from https://micasense.com/what-is-ndre/.

MJH Life Sciences. (n.d.). *Laser-induced breakdown spectroscopy for soil measurements: Recent progress and potential*. Spectroscopy Online. Retrieved April 22, 2022, from https://www.spectroscopyonline.com/view/laser-induced-breakdown-spectroscopy-soil-measurements-recent-progress-and-potential

McMillan, N. (2018, June 15). *Laser-induced breakdown spectroscopy (LIBS)*. Methods. Retrieved September 23, 2022, from https://serc.carleton.edu/msu_nanotech/methods/libs.html

Näsi, R., Viljanen, N., Kaivosoja, J., Alhonoja, K., Hakala, T., Markelin, L., & Honkavaara, E. (2018). Estimating biomass and nitrogen amount of barley and grass using UAV and aircraft based spectral and photogrammetric 3D features. *Remote Sensing*, *10*(7), 1082.

National Centers for Environmental Information (NCEI). (n.d.). *Climate Data Online*. Climate Data Online (CDO) - The National Climatic Data Center's (NCDC) Climate Data Online (CDO) provides free access to NCDC's archive of historical weather and climate data in addition to station history information. | National Climatic Data Center (NCDC). Retrieved March 25, 2022, from https://www.ncdc.noaa.gov/cdo-web/

Northeast Region Certified crop adviser (NRCCA) study resources. Certified Crop Advisor study resources (Northeast region). (n.d.). Retrieved April 21, 2022, from https://nrcca.cals.cornell.edu/soil/CA3/CA0324.php

NIST LIBS, N. I. S. T. L. I. B. S. (2022). *NIST LIBS Database*. NIST LIBS. https://physics.nist.gov/PhysRefData/ASD/LIBS/libs-form.html

Northeast Region Certified crop adviser (NRCCA) study resources. Certified Crop Advisor study resources (Northeast region). (n.d.). Retrieved April 21, 2022, from https://nrcca.cals.cornell.edu/soil/CA3/CA0324.php

Nrcs. (n.d.). *Web soil survey - home*. Web Soil Survey - Home. Retrieved March 25, 2022, from https://websoilsurvey.sc.egov.usda.gov/

Odenweller, J. B., & Johnson, K. I. (1984). Crop identification using Landsat temporal-spectral profiles. *Remote Sensing of Environment*, *14*(1-3), 39-54.

Patil, V. K., Jadhav, A., Gavhane, S., & Kapare, V. (2021, March). IoT Based Real-Time Soil Nutrients Detection. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 737-742). IEEE.

Parreiras, T. C., Lense, G. H. E., Moreira, R. S., Santana, D. B., & Mincato, R. L. (2020). Using unmanned aerial vehicle and machine learning algorithm to monitor leaf nitrogen in coffee.

Project jupyter. Project Jupyter. (n.d.). Retrieved April 21, 2022, from https://jupyter.org/

Pollatos, V., Kouvaras, L., & Charou, E. (2020). Land Cover Semantic Segmentation Using ResUNet. *arXiv preprint arXiv:2010.06285*.

Quantum. (2022, June 22). *Crop field boundary detection: Approaches and main challenges*. Medium. Retrieved September 23, 2022, from

https://medium.com/geekculture/%D1%81r%D0%BE%D1%80-field-boundary-detection-approaches-and-main-challenges-46e37dd276bc

R, B. (2022). *Know Your Garden Soil: How to Make the Most of Your Soil Type*. Eartheasy Guides & Articles. https://learn.eartheasy.com/articles/know-your-garden-soil-how-to-make-the-most-of-your-soil-type/

Ray, S. (2021, August 26). *Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Sarparast, M. (2019, May 1). *MSAVI: Modified Soil-adjusted vegetation index in LSRS: Land Surface Remote Sensing*. MSAVI: Modified Soil-Adjusted Vegetation Index in LSRS: Land Surface Remote Sensing. Retrieved March 21, 2022, from https://rdrr.io/cran/LSRS/man/MSAVI.html

Silver, W. L., Perez, T., Mayer, A., & Jones, A. R. (2021). The role of soil in the contribution of food and feed. Philosophical Transactions of the Royal Society B: Biological Sciences, 376(1834), 20200181. https://doi.org/10.1098/rstb.2020.0181.

Sgma.water.ca.gov. (n.d.). Retrieved April 22, 2022, from https://sgma.water.ca.gov/webgis/?appid=SGMADataViewer

Sruthi, S., & Aslam, M. M. (2015). Agricultural drought analysis using the NDVI and land surface temperature data; a case study of Raichur district. *Aquatic Procedia*, *4*, 1258-1264.

Singh, A. (2020, November 28). *Ensemble Learning | Ensemble Techniques*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-model

Singh, J., Devi, U., Hazra, J., & Kalyanaraman, S. (2018, July). Crop-identification using sentinel-1 and sentinel-2 data for indian region. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 5312-5314). IEEE.

Tao, X., Wang, B., Zhang, L., & Zhang, J. Q. (2007, July). A new scheme for decomposition of mixed pixels based on nonnegative matrix factorization. In *2007 IEEE International Geoscience and Remote Sensing Symposium* (pp. 1759-1762). IEEE.

Tatsumi, K., Yamashiki, Y., Torres, M. A. C., & Taipe, C. L. R. (2015). Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data. *Computers and Electronics in Agriculture*, *115*, 171-179.

T. (2022, January 10). *Types of Crops | Classification and Basics of Agriculture*. CropForLife. https://cropforlife.com/classification-types-of-crops-basics-of-agriculture/

Using voilà. Using Voilà - voila 0.3.5 documentation. (n.d.). Retrieved April 22, 2022, from https://voila.readthedocs.io/en/stable/using.html

USDA. (2022). *What is Soil? | NRCS Soils*. USDA. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/edu/?cid=nrcs142p2_054280

USGS. (2022). *Landsat Missions - Data | U.S. Geological Survey*. USGS. https://www.usgs.gov/landsat-missions/data

What is the F1-score? Educative. (n.d.). Retrieved April 22, 2022, from https://www.educative.io/edpresso/what-is-the-f1-score

Appendices

Appendix A – System Testing

In Appendix A, we will provide our system testing and also, we will upload our system testing folder at the google drive. Our application will be separated into three main scenarios: Vegetation Index, Irrigation, and NPK.

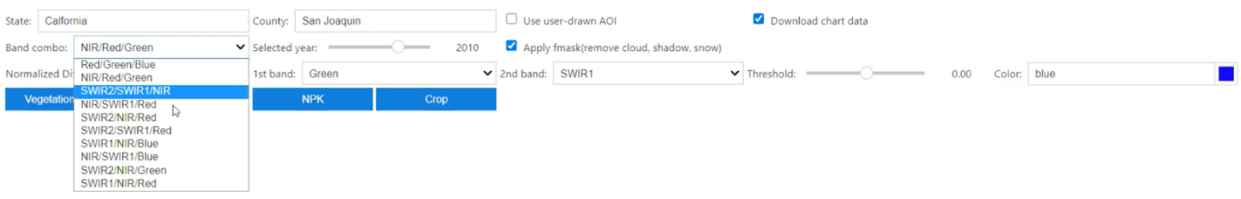
1. Home Page

Figure 1

Web Application Home Page

Figure 2

Users Can Select All the Information (State, County, Band Combo, Year, Color, NDVI, 1st Band, 2nd Band, Apply fmask, Threshold, and Download Chart Data)



2. Vegetation Index

Figure 3

Users Hit Vegetation_Index Button

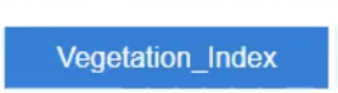


Figure 4

Visualization Chart for The Selected Area is Displayed at the Bottom Right

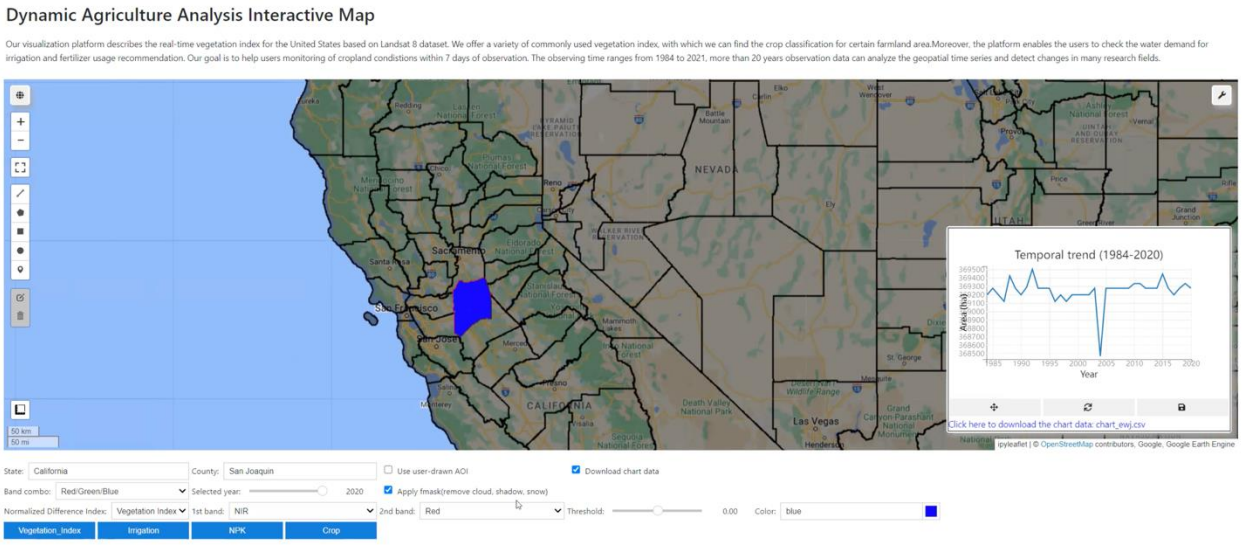


Figure 5

Visualization Chart for The Selected Area is Displayed at the Bottom Right

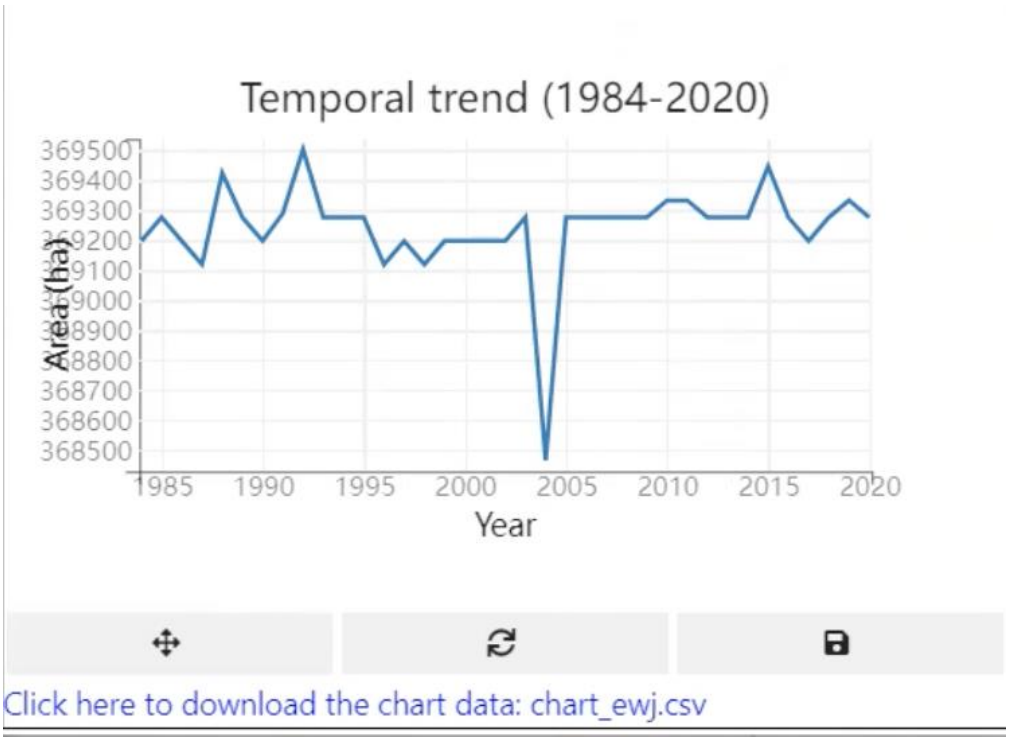


Figure 6

Users Can Use this Button to Download the Chart

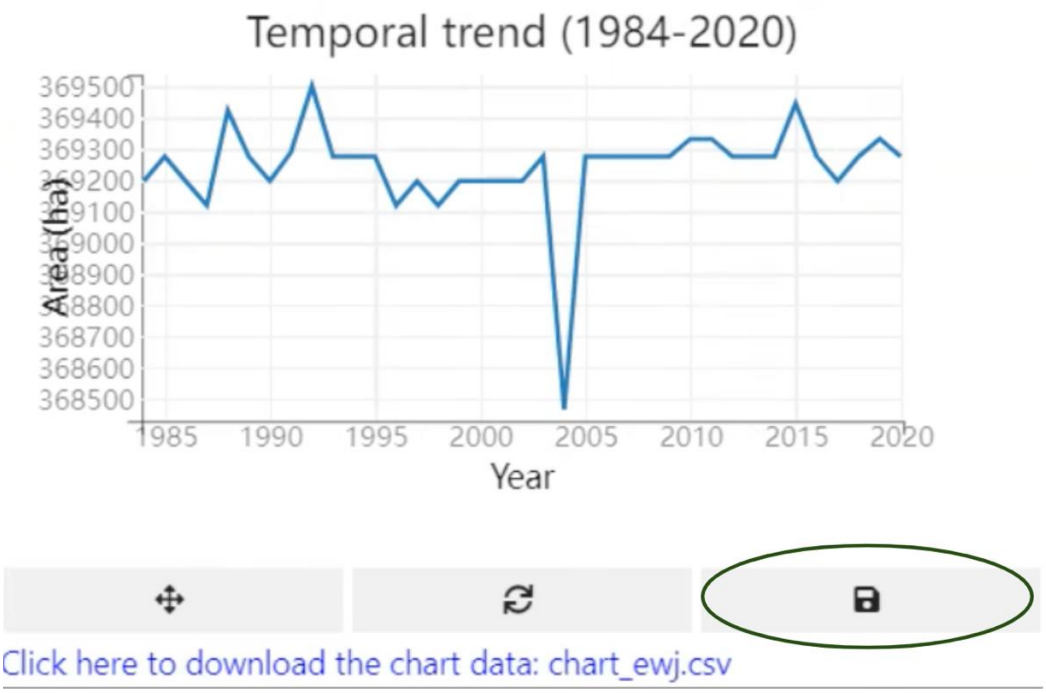


Figure 7

Users Can Click Here to Download the Chart Data

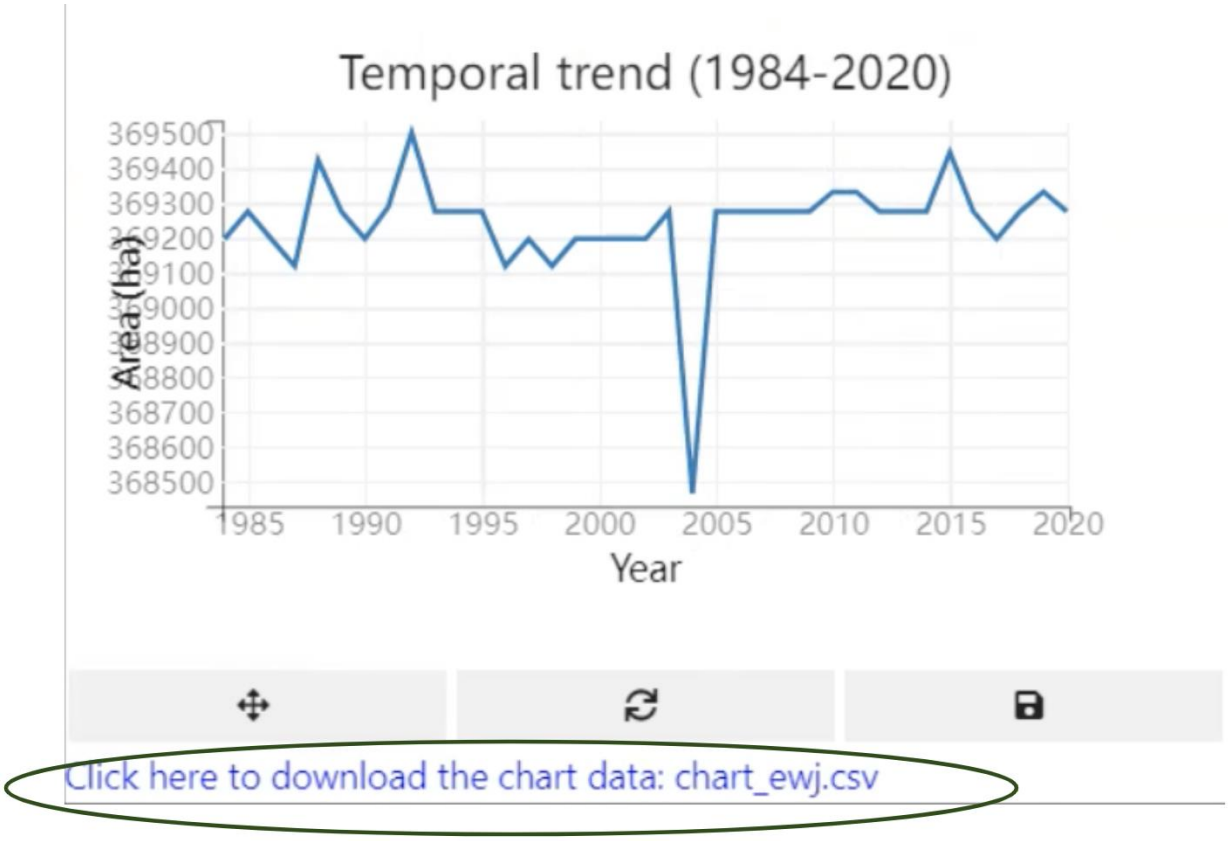


Figure 7

Chart Data Downloaded in CSV Format

	year	area (ha)
2	1984	369197
3	1985	369275.7
4	1986	369197
5	1987	369118.6
6	1988	369423.1
7	1989	369275.7
8	1990	369197
9	1991	369288.9
10	1992	369501.8
11	1993	369275.7
12	1994	369275.7
13	1995	369275.7
14	1996	369118.3
15	1997	369197
16	1998	369118.3
17	1999	369197
18	2000	369197
19	2001	369197
20	2002	3691
21	2003	369275.7
22	2004	368466.1
23	2005	369275.7
24	2006	369275.7
25	2007	369275.7
26	2008	369275.7
27	2009	369275.7
28	2010	369332
29	2011	369332
30	2012	369275.7
31	2013	369275.7
32	2014	369275.7
33	2015	369445.5
34	2016	369275.7
35	2017	369197.3
36	2018	369275.7
37	2019	369332
38	2020	369275.7

3. Irrigation

Figure 8

Users Can Locate to the Irrigation Button

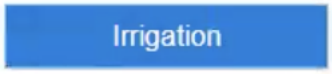


Figure 9

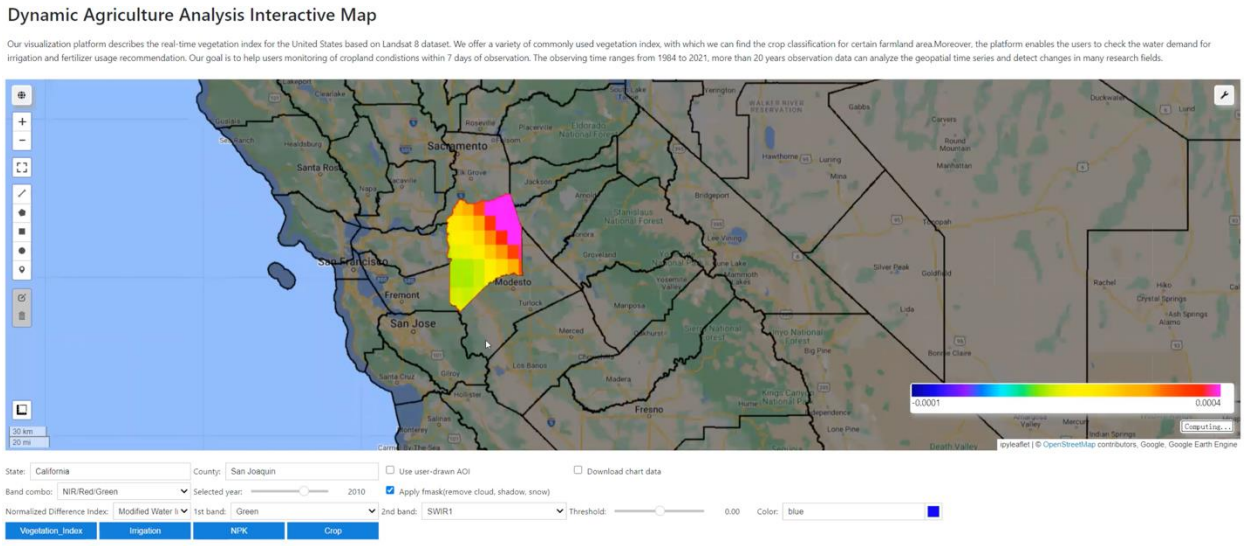


Figure 9

Users Draw an AOI at Any Location

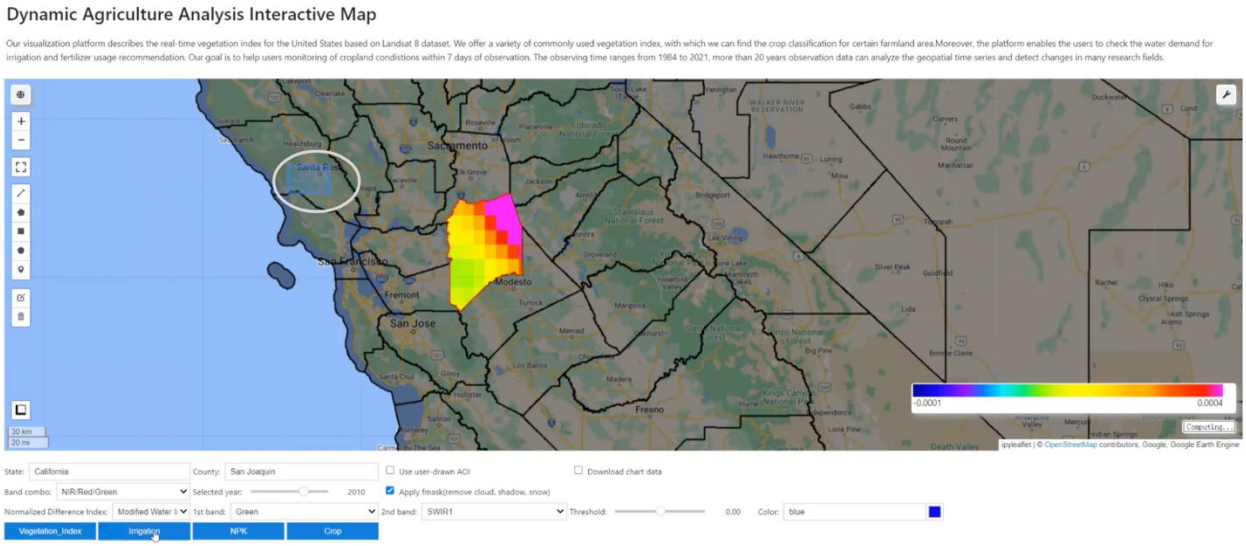
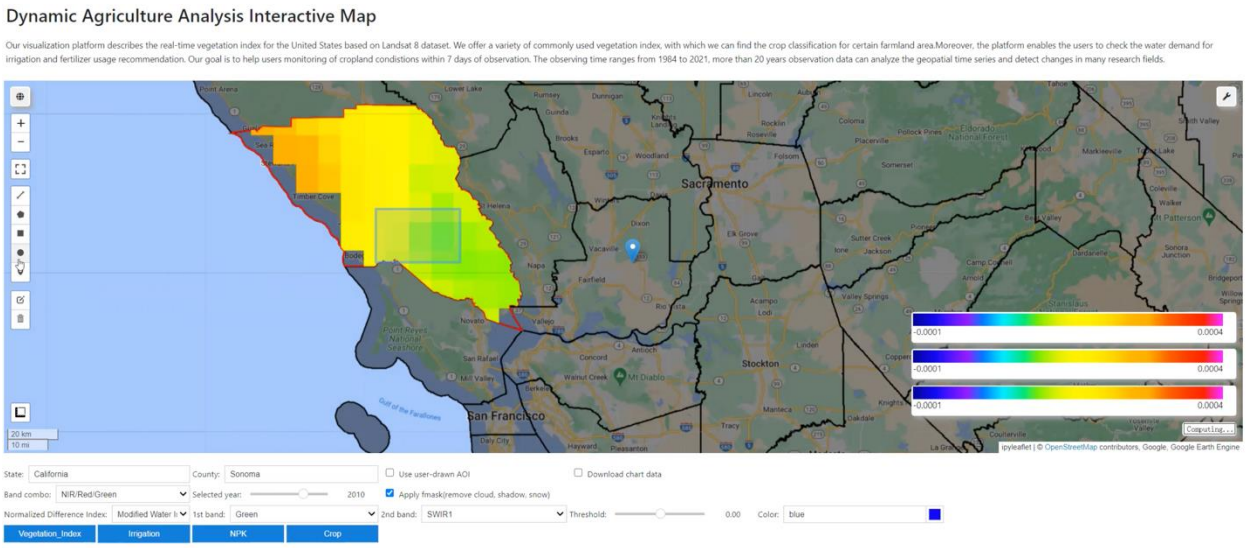


Figure 10

Irrigation of Selected Area Is Displayed



4. NPK

Figure 10

Users Choose the Region of Interest

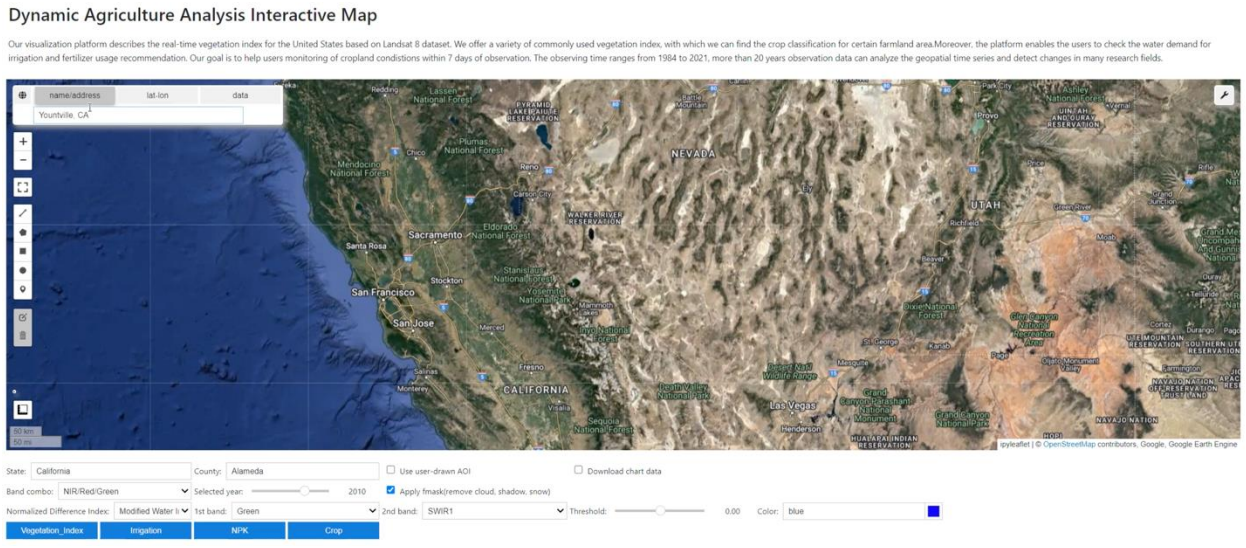


Figure 12

Users Draw AOI

Figure 13

Users Select NPK Button

Figure 14

Application Displayed NPK of Selected Area

Appendix B – Project Data Source and Management Store

We provided all of the datasets that we have used for model training and testing under the Dataset Folder. All of our source code includes the machine learning models and the UI code are uploaded under Machine Learning Program.

Appendix C – Project Program Source Library, Presentation, and Demonstration

Inside this PPTs Folder, we have provided all of our PowerPoints that we have made: one for the final demo in 298A, two demos in 298B (Demo 1 & 2) and one final demo. All of our reports are provided under Documents Folder. We have uploaded all of our reports in 298A, 298B, and the final report.