Brandon Nguyen, Ulises Urbina, Daniel Self
COMPE510 - Fall 2025

COMPE510 Term Project Final Report

1 - Introduction:

For the Semester Term Project in COMPE510, we are tasked with identifying a real-life engineering problem that can be solved by machine learning and an associated dataset. Afterwards, we must perform machine learning tasks, such as classification and regression etc., to the datasets, and report our results and observations. Lastly, each team member must implement a new machine learning algorithm that has not been discussed in class and also develop a GUI. For our real-life engineering problem, our group identified that it is difficult for farmers to forecast how much crop yield will be produced when there are many factors affecting crop growth. Our group believes the solution is to. By the end of this semester term project, our group hopes to fulfill the following requirements:. The dataset we found from Kaggle consists of. In conclusion, our group believes that implementing machine learning algorithms like XGBoost (eXtreme Gradient Boosting), K-Nearest Neighbors (KNN) Regression, and Support Vector Regression will help us achieve our goals, followed by a GUI.

2 - Related Work:

Existing research in smart agriculture highlights the value of machine learning for optimizing irrigation schedules, predicting soil moisture, classifying crop diseases, and performing yield forecasting. Many studies employ regression models to estimate continuous environmental values, while others use tree-based ensembles or kernel methods to capture nonlinear patterns in agricultural datasets. Prior work also emphasizes the importance of sensor reliability and anomaly detection to maintain data quality in large-scale deployments. Overall, the literature shows that machine learning methods consistently improve the accuracy and yield-prediction of agricultural decision support systems.

Unlike many prior studies that rely on neural networks or standard classroom algorithms such as decision trees and logistic regression, our term project focuses specifically on methods not covered in class such as Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). Our approach also emphasizes practical model comparison using a real smart-farm dataset based on sensor data from various countries, highlighting the trade-offs in accuracy, generalization, and computational cost between kernel-based, distance-based, and ensemble-based learning.

3 - Problem Formulation:

The objective of this project is to build a predictive model that estimates key smart-farm variables (such as soil moisture, humidity, temperature, or plant-health indices) using historical sensor data from a variety of different countries and continents across the globe. We are given a dataset containing observations collected from agricultural sensor nodes, where each record includes environmental attributes and corresponding output values to be predicted. The goal is to learn a regression function that maps the input attributes to accurate numerical predictions while handling noise, nonlinear relationships, and variations across time. Constraints include limited dataset size, variability in sensor quality, and the requirement to implement new machine-learning methods outside the scope of the course curriculum by ourselves. In the end, the goal is to create a product that determines the best farming practices and predict potential yield.

4 - Methods:

To address the prediction tasks, we implemented several machine learning algorithms not covered in class, including Support Vector Regression (SVR), XGboost, and K-Nearest Neighbors Regression. SVR was selected for its ability to model nonlinear relationships using kernel functions. K-NN provided a distance-based baseline that measures similarity between feature vectors. XGboost was used to utilize its effectiveness on classification and regression. Each method was trained using standardized feature values, tuned through hyperparameter searches, and evaluated using common regression metrics such as MAE, MSE, and R^2.

5 - Data Description and Experimental Setup:

The dataset used in this project originates from a publicly available smart-farm collection and contains numerical sensor measurements such as temperature, humidity, soil moisture, light intensity, and other environmental variables. The dataset includes several thousand samples, each represented as a feature vector with continuous attributes. Basic statistics, such as means, standard deviations, and attribute ranges were computed to understand the distribution and variability of the data. The dataset was divided into training and testing sets, typically using a 70–80% split, and all features were normalized to ensure consistency across models. Hyperparameters for each method (e.g., SVR kernel type, XGBoost learning rate and tree depth, and k-values for k-NN) were selected through iterative experimentation using validation results. All training and evaluation were performed in Python using scikit-learn and XGBoost's dedicated library.

6 - Experimental Results and Analyses:

The experimental results show the performance differences among SVR, XGboost, and KNN. SVR performed strongly when an appropriate kernel was selected, especially in capturing smooth nonlinear trends, but required more tuning. k-NN provided reasonable baseline predictions but was sensitive to feature scaling and offered lower accuracy on complex relationships. XGboost showed strong accuracy across metrics due to its gradient-boosting framework, ability to model non-linear relationships within our dataset, and to prevent overfitting. While SVR excelled in smoother prediction regimes. These results highlight the importance of model selection based on the data characteristics of smart-farm environments.

7 - Application and User Interaction:

The application developed for this project provides a simple interface that allows users to input sensor values and receive predicted environmental metrics or plant-health indicators. The interface displays model outputs, visualizes relevant trends, and provides warnings if the input data resemble previously identified anomalies. This design allows farmers or system operators to quickly evaluate conditions and make informed decisions about irrigation, crop care, or equipment adjustments. The application emphasizes ease of use, clear visualization, and real-time interaction capability.

8 - Conclusion:

This project demonstrates how machine learning models can be applied to smart-farm sensor data to support precision agriculture. Using algorithms outside the classroom material, such as Support Vector Regression, XGboost, and KNN, we developed predictive models that estimate important environmental variables and assist farmers with data-driven decision-making. Experimental results show that ensemble models provide strong performance on noisy agricultural data, while kernel methods offer effective nonlinear modeling when properly tuned. Overall, the project highlights the value of machine learning in enhancing agricultural efficiency and intelligent monitoring systems.

9 - Video Demo/Final Presentation:

Link to Final Presentation Video

10 - Reference:

Soundankar, Atharva. “🌾 Smart Farming Sensor Data for Yield Prediction.” Kaggle, 15 Apr. 2025,

www.kaggle.com/datasets/atharvasoundankar/smart-farming-sensor-data-for-yield-prediction.

“Understanding Support Vector Machine Regression.” MATLAB & Simulink, www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html. Accessed 4 Dec. 2025.

“XGBoost Documentation.” XGBoost Documentation - Xgboost 3.1.1 Documentation, xgboost.readthedocs.io/en/stable/. Accessed 4 Dec. 2025.

“XGBoost.” GeeksforGeeks, GeeksforGeeks, 24 Oct. 2025, www.geeksforgeeks.org/machine-learning/xgboost/.

“K-Nearest Neighbor(KNN) Algorithm.” GeeksforGeeks, GeeksforGeeks, 23 Aug. 2025, www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/.

“Support Vector Regression (SVR) Using Linear and Non-Linear Kernels in Scikit Learn.” GeeksforGeeks, GeeksforGeeks, 6 Aug. 2025, www.geeksforgeeks.org/machine-learning/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/.

11 - Individual Contributions:

For the semester term project, our group believes that the workload was evenly divided among each member. As stated in the introduction, each team member must implement a new machine learning algorithm that has not been discussed in class and also develop a GUI as one of the requirements of this term project. Displayed below is the breakdown of individual team member contributions regarding the term project.

Summary:

Brandon Nguyen: XGBoost algorithm (main), K-Nearest Neighbor Regression (help), Final Report and Slide Presentation

Daniel Self: K-Nearest Neighbor Regression (main), GUI Development (main), Final Report and Slide Presentation

Ulises Urbina: Support Vector Regression (main), GUI Development (help), Final Report and Slide Presentation