CIS 890 Introduction to Food Informatics

Aryan Singh Dalal and Prof. Dr Hande Küçük McGinty

Progress Report

Objective : The aim of this project is to develop a predictive model for the optimal application of nitrogen in agriculture. Nitrogen plays a crucial role in plant growth by facilitating photosynthesis and contributing to essential plant components like proteins and DNA. However, excessive use of nitrogen fertilizers can lead to salt accumulation that harms plant roots and foliage, making them more vulnerable to pests and diseases. Therefore, this project leverages open-source data to create a prediction model that will guide appropriate nitrogen application to crops, ensuring healthier and more resilient plant growth.

Approach's Overview:

1. Utilizes Machine Learning, specifically Linear Regression, to predict nitrogen usage in agriculture.
2. Employs data preprocessing to handle missing data and select relevant features.
3. Standardizes features through scaling for consistent analysis.
4. Utilizes Linear Regression, assuming a linear relationship between variables.
5. Evaluates model performance with MAE and MSE, aiming to optimize nitrogen application in crop cultivation.

Status of Deliverables: It begins by importing essential libraries for data processing, machine learning, and visualization. However, expected data was not provided by the specific domain expert. This resulted in a unique problem of insufficient data. To tackle this problem, open source data found on Kaggle was used. The new data found, had no measuring units and assumptions were made to overcome this, after this the dataset is loaded from a CSV file and preprocessed, with missing data, handled and split the data into training and testing sets. Categorical and numerical columns are identified, and preprocessing transformers are created for both types of data. These transformers are combined into a preprocessor using Column Transformer. The code then establishes a machine learning pipeline, including data preprocessing and a Random Forest Regressor model. The model is trained on the training data, used to make predictions on the test data, and evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R2). Lastly, the code visualizes the model's performance by plotting residuals, allowing for an assessment of how well the model's predictions align with actual values.

Choice of Model: In this project, a diverse set of regression models is employed to predict nitrogen usage in agriculture. Linear Regression establishes a baseline prediction, while Ridge and Lasso Regressions address multicollinearity issues and enhance model robustness. Support Vector Regression (SVR) is leveraged to capture complex non-linear relationships, whereas Random Forest Regression and Gradient Boosting Regression handle intricate data interactions and non-linearity. By utilizing this ensemble of models, the project aims to provide comprehensive insights into optimizing nitrogen application, effectively addressing different data challenges and achieving accurate predictions for crop cultivation.

Status of Deliverables: Multiple models have been compared, evaluated, and the most optimal one has been saved. Simultaneously, the user interface is prepared and the process of constructing the Neo4J graph is underway.

Status of Outcome: The project's outcome materialized in the successful prediction of nitrogen levels, marking its completion immediately upon the model's development.