AI SYSTEM IMPLEMENTATION

The crop recommendation system was implemented by developing a machine learning (ML) recommendation model. This section gives an overview of how the model was created, explaining the main steps and methods used.

The main goal here is to explain the tasks involved in making a precise and reliable ML model for crop and fertilizer recommendation. It talks about understanding farming, looking at available data, preparing the data and selecting important features, training the model, and then using it in the recommendation system.

Additionally, the section highlights the significance of rigorous testing and evaluation to ensure that the model's performance meets the desired standards.

Problem Understanding and Domain Analysis

The project began with the aim of making use of soil properties and weather to provide suitable and accessible crop recommendations to farmers. To develop an accurate model, a study was conducted to understand how soil properties affect the farming conditions of various crops.

The study uncovered some factors such as Nitrogen, Phosphorus, Potassium, temperature, humidity, pH, and rainfall as essential conditions for accurate recommendation. These factors were put into two categories, namely; soil nutrients, and environmental factors.

To supplement the main feature -crop recommendation- of the system, fertilizer recommendation was added to provide the user with a solution in case of deficiency.

Data collection, pre-processing, and analysis

The dataset used is from Kaggle and Global Yield Gap Atlas. Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC. It is a platform that provides reliable datasets for building AI models and taking part in competitions. The specific dataset used can be found here: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset.

The Global Yield Gap Atlas (GYGA) project provides a global dataset on yield gaps worldwide, including sub-Saharan Africa. The specific dataset can be found at https://www.yieldgap.org/.

Pre-processing is crucial in preparing the data for modeling since it lays the foundation for building accurate, reliable, and interpretable AI models. It includes:

i. **Scaling:** The notebooks mention scaling the data, which normalizes the feature values using a Standard Scaler. This step is essential because it ensures features with larger ranges don’t unduly influence the model. All the numerical values such as Nitrogen, phosphorus, pH, rainfall, humidity, temperature, and potassium were all scaled to a range of -1 to 1.
ii. **Feature Selection and Engineering:** Feature selection and engineering are critical steps in the data preparation process. They involve choosing the most relevant features and creating new ones to improve model performance. Examining the relationships between different soil and environmental factors to identify which features are most strongly

associated with crop suitability. This helps in selecting the most informative features for the model. The most prominent features identified during training that contributed to the high performance of the model were Rainfall and Nitrogen.

The Kaggle dataset has 22 labels for crops and while the fertilizer dataset has 7 labels for fertilizer, including Urea, DAP, 14-35-14, 28-28, 17-17-17, 20-20, and 10-26-26. Training data: 1540

1	N	P	K	temperature	humidity	ph	rainfall	label
2	90	42	43	20.87974371	82.00274423	6.502985292000001	202.9355362	rice
3	85	58	41	21.77046169	80.31964408	7.038096361	226.6555374	rice
4	60	55	44	23.00445915	82.3207629	7.840207144	263.9642476	rice
5	74	35	40	26.49109635	80.15836264	6.980400905	242.8640342	rice
6	78	42	42	20.13017482	81.60487287	7.628472891	262.7173405	rice
7	69	37	42	23.05804872	83.37011772	7.073453503	251.0549998	rice

Figure 1: Snapshot of the crop dataset used

Training Data: 99 data points

Temparature	Humidity	Moisture	Soil Type	Crop Type	Nitrogen	Potassium	Phosphorous	Fertilizer Name
26	52	38	Sandy	Maize	37	0	0	Urea
29	52	45	Loamy	Sugarcane	12	0	36	DAP
34	65	62	Black	Cotton	7	9	30	14-35-14
32	62	34	Red	Tobacco	22	0	20	28-28
28	54	46	Clayey	Paddy	35	0	21	Urea
26	52	35	Sandy	Barley	12	10	22	17-17-17
25	50	64	Red	Cotton	9	0	23	20-20

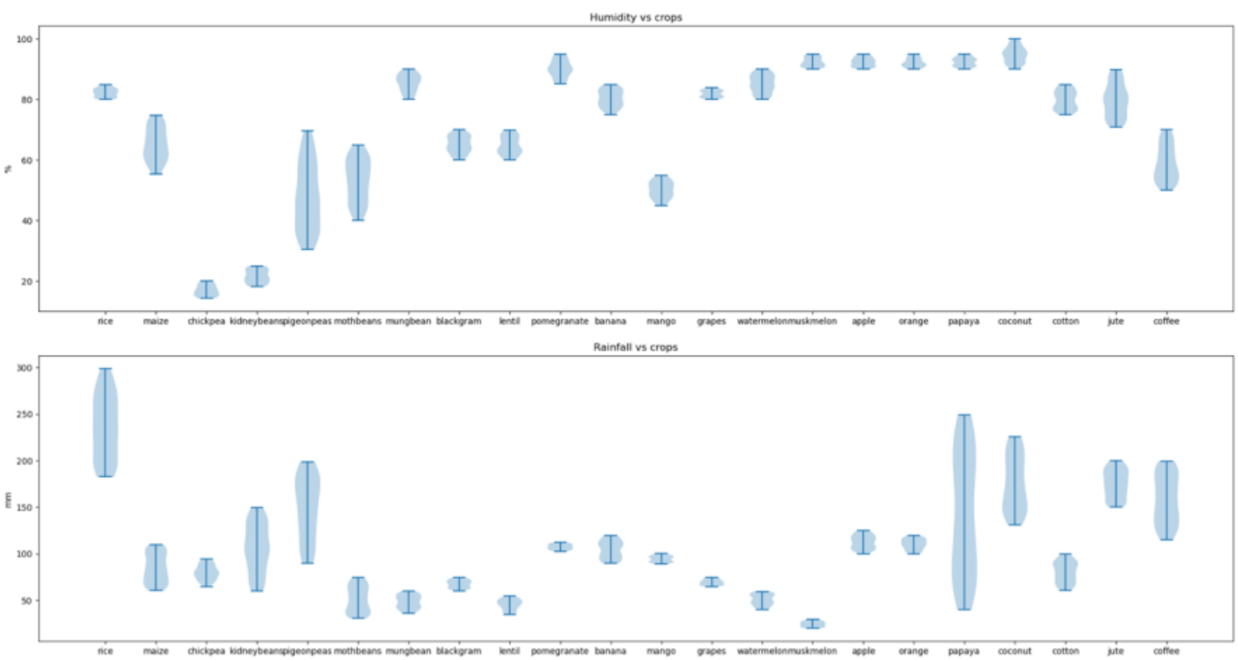
Figure 2: A Snapshot of Fertilizer dataset used

The GYGA dataset contained 348 rows containing 5 crops, each with yield obtained in five years. The crops contained in the dataset were maize, rice, millet, sorghum, wheat, and sugarcane.

STATIONN.	LONGITUD	LATITUDE	ELEVATION	COUNTRY	CROP	YA	YW	YW-YA	YP	YP-YA	WPP	WPA	CROPPING
Bobo-Diou	-4.317	11.167	445	Burkina Fa	Rainfed mi	0.822404	5.409494	4.58709	5.660274	4.83787	14.22598	2.162773	1
Bogandé	-0.137	12.974	281	Burkina Fa	Rainfed mi	0.68401	2.990438	2.306428	4.190429	3.506418	9.713214	2.221728	1
Boromo	-2.933	11.75	243	Burkina Fa	Rainfed mi	0.953076	5.036128	4.083051	5.442717	4.489641	13.25604	2.508678	1
Dori	-0.0333	14.0333	288	Burkina Fa	Rainfed mi	0.668488	1.91642	1.247932	5.04029	4.371802	6.462804	2.254364	1
Dédougou	-3.483	12.467	299	Burkina Fa	Rainfed mi	0.992483	3.202434	2.20995	5.188304	4.19582	8.796	2.726015	1
Fada Ngou	0.367	12.033	294	Burkina Fa	Rainfed mi	0.765633	3.436115	2.670482	4.301383	3.53575	11.23046	2.502366	1
Gaoua	-3.183	10.333	339	Burkina Fa	Rainfed mi	0.637542	5.072043	4.434501	5.110987	4.473446	14.53788	1.827372	1
Ouahigouy	-2.417	13.567	315	Burkina Fa	Rainfed mi	0.818752	3.161314	2.342561	4.313675	3.494923	9.840917	2.54871	1
Pô	-1.15	11.15	322	Burkina Fa	Rainfed mi	0.983713	4.441649	3.457937	5.553187	4.569474	12.32117	2.728827	1
bur_rfmt1	-1.72896	14.02031	308	Burkina Fa	Rainfed mi	0.584461	2.641314	2.056853	4.374257	3.789797	7.047733	1.559498	1
Adet	37.48	11.27	2240	Ethiopia	Rainfed mi	1.704167	5.460698	3.756531	6.022151	4.317984	10.53179	3.286745	1
Assosa	34.52	10.07	1575	Ethiopia	Rainfed mi	1.235667	6.326214	5.090547	6.342326	5.106659	13.36318	2.61016	1
Ayira	35.33	9.06	1700	Ethiopia	Rainfed mi	1.850833	5.950999	4.100166	6.084767	4.233934	12.23072	3.803901	1
Bahir Dar	37.38	11.58	1790	Ethiopia	Rainfed mi	1.924333	3.223575	1.299242	4.204535	2.280202	8.185381	4.886314	1
Gelemso	40.525	8.809	1810	Ethiopia	Rainfed mi	1.841833	5.17578	3.333946	7.056279	5.214446	11.00401	3.915845	1
Gondar	37.4715	12.59	1967	Ethiopia	Rainfed mi	2.289833	3.696909	1.407075	4.269884	1.98005	10.04922	6.224397	1
Gore	35.53	8.02	1880	Ethiopia	Rainfed mi	1.5175	5.743185	4.225685	5.746802	4.229302	14.22356	3.758238	1
Kobo	39.63	12.15	1500	Ethiopia	Rainfed mi	1.5174	1.165814	0	2.834826	1.317426	3.558881	4.632168	1
Melkassa	39.33	8.4	1550	Ethiopia	Rainfed mi	1.9734	1.901179	0	5.469709	3.496309	5.384526	5.58907	1
Nekemte	36.54	9.09	2110	Ethiopia	Rainfed mi	2.243667	7.197998	4.954331	7.240465	4.996798	15.19676	4.736937	1

Figure 3: A snapshot of the GYGA Dataset

Statistical analysis of the dataset revealed crops, such as rice and papaya to have a stable preference for a specific range of humidity levels while being sensitive to rainfall fluctuation. The study also further revealed temperature to be the most sensitive variable in determining crop recommendation.



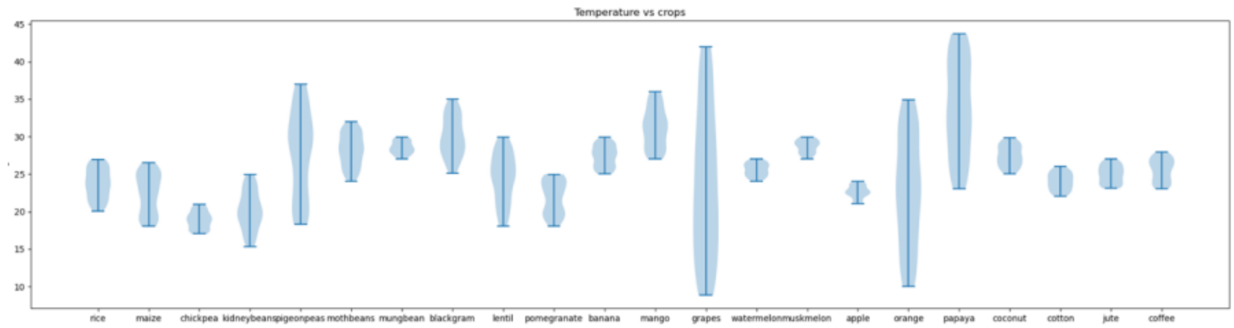


Figure 4: Violin Plots of environmental factors against crop requirements

On examining the relation between the NPK values, it was seen that all correlations between the nutrients except that between potassium and phosphorus were negative correlation indicating an antagonistic relation where a single nutrient affects how other nutrients are absorbed.

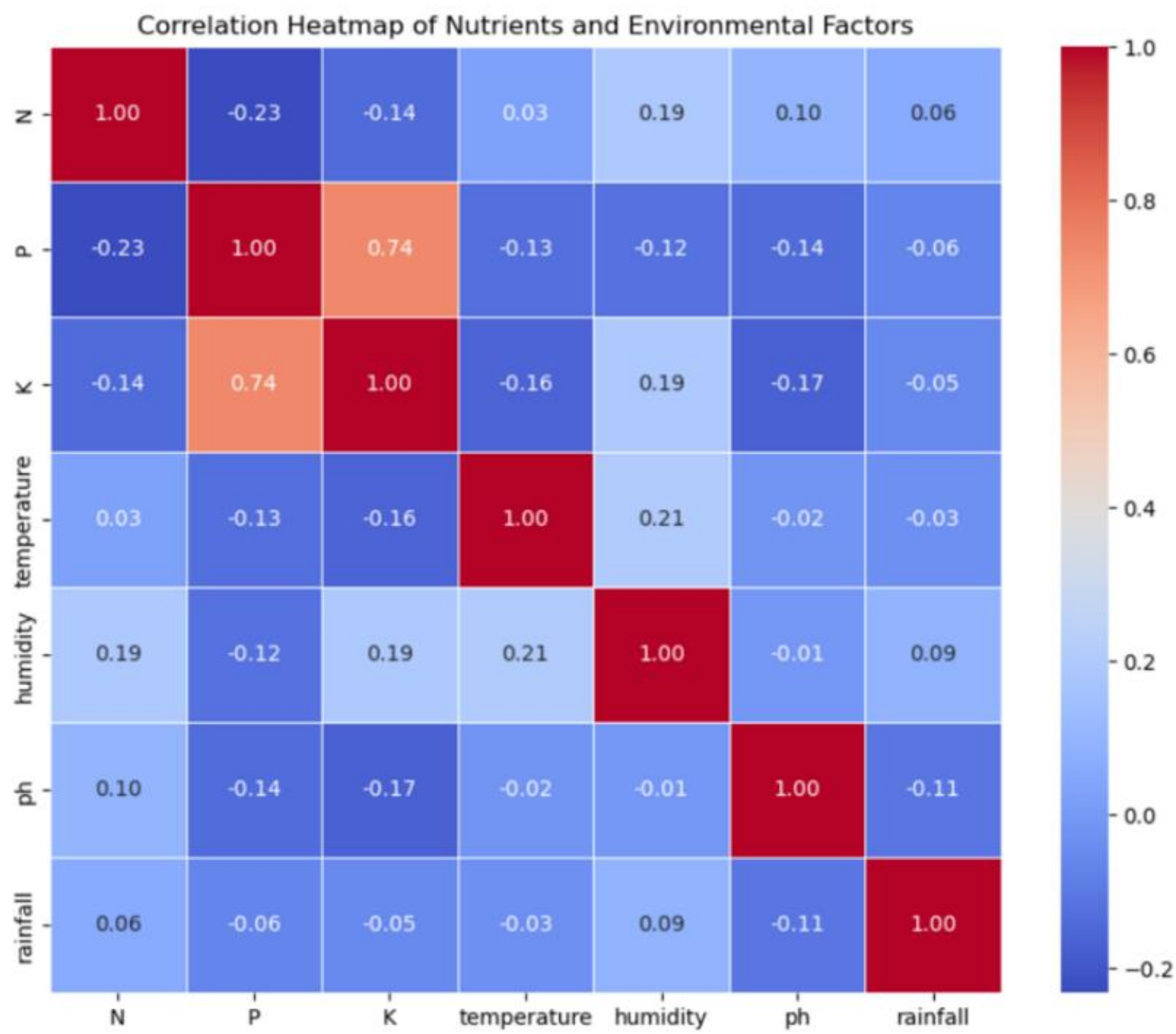


Figure 5: A heatmap showing the correlation between features

Model Selection, Evaluation and Training

In the crop recommendation model, the Random Forest Classifier was chosen for its effectiveness in handling non-linear data and its capability to perform both classification and regression tasks. This method is adept at addressing overfitting, particularly in datasets with high dimensionality

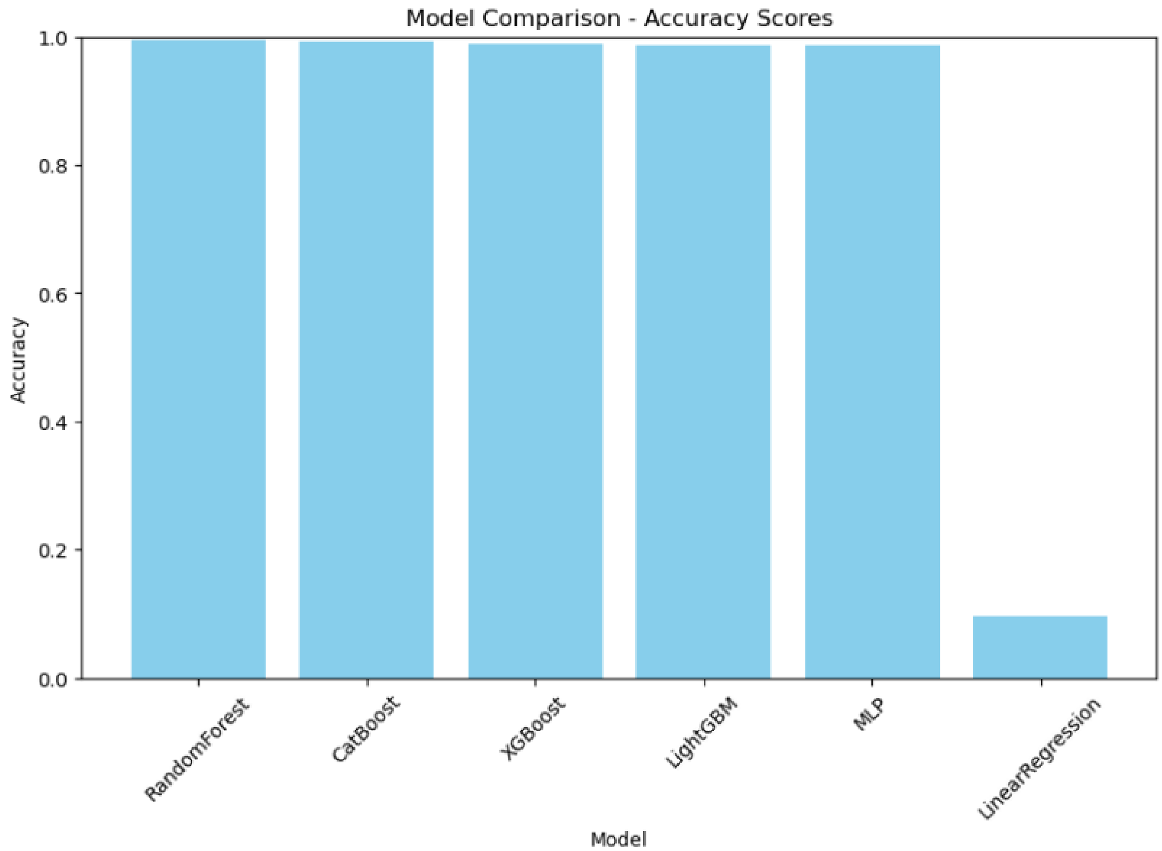


Figure 6: The comparison of models' accuracy.

On the other hand, CatBoost was chosen because in comparison to other models, it gave a higher prediction accuracy.

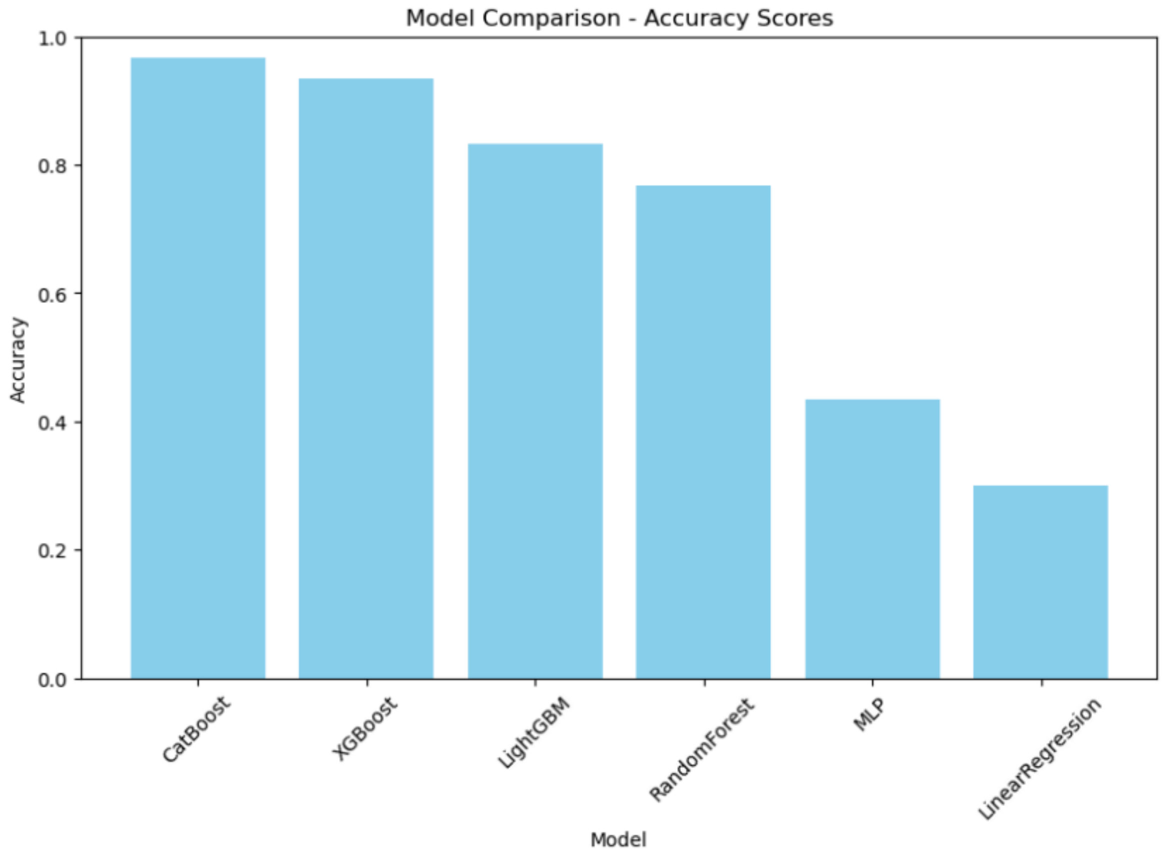


Figure 7: : The comparison of model accuracy for fertilizer.

On yield estimation, the Boosting algorithm CatBoost used for regression, had the best performance compared to the other models with a mean absolute error of 3.48 tonnes/hectare The model underwent cross-validation to assess its performance.

A 5-fold cross-validation technique was employed, testing the model's accuracy and stability across various dataset subsets. These metrics, including accuracy scores and cross-validation results, were used to measure the model's effectiveness in predicting crop types.

For fertilizer recommendation, the dataset was divided into five sections; four sections for training and the remaining for testing. To compensate for insufficient data this process was iterated five times and the model accuracy was computed averaging 0.716.

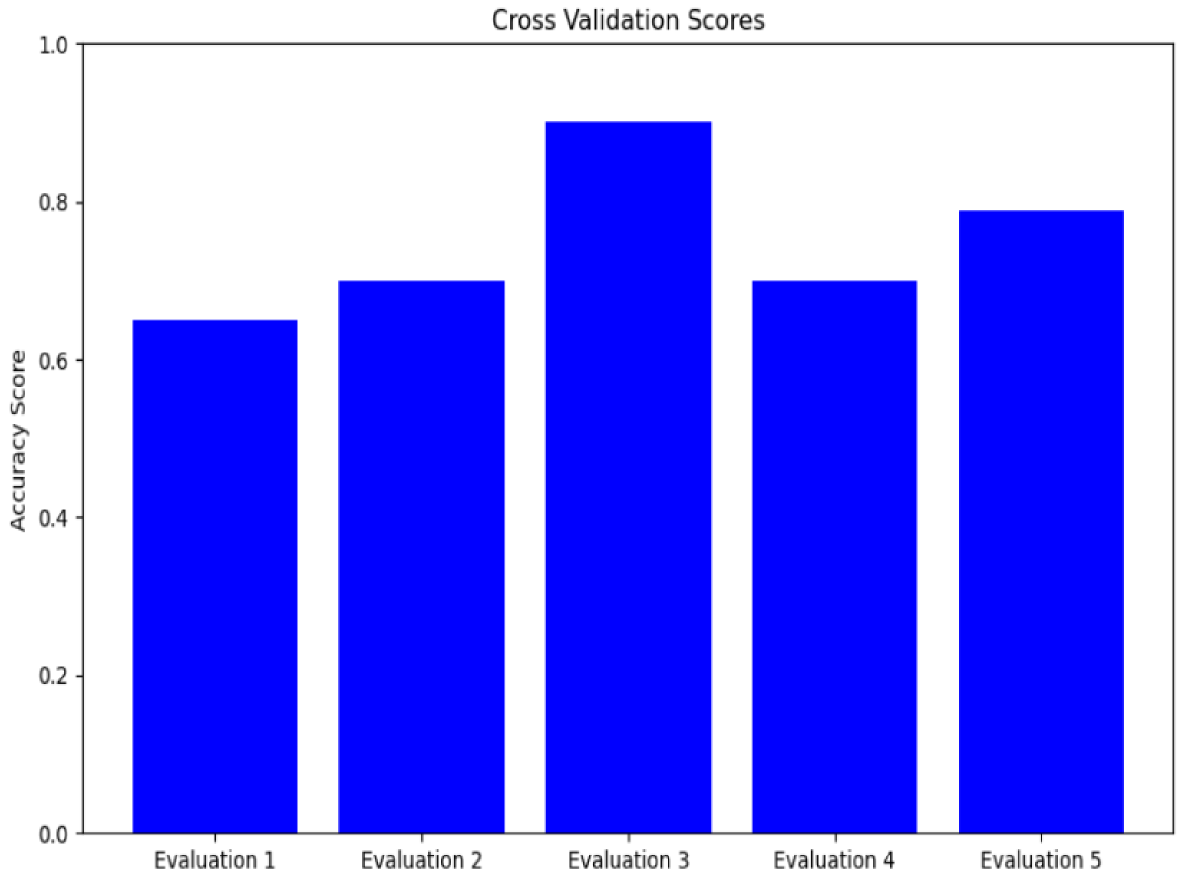


Figure 8: Cross-validation scores for fertilizer recommendation model.

In the assessment of the fertilizer recommendation model, a confusion matrix was drawn to explain the precision of the model in predicting a certain fertilizer.

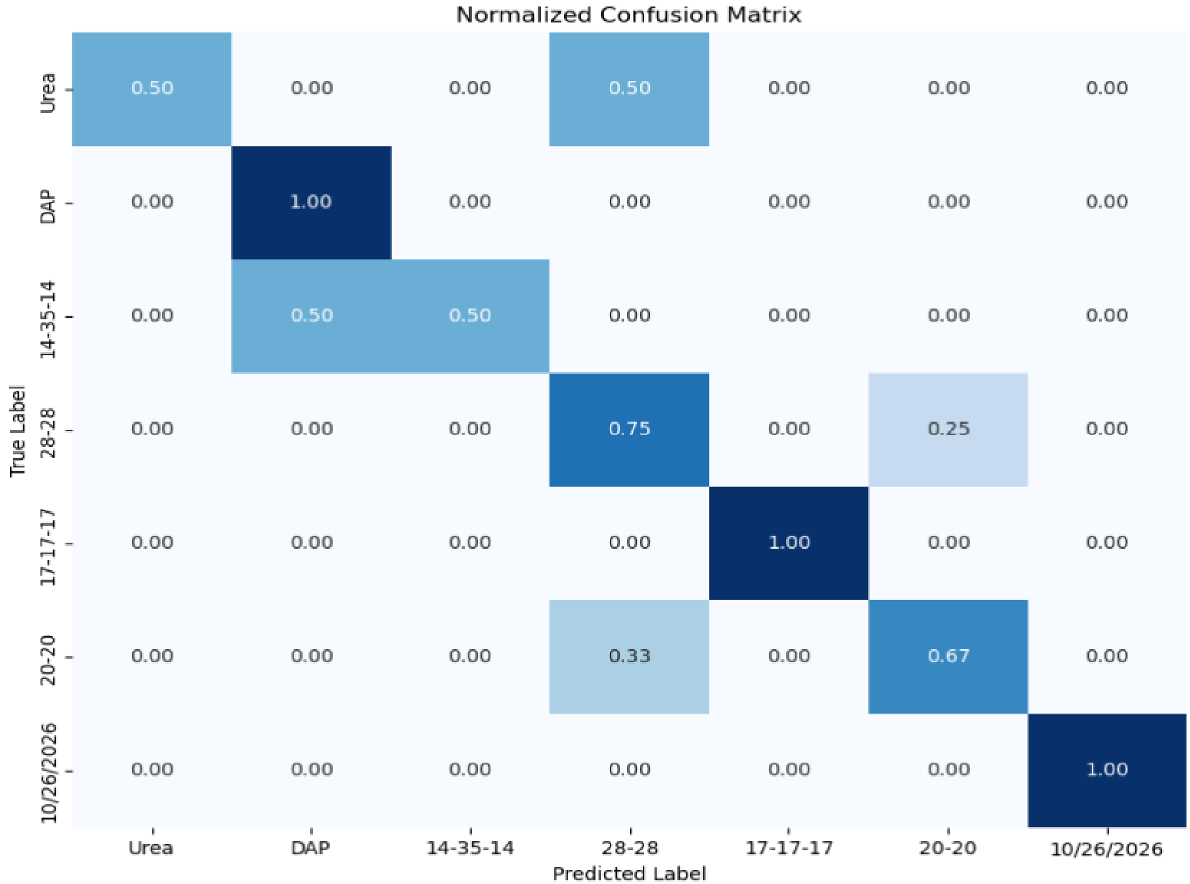


Figure 9: One of the confusion matrix for fertilizer recommendation model.