Crop Combination and Market Price Prediction

CS6611 Creative and Innovative Project

Submitted by ,

2022103525 – GURUMOORTHI R

2022103002 – MATHAN KUMAR P

2022103507 – SANJAY TG

2022103526 – SRI BALAJI J

Introduction

This report, part of the CS6611 Creative and Innovative Project titled 'Crop Combination and Market Price Prediction,' evaluates machine learning models for recommending main crops and their sub-crops. The main crop model uses a Random Forest Classifier, while the sub-crop model employs a Euclidean Distance-based approach (SubCropRecommender), akin to KNN, to rank sub-crops. The report compares these models against alternatives, presents detailed performance metrics, and addresses data challenges, particularly for sub-crop datasets, using the Crop Recommendation dataset and sub-crop datasets.

Methodology

The Crop Recommendation dataset (~2200 samples, 22 crops) was preprocessed with LabelEncoder and StandardScaler, split into 80% training and 20% testing sets for main crop prediction. Models compared include Logistic Regression, KNN, SVM (Linear and RBF), Decision Tree, Random Forest, Naive Bayes, and XGBoost. Metrics for main crop include accuracy, precision, recall, F1-score (macro-averaged), average confidence, low confidence rate (<0.7), and 5-fold cross-validation accuracy. For sub-crop recommendation, the SubCropRecommender uses Euclidean distance, compared with KNN and Random Forest on sub-crop datasets (30-200 samples) with a 30-sample and 3-sub-crop minimum threshold. Sub-crop metrics include top-3 accuracy, precision@3, recall@3, F1-score@3, MRR, NDCG@3, hit rate@3, average Euclidean distance, diversity, and coverage.

Data Challenges

Main crop prediction used a robust dataset (~2200 samples, 22 crops), ensuring reliable metrics. However, sub-crop recommendation faced significant challenges due to small dataset sizes (30-50 samples for most crops, 200 for Grapes but only 2 sub-crops) and missing files (e.g., Black Gram Dal). Many datasets were skipped due to insufficient samples (<30), too few unique sub-crops (<3), or file errors, leading to limited evaluation. These issues highlight the need for larger, standardized sub-crop datasets.

Results

The following sections present model comparisons and performance metrics for main crop prediction and sub-crop recommendation. Main crop results are robust, while sub-crop results are constrained by data limitations, as detailed in the Sub-Crop Dataset Summary.

Main Crop Model Comparison

Model	Accuracy (%)
Logistic Regression	96.67
KNN (k=5)	97.38
SVM (Linear)	97.62
SVM (RBF)	97.38
Decision Tree	98.57
Random Forest	99.05
Naive Bayes	99.05
XGBoost	98.57
Main Crop Model (Random Forest)	99.05

Sub-Crop Model Comparison

Model	Accuracy (%)
Euclidean Distance (SubCropRecommender)	97.91
KNN (k=5)	97.08
Random Forest	75.51

Main Crop Performance Metrics (Random Forest)

Metric	Value
Accuracy (%)	99.05
Precision (Macro) (%)	99.13
Recall (Macro) (%)	99.09
F1-Score (Macro) (%)	99.02
Average Confidence	0.96
Low Confidence Rate (%)	1.43
CV Accuracy (%)	97.86

Sub-Crop Performance Metrics (Euclidean Distance)

Metric	Value
Top-3 Accuracy (%)	97.91
Precision@3 (%)	49.25
Recall@3 (%)	97.91
F1-Score@3 (%)	65.54
Mean Reciprocal Rank	0.98
NDCG@3	1.26
Hit Rate@3 (%)	97.91
Average Euclidean Distance	0.02
Diversity	6.57
Coverage (%)	100.00

Sub-Crop Dataset Summary

Main Crop	Samples	Unique Sub-Crops	Status
Rice	500	5	Evaluated
Maize	300	3	Evaluated
Bengal Gram (Gram)(Whole)	400	4	Evaluated
Pegeon Pea (Arhar Fali)	500	5	Evaluated
Moath Dal	500	5	Evaluated
Green Gram (Moong)(Whole)	400	5	Evaluated
Black Gram Dal (Urd Dal)	500	5	Evaluated
Lentil (Masur)(Whole)	400	4	Evaluated
Pomegranate	500	5	Evaluated
Banana	500	5	Evaluated
Mango	500	5	Evaluated
Grapes	200	2	Skipped
Water Melon	500	5	Evaluated
Karbuja (Musk Melon)	300	3	Evaluated
Apple	500	5	Evaluated
Orange	300	3	Evaluated
Papaya	400	4	Evaluated
Coconut	500	5	Evaluated
Cotton	500	5	Evaluated
Jute	500	5	Evaluated
Coffee	400	4	Evaluated

Summary

The Main Crop Model (Random Forest) excelled in main crop prediction, with Random Forest achieving 99.05% accuracy. For sub-crop recommendation, Euclidean Distance (SubCropRecommender) led with 97.91% top-3 accuracy, though results were limited by small datasets (30-50 samples) and missing files, as shown in the Sub-Crop Dataset Summary. The Random Forest model demonstrated robust performance across metrics, while the SubCropRecommender's Euclidean Distance approach requires enhanced data for reliable evaluation.

Discussion

The Random Forest model and XGBoost outperformed other models in main crop prediction, leveraging ensemble techniques to capture complex feature interactions, with accuracies above 98%. The SubCropRecommender's Euclidean Distance approach, akin to KNN, showed potential but was hindered by small datasets, resulting in limited or zero metrics for most crops. KNN and Random Forest for sub-crops also faced data constraints, emphasizing the need for larger, standardized sub-crop datasets. Cross-validation and confidence metrics confirm the main crop model's reliability, while sub-crop metrics like NDCG@3 and diversity highlight ranking quality when data is sufficient.

Conclusion

This report, part of the CS6611 project, validates the Random Forest model for main crop prediction and evaluates the SubCropRecommender for sub-crop recommendation. While main crop prediction is highly accurate, sub-crop recommendation requires improved datasets to achieve reliable performance. Future work will integrate price prediction using market data (e.g., from Agmarknet) and expand sub-crop datasets to enhance agricultural decision-making