УДК 004.9+631

А. В. Марущак
В. Б. Мокін

РОЗРОБЛЕННЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ ВРОЖАЮ СІЛЬСЬКОГОСПОДАРСЬКИХ КУЛЬТУР

Вінницький національний технічний університет

Анотація

Проведено розвідувальний аналіз Kaggle-датасету „Crop Production" з врожайності сільськогосподарських культур. Для вирішення поставленої задачі побудовано інтелектуальні моделі передбачення даних XGBoostRegressor, DecisionTreeRegressor та RandomForestRegressor. Оптимальною є XGBoostRegressor. Розроблено та випробувано інформаційну технологію передбачення врожайності сільськогосподарських культур за рахунок удосконалення методів машинного навчання та розвідувального аналізу, що дозволяє підвищити точність цього передбачення.

Ключові слова: Python, розвідувальний аналіз, машинне навчання, сільськогосподарська сфера, врожайність.

Abstracts

An exploratory analysis of the Kaggle dataset „Crop Production" on crop yields was carried out. To solve this problem, we created data intelligent prediction models: XGBoostRegressor, DecisionTreeRegressor and RandomForestRegressor. The optimal is XGBoostRegressor. The information technology for predicting crop yields by improving machine learning and intelligence analysis methods is developed and tested. It allows for an increase in the accuracy of this prediction.

Keywords: *Python, intelligence analysis, machine learning, agricultural sector, yield.*

Вступ

Сільське господарство – це практика вирощування природних ресурсів для підтримки життя людини та забезпечення економічної стабільності. В цьому процесі аналітичні навички та використання інформаційних технологій, зокрема, для аналізу великих обсягів даних, моделювання та передбачення врожайності сільськогосподарських культур. Застосування сучасних методів, таких як машинне навчання та аналіз даних, допомагає сільськогосподарським підприємствам передбачувати майбутні складнощі та заздалегідь готовитися до цього.

Використовуючи методи машинного навчання можна враховувати різні фактори, такі як якість ґрунту, погодні умови, сорти культур, вплив шкідників, тощо. Такі дані допоможуть в подальшому, підприємствам готуватися до несподіваних викликів та оптимізувати об'єми затрачених ресурсів.

Метою дослідження є підвищення точності передбачення врожайності сільськогосподарських культур за рахунок використання методів машинного навчання.

Розвідувальний аналіз даних

Дослідження проводилось з датасетом Kaggle Dataset „Crop Production" (https://www.kaggle.com/datasets/imtkaggleteam/crop-production). Було проведено розвідувальний аналіз, на основі якого буде виконано передбачення врожайності зернових культур на території Європи.

Вхідні дані відфільтровано за такими параметрами, як: зернова культура – пшениця, ключова ознака – врожайність. Виконано перевірку між параметрами значень по роках. На рисунку 1 наведено Heatmap, де видно, що кореляції даних є незначною від 1961 по 1991 роки. Велику зміну дана сфера зазнала після 2000-х років, спостерігається стрімке зростання врожайності, яке поступово зростає щороку.

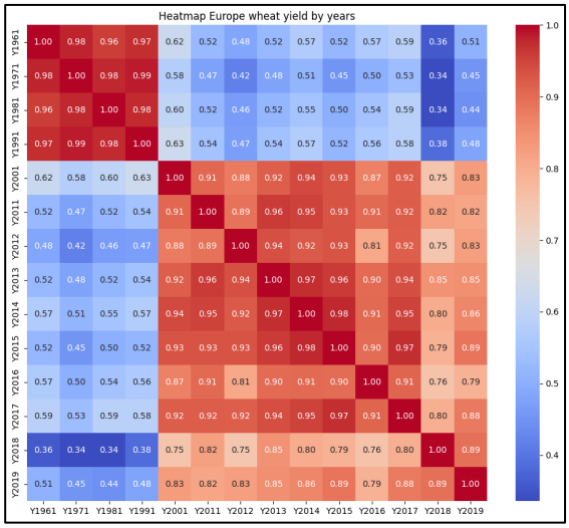

Рисунок 1. Heatmap врожайності пшениці у Європі

На рисунку 2 зображено відношення посівних площ соняшника та пшениці за 2019 рік на території Європи. На діаграмі чітко видно межу посівних площ соняшника та пшениці території України порівняно з усією Європою. Отже, Україна забезпечувала більшу частину сировини на глобальному ринку.

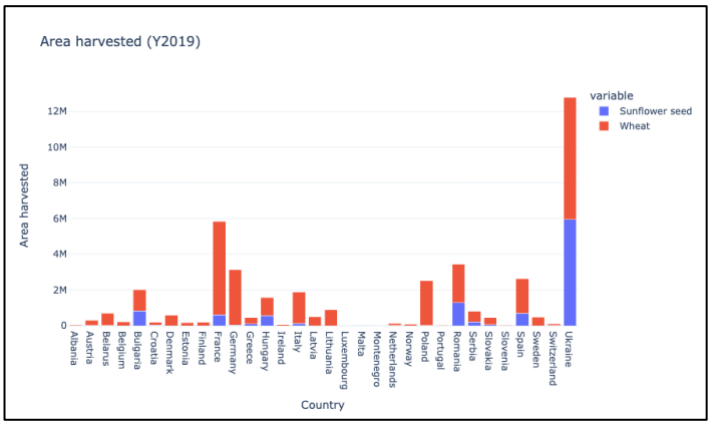

Рисунок 2. Посівні площі соняшника та пшениці за 2019 рік

Після розвідувального аналізу вхідних даних, було визначено, що для вирішення мети роботи потрібно використовувати такі методи машинного навчання, як [3]: eXtremeGradientBoosting (XGBoost), DecisionTreeRegressor, RandomForestRegressor.

Результати дослідження

Було розроблено ноутбук на Python, з використанням платформи Kaggle, який будує та застосовує використовує вищеописані моделі для передбачення даних для покращення точності врожаю сільськогосподарських культур. На рисунку 3 показано формування вхідних даних на навчальну та тестову вибірки.

```
# Вибір ознак (features) та цільової змінної (target)
features = ['Area Code','Item Code', 'Y2017', 'Y2018']
target = 'Y2019'
# Розділення даних на ознаки та цільову змінну
train_all = df[features]
target_all = df[target]
# Розділення на навчальні та тестові дані (80% на навчання, 20% на тестування)
train, valid, target_train, target_valid = train_test_split(train_all, target_all, test_size=0.2, random_state=42)
# Виведення результатів
print("Навчальні дані (X_train):")
print(train.info())
#print("\nНавчальні мітки (target_train):")
#print(target_train.info())
print("\nТестові дані (valid):")
print(valid.info())
#print("\nТестові мітки (target_valid):")
#print(target_valid.info())
```

Рисунок 3. Посівні площі соняшника та пшениці за 2019 рік

Навчання моделей даних проводилося з використанням різного відсотку тестових даних, а саме: 10%, 20%, 30% та 40%, було отримано результати, які сформовано та візуалізовано в таблиці 1.

Таблиця 1 – Точність моделей з різним відсотком тестових даних

	10%	20%	30%	40%
XGBoostRegressor	95.91	95.94	96.09	95.28
DecisionTreeRegressor	98.72	93.86	84.53	97.47
RandomForestRegressor	96.31	93.86	95.5	91.7

Зведені дані у таблиці показують, що моделі були добре навчені, також підлаштовуються під різний відсоток даних і дають високі показники точності. Для найбільш ефективної моделі даних результат навчання було візуалізовано за допомогою кривих навчання. На яких можна оцінити стабільність моделі, а також виявити проблеми, пов'язані з перенавчанням або недодаванням. На рисунку 4 показано криві навчання моделі XGBoostRegressor.

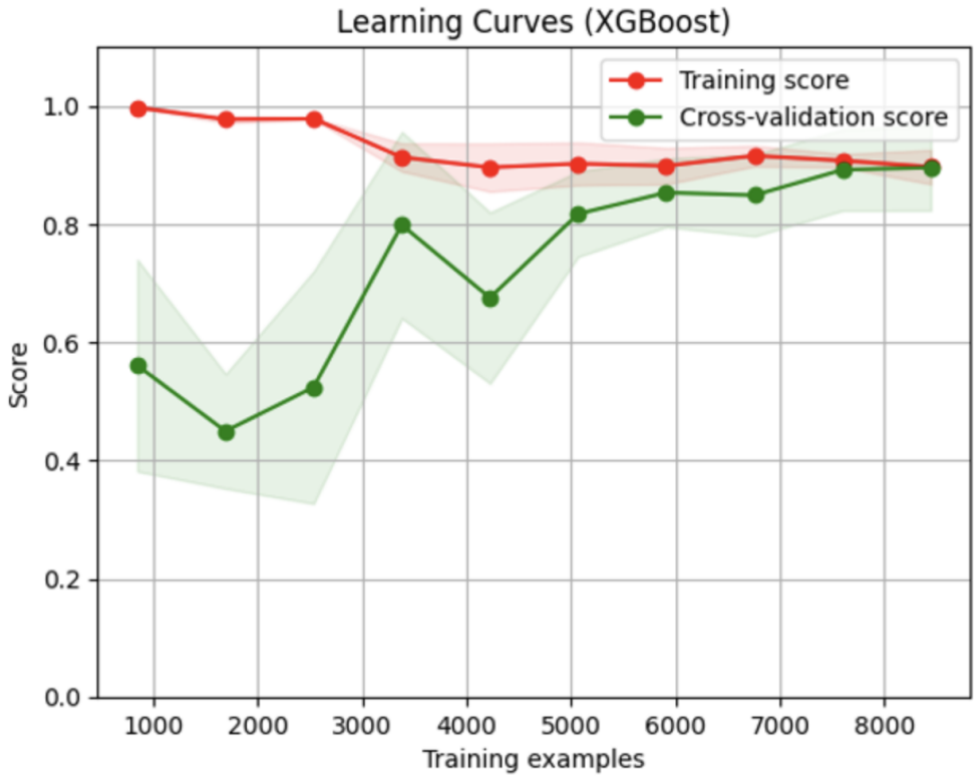

Рисунок 4. Криві навчання для моделі XGBoostRegressor

Висновки

Виконано налаштування системи розробки та проведено формування графіків з підготовлених

даних. Отримана візуалізація даних показала, що датасети є актуальними й містять важливу інформацію для виконання поставленої задачі в магістерській кваліфікаційній роботі.

Зроблено розвідувальний аналіз обраної тематики й датасету. Сформовано теплову карту масиву даних, отримано наглядну інформацію, яка показує відношення врожайності пшениці у країнах Європи по роках. Згенеровано додаткові графіки, для розширеного розуміння вхідних даних.

Запрограмувавши моделі та вхідні дані з обраного датасету, було виконано навчання для отримання передбачення врожаю зернових культур країн Європи. З яких оптимальною моделлю даних для виконання поставленої задачі є XGBoostRegressor – 95.94% за метрикою r2_score. Дослідження показало, що дана сфера застосування має високі перспективи в практичному використанні отриманих результатів та потребує подальшого дослідження в цьому напрямку.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Robert GRAYBOSCH, Encyclopedia of Food Grains (Second Edition). 2016. [Електронний ресурс] – Режим доступу до ресурсу: https://www.sciencedirect.com/science/article/abs/pii/B9780123944375000012.

2. Momeni MOHAMADREZA, Crop Production. 2024. [Електронний ресурс] – Режим доступу до ресурсу: https://www.kaggle.com/code/artem1018/crop-production-data-analyze-and-prediction-ml/input.

3. Наука про дані: машинне навчання та інтелектуальний аналіз даних : електронний навчальний посібник комбінованого (локального та мережевого) використання [Електронний ресурс] / В. Б. Мокін, М. В. Дратований – Вінниця : ВНТУ, 2024. – 258 с. – Режим доступу: https://docs.vntu.edu.ua/card.php?id=8163

Марущак Артем Володимирович — студент групи 1ІСТ-23м, Вінницький національний технічний університет, Вінниця, maruskhak@gmail.com

Мокін Віталій Борисович – д-р техн. наук, проф., завідувач кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: vbmokin@vntu.edu.ua

Marushchak Artem Volodymyrovych *– student of group 1IST-23m, Vinnytsia National Technical University, Vinnytsia, maruskhak@gmail.com*

Mokin Vitalii B. **–** Dr. Tech. Sciences, Prof., Head of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: vbmokin@vntu.edu.ua