Weather vs Yield – Prognozowanie plonów rolnych na podstawie warunków pogodowych

1 WSTĘP

Rolnictwo od wieków pełniło istotną rolę w życiu człowieka, stanowiąc fundament bezpieczeństwa żywnościowego i rozwoju cywilizacji. Współczesne wyzwania, takie jak zmiany klimatyczne, niestabilność pogodowa oraz rosnące zapotrzebowanie na żywność, stawiają przed rolnictwem nowe wymagania. W odpowiedzi na te wyzwania, coraz częściej wykorzystuje się zaawansowane technologie, w tym uczenie maszynowe, do optymalizacji produkcji rolniczej i przewidywania plonów.

Projekt *Weather vs Yield* ma na celu stworzenie modelu regresyjnego, który pozwoli na prognozowanie wysokości plonów (t/ha) na podstawie danych pogodowych i agrarnych, takich jak opady, temperatura, nawadnianie, rodzaj gleby i uprawy. Takie podejście może wspierać rolników i decydentów w podejmowaniu świadomych decyzji dotyczących upraw, planowania zasobów oraz adaptacji do zmieniających się warunków klimatycznych.

2 DANE

Do realizacji projektu wykorzystano zbiór danych pochodzący z platformy *Kaggle* [1], zawierający około 100 000 rekordów z informacjami dotyczącymi:

- średnich rocznych opadów,
- średnich rocznych temperatur,
- nawadniania,
- rodzaju gleby,
- typu uprawy,
- wysokości plonów (t/ha).

Dane zostały podzielone na zestawy treningowe i testowe oraz poddane wstępnej analizie eksploracyjnej w celu identyfikacji zależności między zmiennymi. Wykorzystano również techniki inżynierii cech, takie jak kodowanie zmiennych kategorycznych i normalizacja danych numerycznych.

3 METODOLOGIA

W projekcie wykorzystano algorytm regresyjny jakim jest *Random Forest* (las losowy). Model został oceniony przy użyciu metryk takich jak:

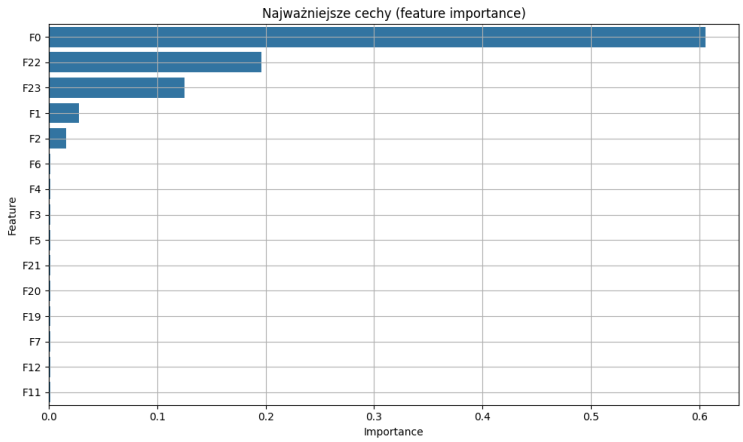
- średni błąd bezwzględny,
- pierwiastek z błędu średniokwadratowego,
- współczynnik determinacji.

Uzyskano następujące wyniki – średni błąd bezwzględny (MAE) wyniósł 0.42, pierwiastek z błędu średniokwadratowego osiągnął wynik 0.52, natomiast współczynnik determinacji (R2) wyniósł 0.91, co oznacza, że model bardzo dobrze wyjaśnia zmienność plonów w danych testowych.

4 WYNIKI

Analiza ważności cech (*feature importance*) wykazała, że największy wpływ na wysokość plonów miały [Rys. 1]:

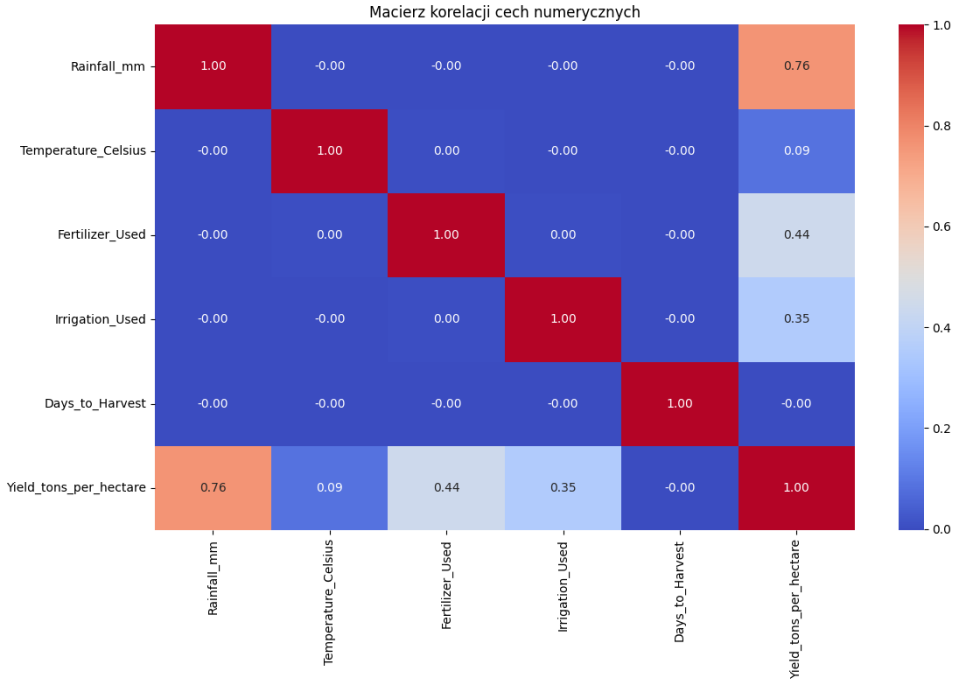
- F0 (średnia opadów deszczu w mm)
- F22 (użycie nawozu)
- F23 (nawadnianie)
- F1 (średnia temperatur)
- F2 (czas do zbiorów)



Rys. 1 Wykres najważniejszych cech

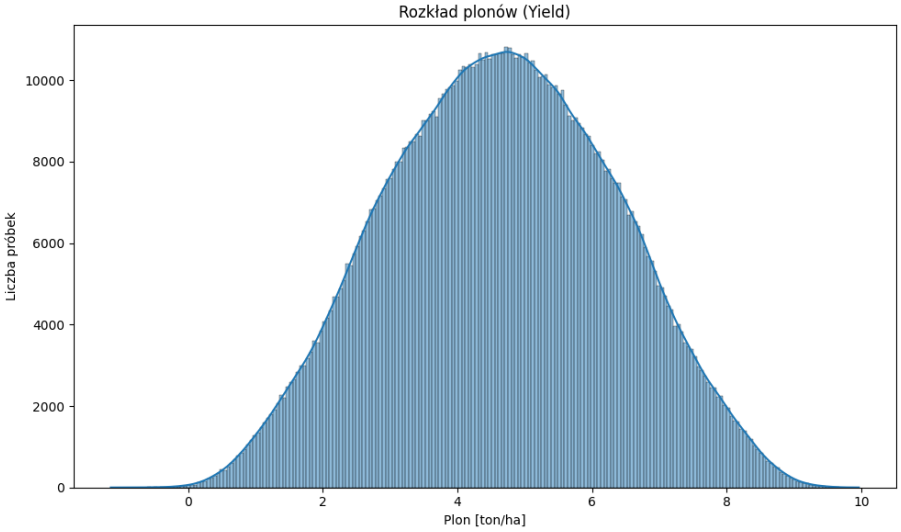
Uzyskana w *Eksploracji Danych* macierz korelacji pokazuje, że [Rys. 2]:

- najwyższą dodatnią korelacje uzyskała średnia opadów deszczu – wartość 0.76,
- stosowanie nawozów uzyskało umiarkowanie dodatnią korelacje – 0.44,
- słabsza, chociaż wciąż pozytywna korelacja to nawadnianie – wartość 0.35,
- znikomą korelacje uzyskała cecha opisująca średnią temperatur – 0.09,
- natomiast zerową korelacje uzyskała cecha czas do zbiorów.



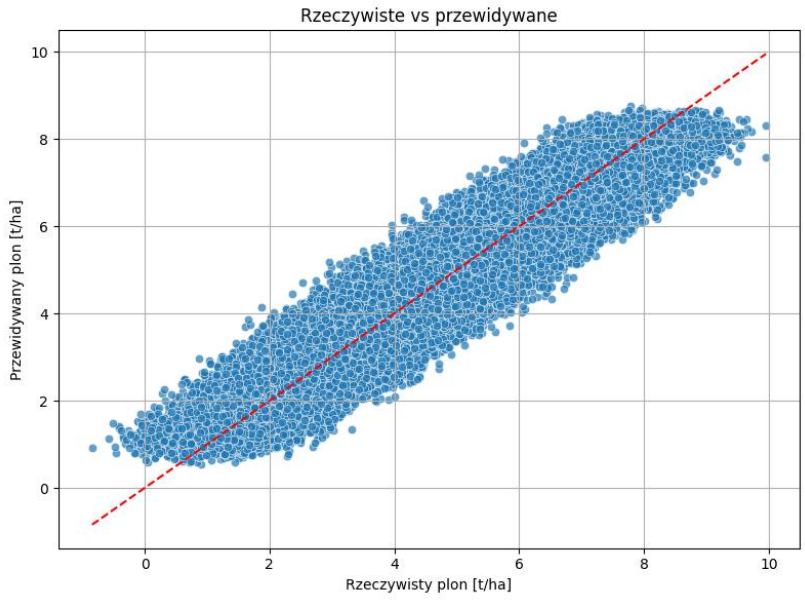
Rys. 2 Macierz korelacji numerycznych

W *Eksploracji Danych* uzyskano również wykres rozkładu zmiennej celu, w przypadku tego projektu jest to wartość zbiorów (tony na hektar pola). Rozkład zmiennej celu przypomina rozkład normalny (rozkład Gaussa) [Rys. 3], większość wartości znajduje się w przedziale 3–7 ton/ha, a szczyt rozkładu występuje w okolicach 5 ton/ha. Takie ukształtowanie rozkładu sprzyja stabilnemu działaniu modeli regresyjnych, ponieważ eliminuje problem asymetrii i wartości odstających.



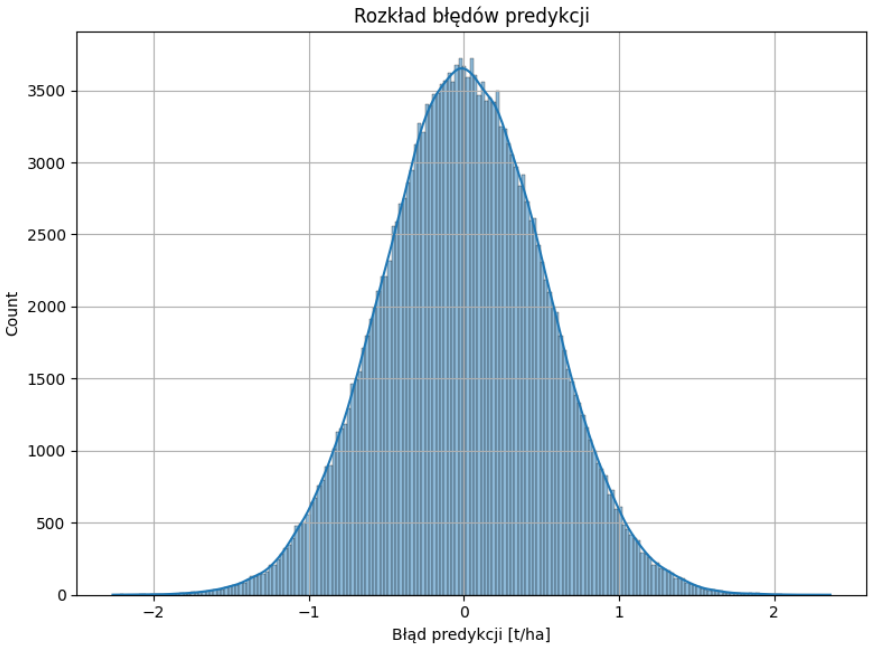
Rys. 3 Wykres rozkładu plonów z surowego zbioru danych

W ewaluacji wytrenowanego oraz utworzonego modelu predykcji, uzyskano wykres „rzeczywistych vs przewidywanych" wartości końcowych [Rys. 4]. Wykres nie przedstawia bardzo dużych rozbieżności względem osi symetrii, co może nam mówić, że model jest dosyć dokładny.



Rys. 4 Wykres "rzeczywiste vs przewidywane"

Ostatnim uzyskanym wykresem – jest wykres rozkładu błędu predykcji [Rys. 5], pokazuje on, że model uzyskał maksymalny błąd w predykcji wyniku końcowego do maksymalnie 2 ton na hektar. Wynik taki jest w akceptowalnej granicy błędu bazując na tym, że zbiór danych posiada 100 tys. próbek z całej kuli ziemskiej, gdzie powierzchnie pól rolnych są zróżnicowane (np. w USA mamy do czynienia z ogromnymi połaciami upraw, natomiast w Polsce średnia wielkość pól uprawnych wynosi około 11.54 ha [2]).



Rys. 5 Wykres rozkładu błędu predykcji

5 WNIOSKI

Projekt *Weather vs Yield* potwierdził, że zastosowanie metod uczenia maszynowego w rolnictwie może znacząco poprawić dokładność prognozowania plonów, co jest kluczowe w kontekście zmieniających się warunków klimatycznych i rosnącego zapotrzebowania na żywność.

Model Random Forest okazał się efektywny w przewidywaniu plonów na podstawie danych pogodowych i agrarnych. W przyszłości warto rozważyć integrację dodatkowych źródeł danych, takich jak dane satelitarne czy informacje o praktykach rolniczych, aby zwiększyć dokładność prognoz i wspierać zrównoważone zarządzanie zasobami rolnymi.

6 ŹRÓDŁA

[1] https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield/data (stan na 23.05.2025 r.)

[2] https://www.gov.pl/web/arimr/srednia-powierzchnia-gruntow-rolnych-w-gospodarstwie-w-2024-roku (stan na 23.05.2025 r.)