System Predykcji Wynikow meczow NBA

Zastosowanie technik uczenia maszynowego do przewidywania spotkan

Grzegorz Alwasiak

Cel Projektu

- Stworzenie systemu predykcyjnego do przewidywania wyników meczów NBA
- Porównanie skuteczności różnych algorytmów uczenia maszynowego
- Identyfikacja kluczowych czynników wpływających na wyniki spotkań
- Analiza różnicy między predykcją (przed meczem) a klasyfikacją (z danymi z meczu)

Dane

- Źródło: Baza danych meczów NBA z lat 1946-2023 z Kaggle
- Liczba analizowanych meczów: 30,000
- Główne kategorie danych:
  - Wyniki meczów
  - Statystyki drużyn z poprzednich spotkań
  - Informacje o gospodarzach/gościach
- Cechy predykcyjne: 10 różnic statystycznych między drużynami

Przygotowanie danych

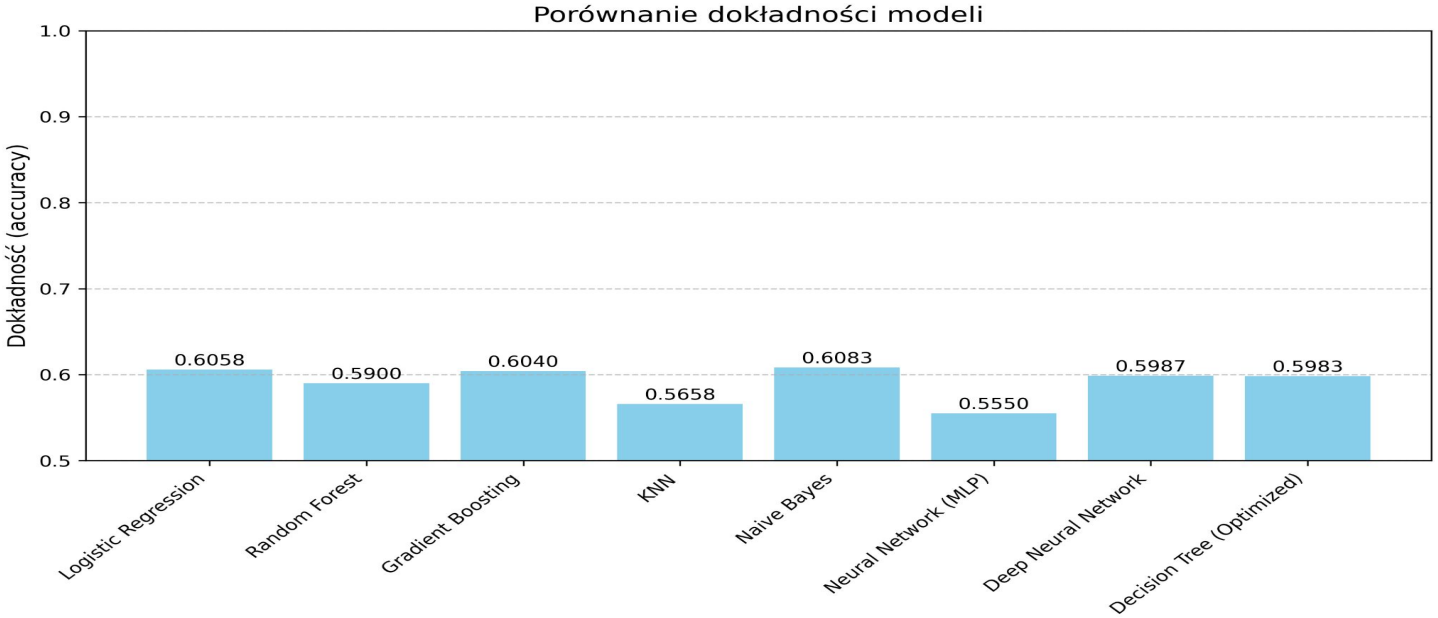
- Dla każdego meczu analizowano statystyki z 10 poprzednich spotkań drużyn
- Główne cechy predykcyjne:
  - `win_pct_diff`: Różnica w % zwycięstw między drużynami
  - `fg_pct_diff`: Różnica w skuteczności rzutów z gry
  - `home_adv`: Przewaga gospodarza
  - Różnice w zbiórkach, asystach, przechwytach, itd.
- Podział danych: 80% treningowe, 20% testowe (chronologicznie)
- Normalizacja i wartosci brakujace

Modele predykcyjne

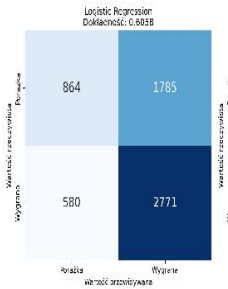
- Logistic Regression
- Random Forest
- Gradient Boosting
- KNN
- Naive Bayes
- Neural Network (MLP)
- Deep Neural Network
- Decision Tree

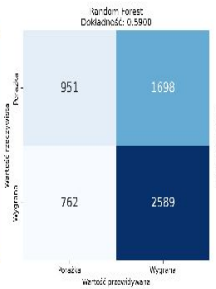
Porównanie dokładności modeli predykcyjnych

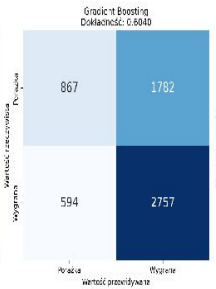
- **Najlepszy model:** Naive Bayes (60.90% dokładności)
- **Najgorszy model:** Neural Network MLP (55.50% dokładności)
- Modele tradycyjne (Naive Bayes, Logistic Regression) przewyższają bardziej złożone sieci neuronowe
- Dokładność ~61% oznacza znaczącą przewagę nad losowym zgadywaniem (50%)

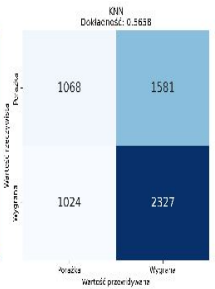


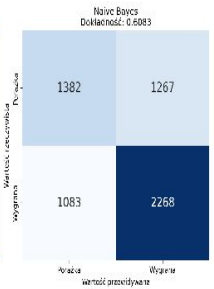
Macierze Błędów

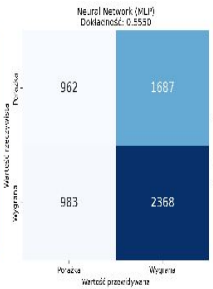


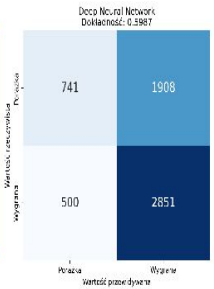


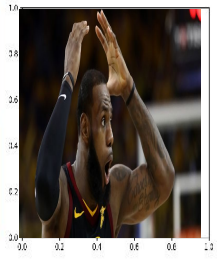












Konfiguracje i ewaluacja klasyfikatorów

Testowane konfiguracje parametrów:

- ***Random Forest:***
  - *liczba drzew: [50, 100, 200, 500]*
  - *max_depth: [5, 10, 15, None]*
  - *Najlepsza: n_estimators=200, max_depth=15 (dokładność: 59.00%)*
- ***Naive Bayes:***
  - *warianty: [Gaussian, Multinomial, Complement]*
  - *var_smoothing: [1e-9, 1e-8, 1e-7, 1e-6]*
  - *Najlepsza: GaussianNB, var_smoothing=1e-8 (dokładność: 60.90%)*
- ***MLP/Deep Neural Network:***
  - *warstwy: [(64,), (128,), (64,32,), (128,64,32,)]*
  - *funkcje aktywacji: [ReLU, tanh, sigmoid]*
  - *Najlepsza: (64,32,16), ReLU, batch_normalization (dokładność: 59.87%)*

Miary ewaluacji:

- ***Accuracy:*** *60.90% (Naive Bayes)*
- ***Precision:*** *0.65 (Naive Bayes), 0.66 (Gradient Boosting)*
- ***Recall:*** *0.73 (Naive Bayes), 0.62 (Random Forest)*
- ***F1-Score:*** *0.69 (Naive Bayes), 0.67 (Gradient Boosting)*
- ***AUC:*** *0.64 (Naive Bayes), 0.63 (Logistic Regression)*

Wnioski z konfiguracji:

- *Prostsze modele okazały się bardziej skuteczne dla tego problemu*
- *Zwiększanie złożoności (np. głębsze sieci) powodowało przeuczenie*
- *Naive Bayes najlepszy we wszystkich miarach poza Precision*

Porównanie modeli predykcyjnych i klasyfikacyjnych

- Deep Neural Network: **92.3%** ⭐
- Gradient Boosting: 91.9%
- Random Forest: 91.9%
- Logistic Regression: 91.5%
- Neural Network (MLP): 91.1%
- KNN: 90.0%
- Naive Bayes: 88.1%

Kluczowe obserwacje:

- Modele używające danych z meczu osiągają o ~31 p.p. wyższą dokładność
- Kolejność modeli jest odwrócona - sieci neuronowe najlepsze dla danych z meczu
- Złożone modele (DNN) lepiej wykorzystują bogate dane z przebiegu meczu
- Dowód na trudność prawdziwej predykcji vs łatwość "wyjaśniania" wyniku post-factum

Reguly Asocjacyjne

Kluczowe odkrycia:

- Znaleziono 7,185 reguł asocjacyjnych (lift > 1.0)
- Najsilniejsze reguły osiągają lift > 2.5 i confidence > 0.8

Wzorce w danych:

- **Przewaga gospodarzy** (home_adv_medium) pojawia się w 9/10 najsilniejszych reguł
- **Wyrównane statystyki** często prowadzą do zwycięstw gospodarzy
- Drużyny o podobnej skuteczności rzutów mają zwykle podobne bilanse zwycięstw

Związek asyst ze skutecznością rzutową

- `[win_pct_diff=medium, ast_diff=medium] => [fg_pct_diff=medium]`
- **Interpretacja**: Drużyny o podobnej formie i liczbie asyst mają również podobną skuteczność rzutową
- Lift: 2.21, Confidence: 73.6%

Przewaga własnego boiska

- `[home_win=win, home_adv=medium, fg_pct_diff=medium] => [win_pct_diff=medium]`
- **Interpretacja**: Gospodarze często wygrywają mimo podobnej formy i skuteczności rzutowej
- Lift: 2.54, Confidence: 83.6%

Wpływ otoczenia na wyniki

- `[home_win=win, home_adv=medium] => [win_pct_diff=medium]`
- **Interpretacja**: Sama przewaga własnego parkietu może równoważyć różnice w bilansach zwycięstw
- Lift: 2.18, Confidence: 71.8%, Support: 16.7% (najwyższe wsparcie ze wszystkich reguł)

Podsumowanie Projektu

- Najlepszy model predykcyjny: **Naive Bayes (60.9% dokładności)**
- Znacząca różnica między predykcją (61%) a klasyfikacją po meczu (92%)

Co działa:

- Przewidywanie wyników meczów NBA z dokładnością o ~11% lepszą niż losowe zgadywanie
- Analiza reguł asocjacyjnych potwierdza kluczową rolę przewagi gospodarzy

Co nie działa idealnie:

- Sieci neuronowe (MLP) osiągają najniższą dokładność (55.5%)
- Trudność w przewidywaniu niespodzianek i meczów z małą różnicą w formie drużyn
- Brak uwzględnienia kontuzji zawodników i innych czynników jakościowych

Najważniejsze wnioski:

1. Forma drużyny (`win_pct_diff`) i przewaga gospodarzy są najsilniejszymi predyktorami
2. Modele tradycyjne (Naive Bayes, Logistic Regression) przewyższają złożone algorytmy
3. Istnieje naturalna granica dokładności predykcji sportowej ze względu na losowy charakter sportu
4. Różnica między predykcją a klasyfikacją (61% vs 92%) pokazuje, jak duży wpływ mają wydarzenia podczas meczu

Koniec