Prezentacja do Projektu

Eksploracja i wizualizacja danych

Michał Brodacki, s32038

Polsko–Japońska Akademia Technik Komputerowych

16 Grudnia 2023

Spis Treści

Cel Projektu

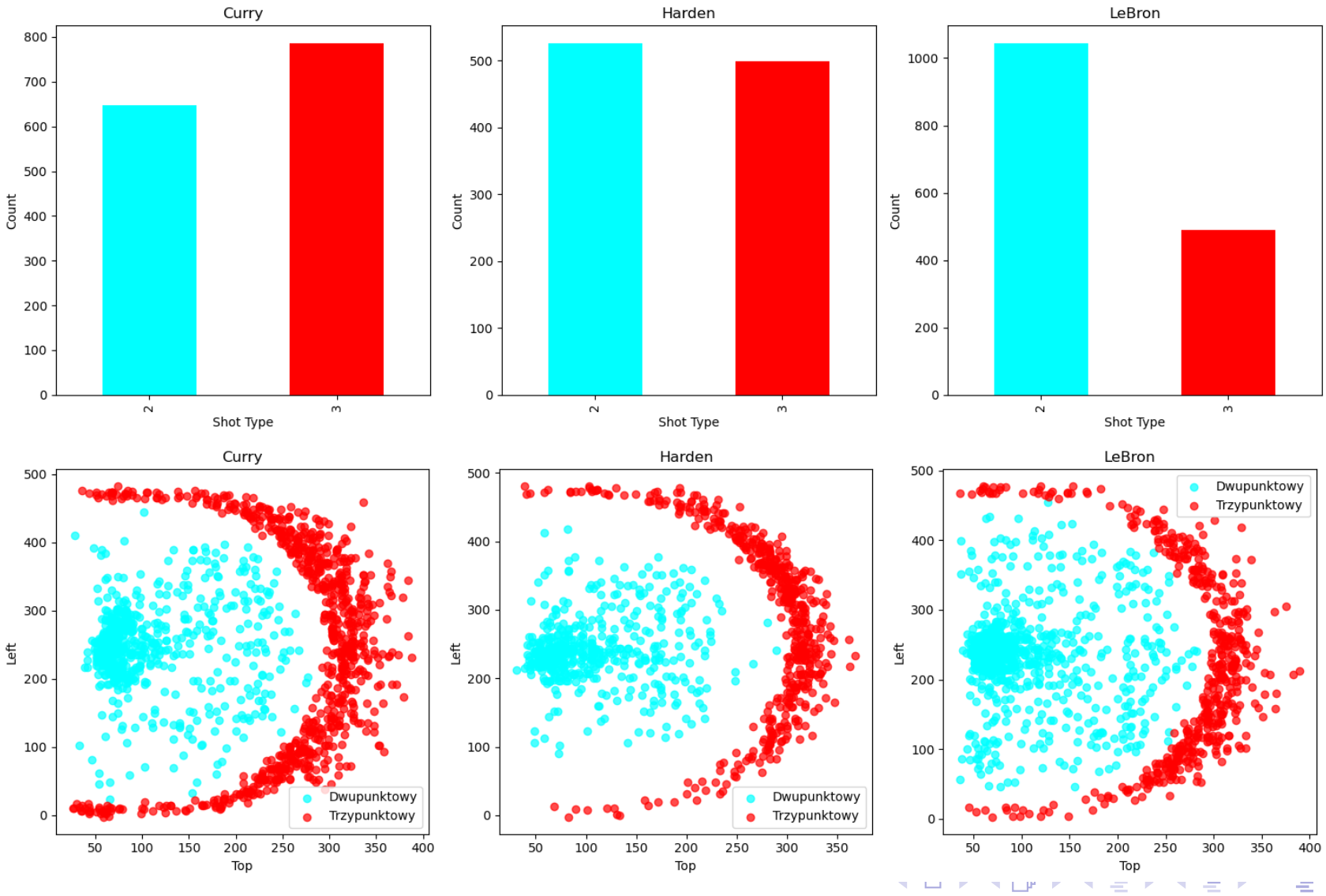
Celem projektu jest rozpoznanie na podstawie danych pozycyjnych koszykarza oddającego rzut, czy będzie on liczony jako 3–punktowy, czy jako 2–punktowy.

Dane

Do projektu wykorzystam dane NBA 2023 Player Shot Dataset dostępne pod linkiem: *https://www.kaggle.com/datasets/dhavalrupapara/nba-2023-player-shot-dataset/?select=2_james_harden_shot_chart_2023.csv*. Zawierają one informacje na temat sytuacji rzutowych trzech koszykarzy występujących w **NBA**.

	top	left	date	qtr	time_remaining	result	shot_type	distance_ft	lead	lebron_team_score	opponent_team_score	opponent	team	season	color
0	310	203	Oct 18, 2022	1st Qtr	09:26	False	3	26	False	2	2	GSW	LAL	2023	red
1	213	259	Oct 18, 2022	1st Qtr	08:38	False	2	16	False	4	5	GSW	LAL	2023	red
2	143	171	Oct 18, 2022	1st Qtr	08:10	False	2	11	False	4	7	GSW	LAL	2023	red
3	68	215	Oct 18, 2022	1st Qtr	05:24	True	2	3	False	12	19	GSW	LAL	2023	green
4	66	470	Oct 18, 2022	1st Qtr	01:02	False	3	23	False	22	23	GSW	LAL	2023	red

Wstępna ocena danych

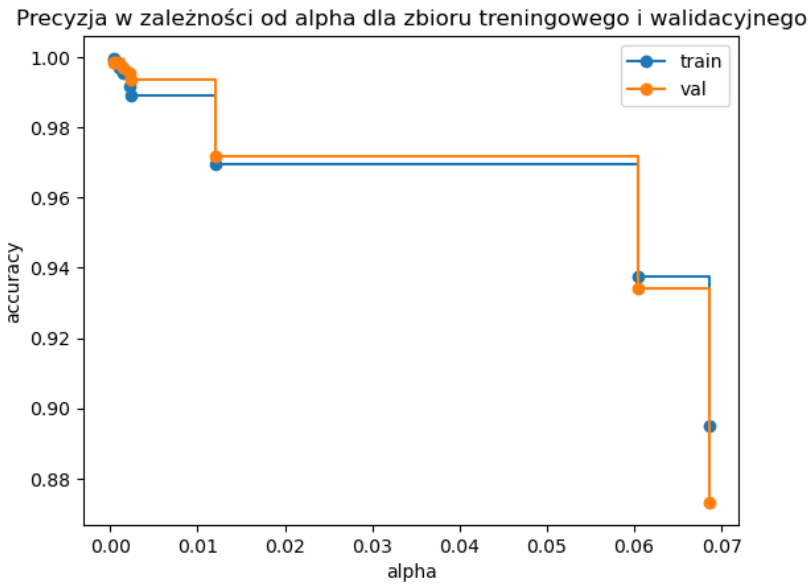


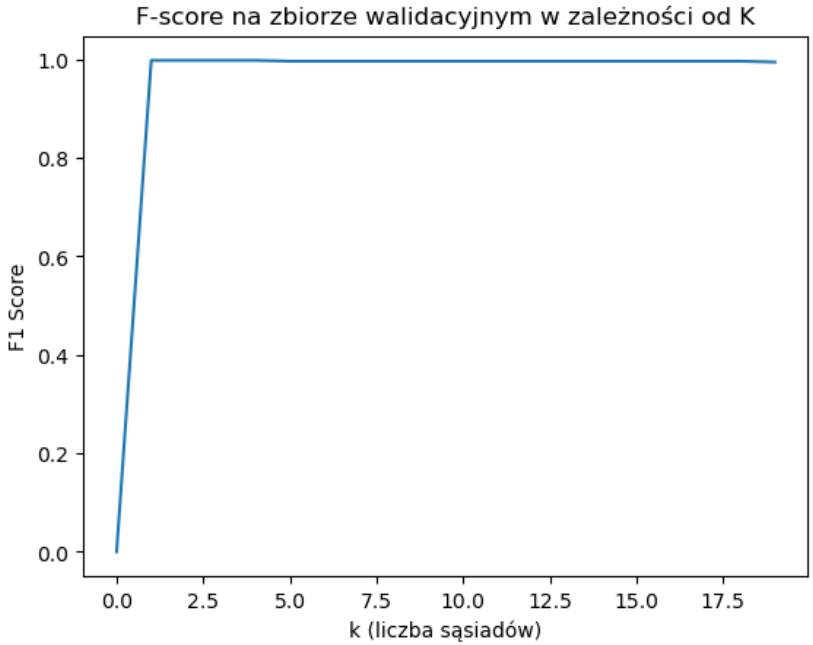
Przygotowanie Danych

Na tym etapie wykonane zostały następujące czynności:

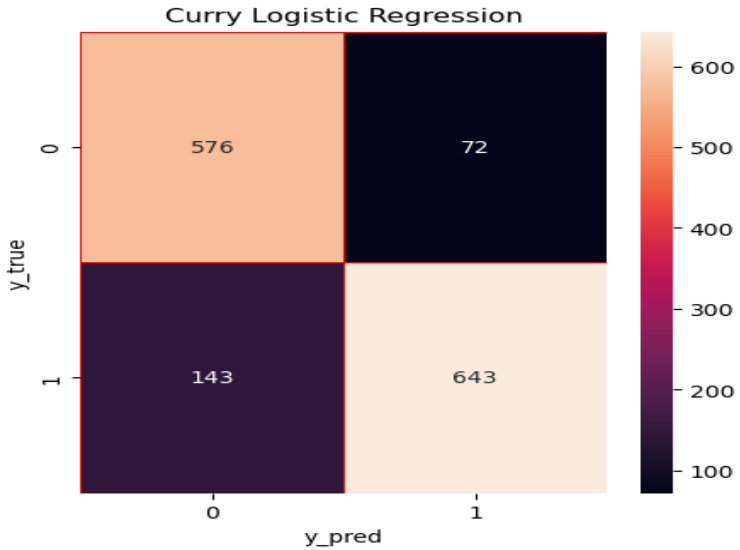
- Sprawdzono czy zbiór danych zawiera brakujące wartości — nie zawiera
- Zbadano kodowanie zmiennej, która będzie później estymowana *Curry["shot_type"].dtype* i wyszło *dtype('int64')*.
- Na potrzeby klasyfikacji stworzono nową zmiennę *is_three* , którą zastąpiono zmienną *shot_type* (zmienna typu Bool).
- Podzielono zbiór na częsci: treningową, walidacyjną i testową na dwa sposoby:
  1. W pierwszej wersji rzuty jednego zawodnika stanowiły zbiór treningowy, drugiego walidacyjny, a trzeciego testowy.
  2. W drugiej złączono wszystkie *dataframe'y* i podzielono na trzy zbiory całość.

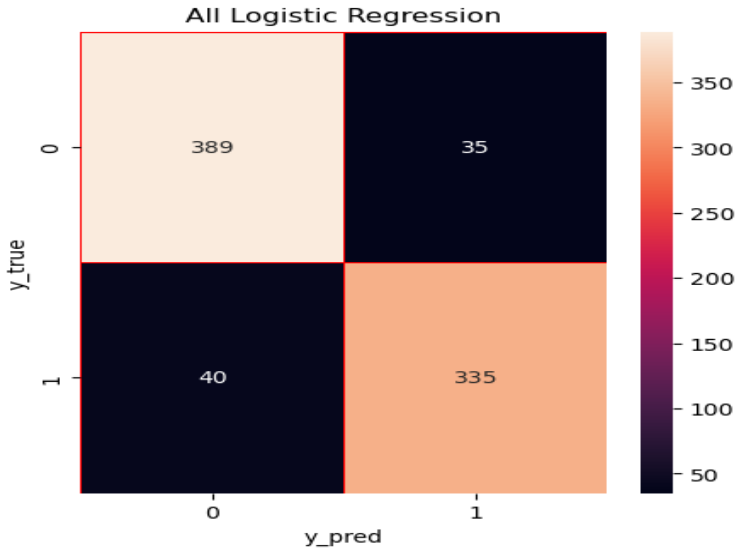
Modelowanie

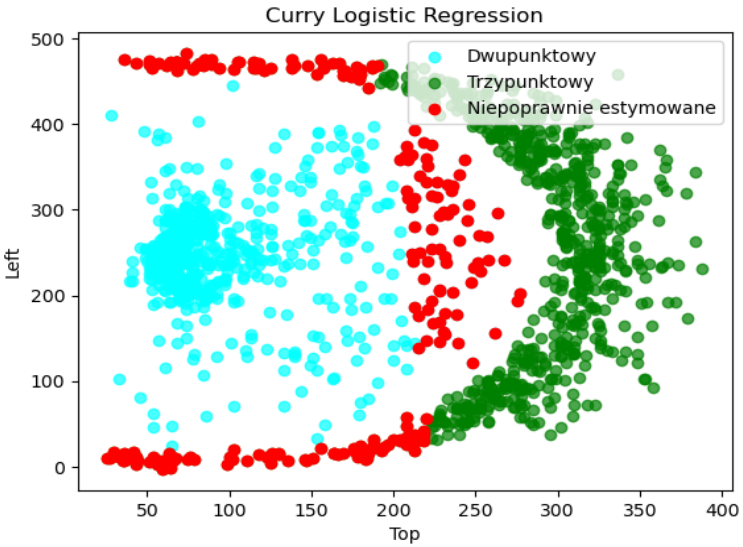


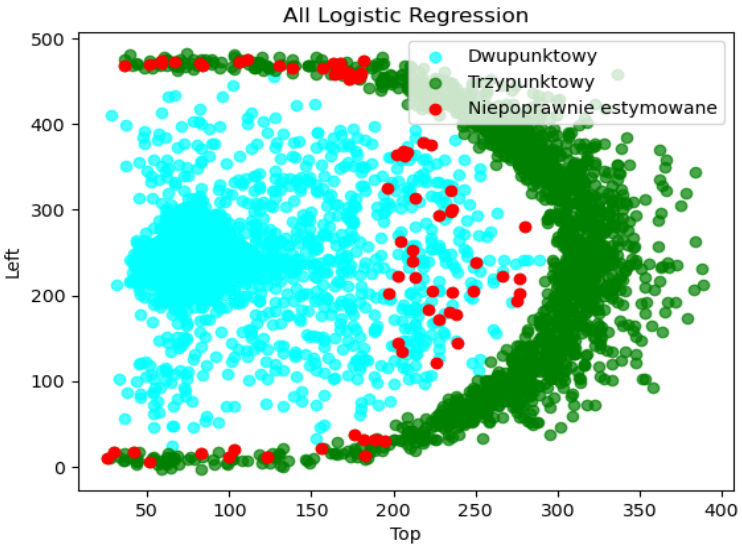


Regresja Logistyczna

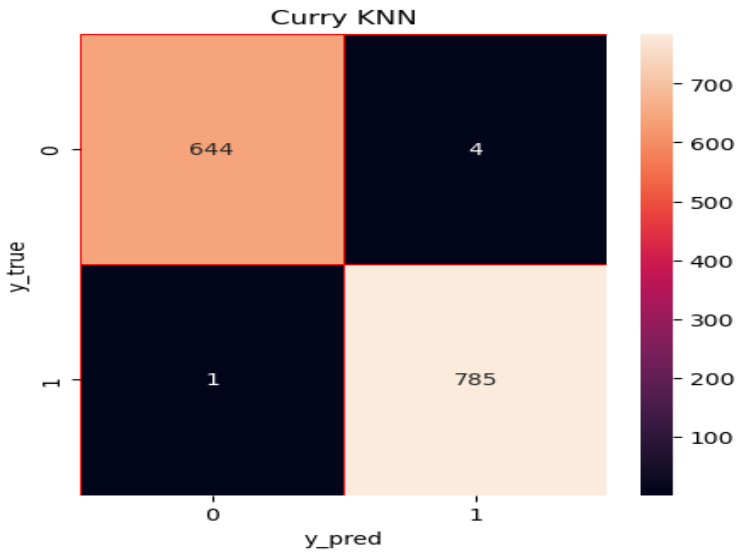


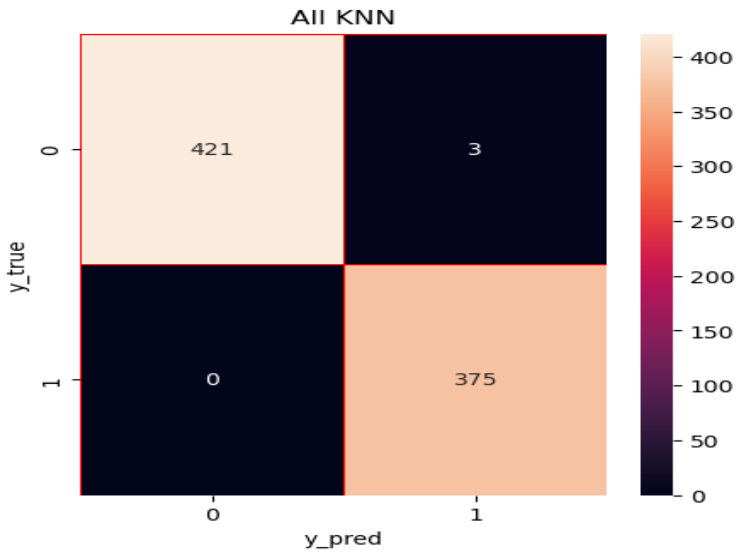


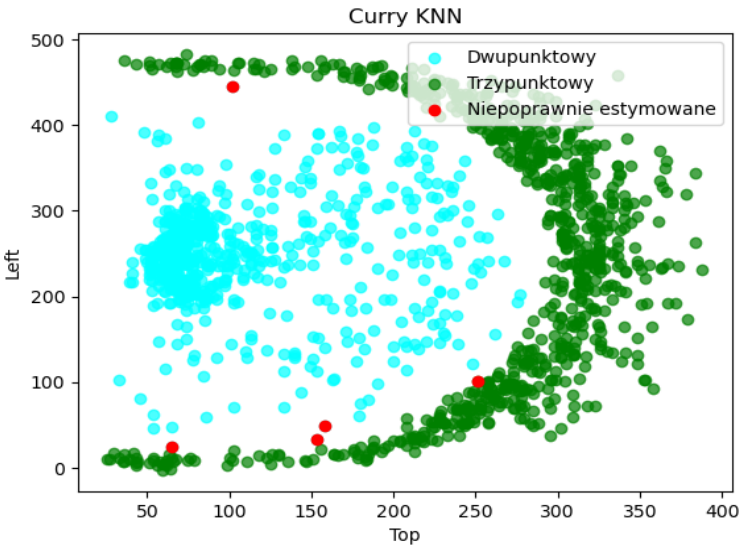


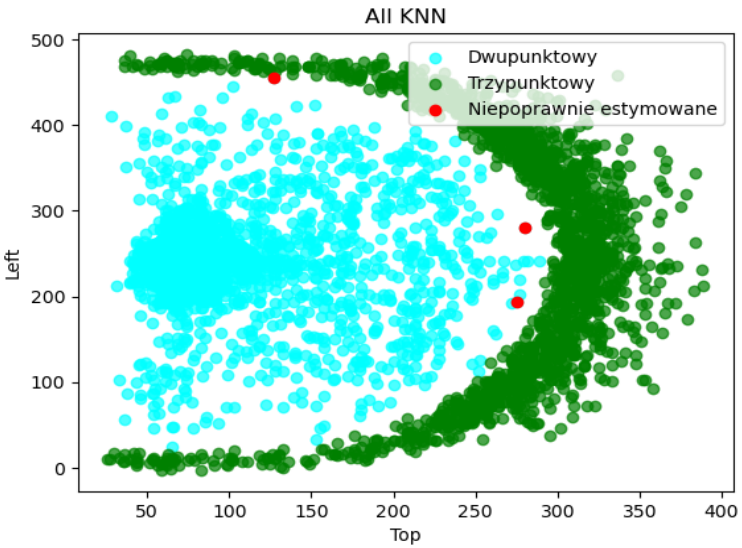


K–najbliższych sąsiadów

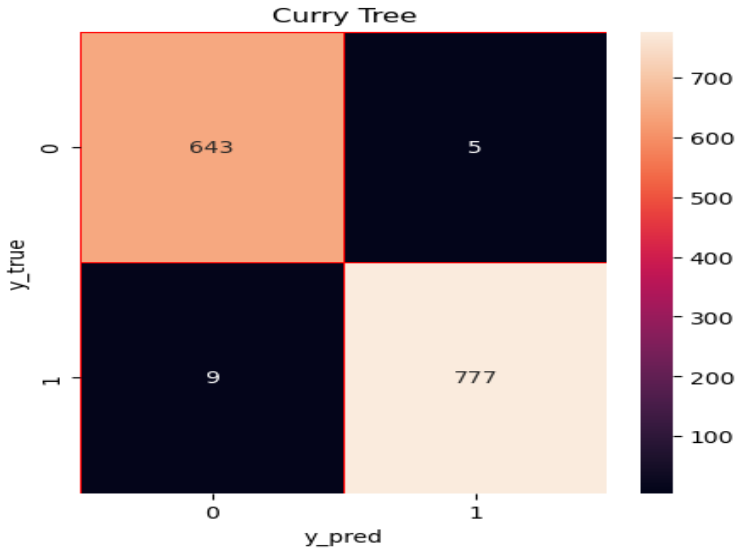


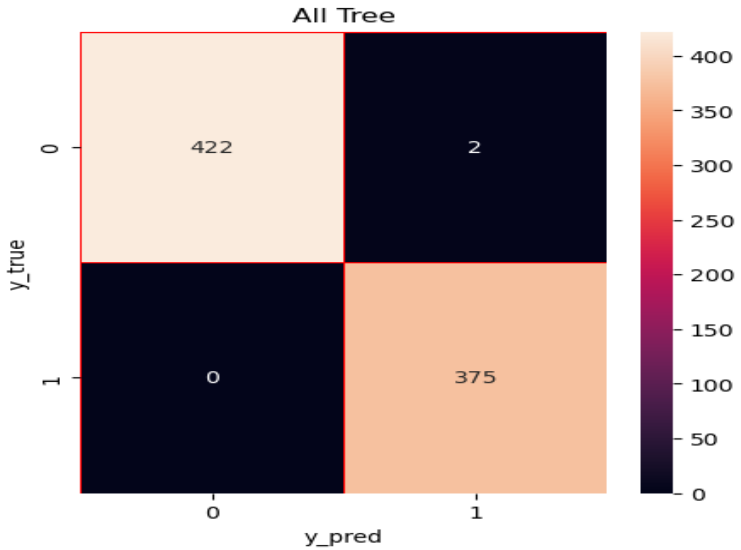


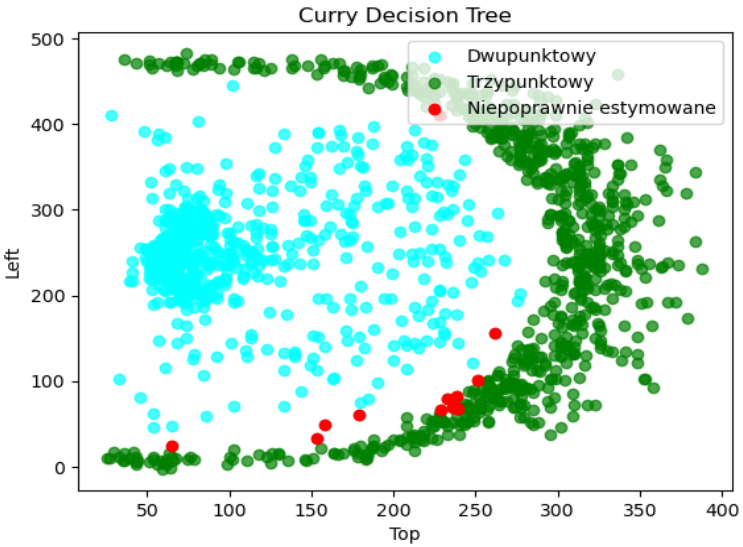


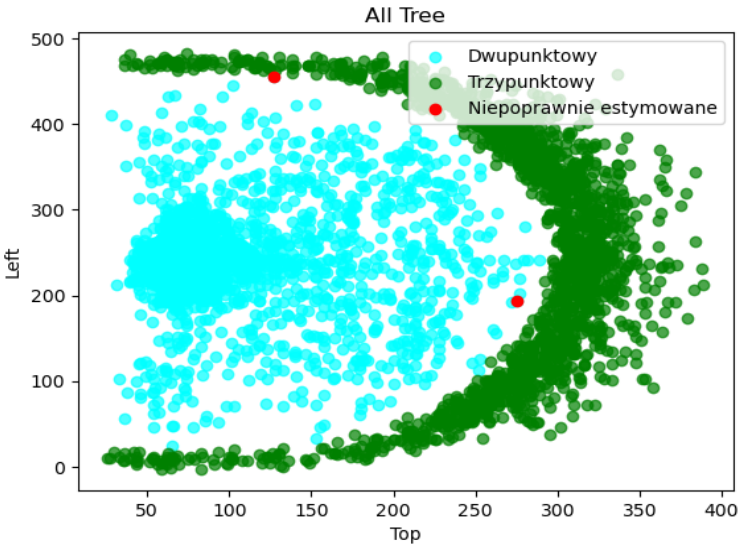


Drzewa


Curry Tree
643
5
9
777
y_true
y_pred
0
1
700
600
500
400
300
200
100


All Tree
422
2
0
375
y_true
y_pred
0
1
400
350
300
250
200
150
100
50
0


Curry Decision Tree
Dwupunktowy
Trzypunktowy
Niepoprawnie estymowane
Left
Top
500
400
300
200
100
0
50
100
150
200
250
300
350
400


All Tree
Dwupunktowy
Trzypunktowy
Niepoprawnie estymowane
Left
Top
500
400
300
200
100
0
50
100
150
200
250
300
350
400

Wdrożenie – Wnioski

- Nie jest potrzebne konstruowanie bardziej złożonych modeli obliczeniowo, gdyż te proste działają bardzo dobrze
- KNN radzi sobie najlepiej na danych zawodników których nie zna, natomiast drzewo decyzyjne na nowych akcjach zawodników, których już zna, regresja logistyczna odstaje od dwóch powyższych
- Drzewo będzie miało dużą głębokość, przez co będzie złożone obliczeniowo, wynika to z owalnego kształtu danych.
- Link do Githuba:
  https://github.com/MichalBrodackiPJA/Eksploracja-i-Wizualizacja-Danych/tree/master/Projekt_koncowy

Dziękuję za uwagę!