Notatki do prezentacji

## Dane

- Pokazać na kolumny których użyje lekko omówić inne

## Wstępna Ocena Danych

- Na podstawie wykresów słupkowych powiedzieć, że niepotrzebna będzie próbkowanie tej samej ilości z obu grup (gdyż ich liczność się wiele nie różni)

- Na podstawie wykresów typu scatter (punktowych), powiedzieć, że możemy odróżnić klasy. Dodać, że K-means (nienadzorowane) nie sprawdzi się bo centroid będzie w środku drugiego zbioru

- Nawiązać do histogramów, że mają różne rozkłady dla różnych zawodników (można to zaobserwować w moim repo)

## Przygotowanie danych

Właściwie to przeczytać, coś dokomentować ewentualnie

## Modelowanie

- Powiedz o trzech modelach, chciałeś bardziej skomplikowane (ale nie było po co o czym się przekonacie)

- Powiedzieć, że robiło się więcej, ale to dwa przykładowe sposoby wybierania najlepszego modelu

- Dla KNN najlepsze wyszło 2 sąsiadów (na podstawie f-score = 0.994) (może zbyt podobne do siebie były

- Dla decision Tree kara ccp\_alpha=0.00038, dała najlepsze accuracy na zbiorze walidacyjnym.

(ta alpha jest mała, ale to przez owal w granicy danych)

## Ewaluacja

f-score

1knn – 0.9968

1Tree – 0.991

1Regresja – 0.8567

Allknn – 0996

AllTree – 0.997

AllRegresja – 0.899

- Regresja wyszła najgorzej

- KNN wyszło lepiej

- Drzewa wyszły najlepiej jak już znały, gorzej w innej sytuacji

## Wnioski

Przeczytać, KNN bym użył, dzięki