Wstęp do uczenia maszynowego
Raport z projektu nr 2

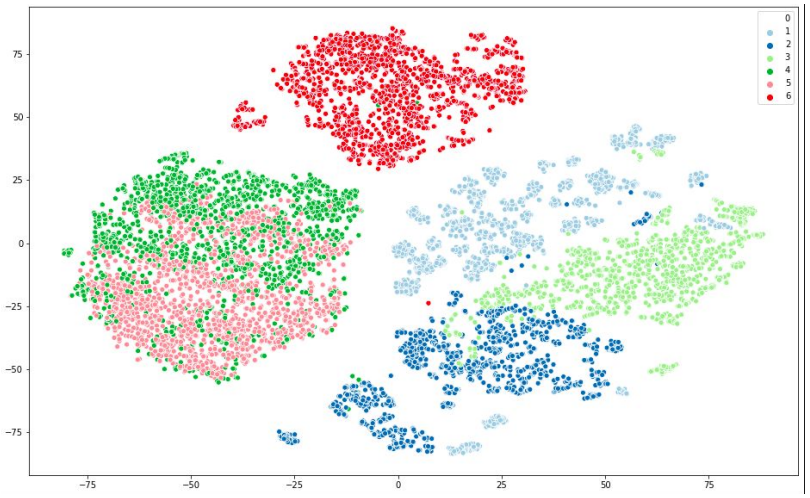
Filip Chrzuszcz Piotr Fic

Wstęp

Przedmiotem naszych analiz jest zbiór Human Activity. Zawiera on dane dotyczące ludzkiej aktywności, która była badana na 30 ochotnikach. Mieli oni zamontowane opaski na nadgarstkach, które za pomocą akcelerometru oraz żyroskopu analizowały ich ruchy. Dane zostały przeskalowane do przedziału [-1.1] oraz wyodrębniono z nich wiele różnych statystyk, tak że zbiór z którym pracowaliśmy miał 561 kolumn. Same aktywności dostały etykiety dzielące je na 6 grup, ale z racji charakteru projektu, który polegał na użyciu algorytmów klasteryzacji zdecydowaliśmy się ich nie używać. Wszystkie dane są typu numerycznego oraz nie zawierają żadnych braków.

Eksploracja danych

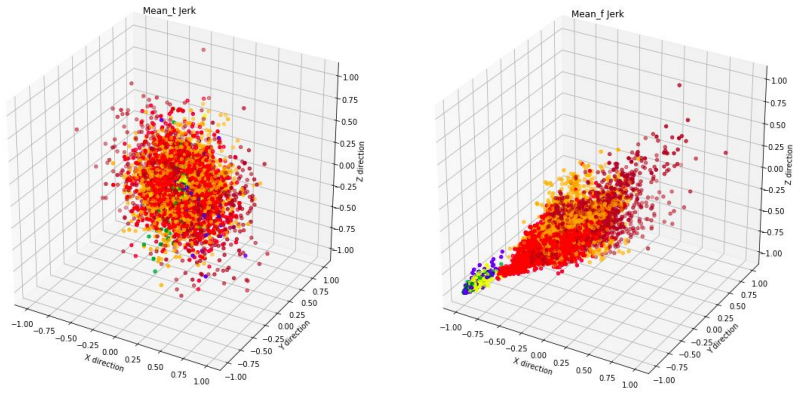
Z racji sporej ilości kolumn nie jest możliwe analizowanie ich po kolei. Najpierw zajęliśmy się sprawdzeniem czy dane są wizualnie separowalne. W tym celu zdecydowaliśmy się skorzystać z PCA oraz t-SNE. Narzędzia te pozwalają lepiej zwizualizować dane, tak aby można było ujrzeć wszelkie klastry jakie mogą one tworzyć. Zwłaszcza wynik uzyskany przez t-SNE jest satysfakcjonujący. Widać dość wyraźnie 3-4 klastry na których będzie można działać.



Dane narysowane po t-SNE

Następnie zdecydowaliśmy się na analizę przykładowych kilku pomiarów z akcelerometru i żyroskopu, tak aby móc zobaczyć na jakich danych będziemy pracować. Z racji że dane były zbierane w 3 wymiarach, to najłatwiej je było wyświetlać również w 3 wymiarach. Mamy do dyspozycji przeróżne statystyki takie jak zryw, entropia, czy też wychylenie w każdą ze stron. Te dane były następnie agregowane przez różnego rodzaju statystyki jak rozstęp międzykwartylowy, średnia, ale także transformata Fouriera. Spora część danych na wykresach nie wygląda jakkolwiek separowanie, jednakże są wyjątki. Często wyglądało to tak, iż transformata Fouriera pomagała wizualnie rozróżniać klastry.

Co naturalne najwięcej wykresów zrobiliśmy dla wszelkiego rodzaju średnich, z racji tego że są to dane najłatwiej interpretowalne. Posiadamy również sporo kolumn wysoce ze sobą skorelowanych, więc z pewnością dobrym krokiem w przyszłości będzie pozbycie się części z nich, tak aby ułatwić klasteryzację.



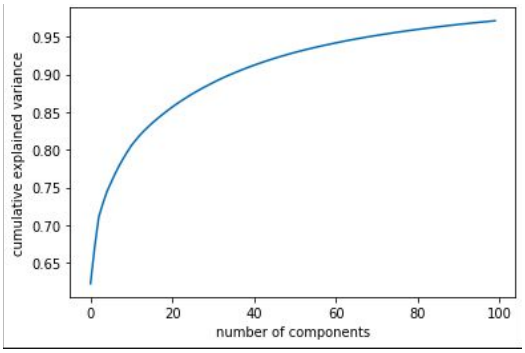
Wykres średniego zrywu w każdą ze stron przed i po transformacie Fouriera

Inżyniera cech

Z racji na dużą ilość kolumn, uznaliśmy iż dobrym pomysłem będzie próba zmniejszenia ich ilości, tak aby zmniejszyć ilość obliczeń, ale także wybrać dane które będą ułatwiały wybranym później algorytmom zadanie klasteryzacji. Zdecydowaliśmy się na sprawdzenie kilku podejść do tego tematu, tak aby w późniejszej fazie, po ich porównaniu zdecydować się na najlepsze.

1. PCA

Jako pierwsze podejście zdecydowaliśmy się użyć danych przetworzonych przez PCA. Jak udało się nam ustalić 80% wariancji jest przedstawiane jedynie przez 11 kolumn. Udało się więc w bardzo dużym stopniu zmniejszyć ilość zmiennych.



Ilość przedstawionej wariancji w zależności od ilości komponentów

2. Niska wariancja

Kolejnym podejściem jakiego użyliśmy bylo odrzucenie zmiennych o małej wariancji wewnątrz danej kolumny. Z racji tego że dane nie były wcześniej standaryzowane, to podejście ma sens Jako próg odcięcia ustaliliśmy 0.2. Jeśli wariancja w danej kolumnie była niższa od tej wartości, to wtedy pozbywaliśmy się jej W ten sposób udało się zredukować ilość kolumn do 55.

3. Factor Analysis

Jest to metoda statystyczna, która w uproszczonym tłumaczeniu ma na celu znalezienie ukrytych zmiennych nieskorelowanych (nazywanych czynnikami), które opisują zależności między wszystkimi zmiennymi ze zbioru danych. Stąd zauważalne jest pewne podobieństwo do PCA. Różnica polega na tym, iż PCA nie szuka “nowych” zmiennych (czynników), lecz stara się znaleźć je wśród już obecnych w zbiorze.
Przed przystąpieniem do analizy czynnikowej możemy sprawdzić czy zbiór posiada potencjał do jej wykorzystania. Służy do tego np. test Kaisera-Meyera-Olkina (KMO). Zbiór wartości wyników KMO to przedział [0, 1]. Im wyższy wynik, tym większy sens wykorzystania factor analysis. KMO obliczony dla naszego zbioru osiągnął wartość 0.979.
Pierwszym etapem factor analysis jest wybranie odpowiedniej liczby czynników. Jednym ze sposobów jest analiza wartości własnych, które im odpowiadają. Na tej podstawie odpowiednia liczba czynników dla naszego zbioru wynosi 63. Oznacza to, że właśnie do takiej liczby zmiennych przekształcenie to zredukowało zbiór danych.

4. Wysoka korelacja

Podczas eksploracji danych zauważyliśmy, że wiele z naszych zmiennych jest ze sobą wysoko skorelowanych. Za warte sprawdzenia podejście uznaliśmy próbę usunięcia części ze zmiennych, dla których istnieje odpowiadająca, wysoko skorelowana zmienna. Pozwala to na redukcję wymiaru i intuicyjnie tłumacząc, usunięcie powtarzalnych informacji ze zbioru. Za próg graniczny przyjęliśmy współczynnik korelacji dla pary zmiennych większy lub równy 0.9. Jeśli był on przekroczony jedna ze zmiennych z pary była usuwana ze zbioru. W ten sposób liczba zmiennych została zredukowana do 251.

5. Transformata Fouriera

Wizualizacje zmiennych z etapu EDA wskazywały, iż dane wydają się być łatwiej separowalne na klastry, kiedy do ich prezentacji wykorzystujemy zmienne, które zostały poddane transformacie Fouriera. Sugerowało to, że może jedynie te zmienne wystarczą do pełnej reprezentacji danych. W związku z tym aby przetestować ten pomysł ostatnim z testowanych podejść było pozostawienie w zbiorze jedynie zmiennych po transformacie. Ograniczyło to ich liczbę do 295.

Rezultaty badania metod redukcji wymiarowości

Za pomocą wyżej opisanych metod otrzymaliśmy 5 przekształconych wersji zbioru. Razem z oryginalnym zbiorem stanowi to 6 wersji do porównania. Dokonaliśmy go za pomocą trzech różnych modeli: k-means, birch oraz agglomerative. Badanie powtórzyliśmy dla liczby klastrów 3 oraz 6. Metryką użytą do oceny był współczynnik silhouette.

Współczynnik silhouette: liczba klastrów 6

	kmeans	agg	birch
orginal	0.132143	0.079372	0.079372
df_pca	0.254071	0.210469	0.201818
df_var	0.217846	0.190949	0.206739
df_factor	0.105811	0.215942	0.215942
df_corr	0.083635	0.046041	0.046041
df_fourier	0.137708	0.116516	0.143183

Współczynnik silhouette: liczba klastrów 3

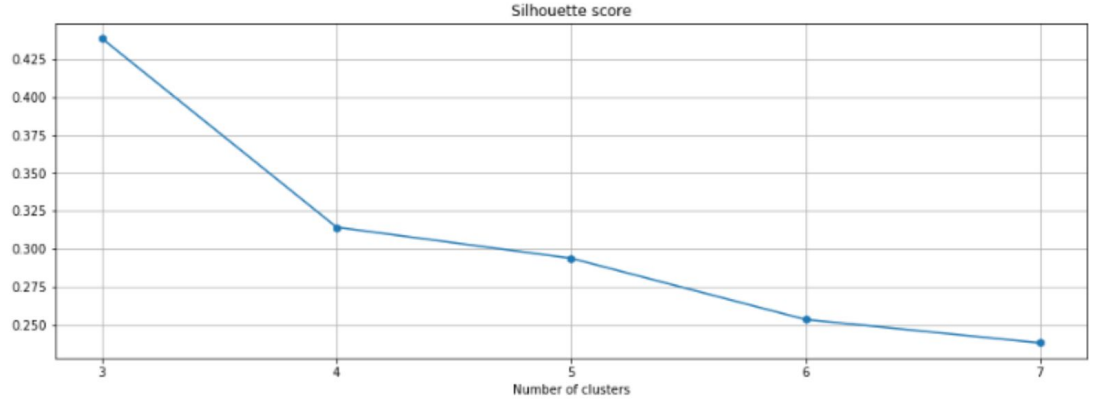
	kmeans	agg	birch
orginal	0.324273	0.306846	0.306846
df_pca	0.438609	0.444234	0.445431
df_var	0.483657	0.481741	0.481111
df_factor	0.308297	0.279283	0.279283
df_corr	0.174962	0.144576	0.144576
df_fourier	0.375957	0.356594	0.355685

W pierwszej kolejności możemy zauważyć ogółem wyższe wyniki dla testów o obranej liczbie klastrów 3. Nie jest to zaskoczeniem, gdyż wskazywały na to metody doboru tego parametru. Bardziej istotnym wnioskiem jest fakt, że wszystkie 3 modele osiągnęły dość podobne rezultaty. Najważniejsze w tym badaniu było jednak wybranie najbardziej optymalnej metody redukcji wymiarowości. Najlepsze poprawy współczynnika, stabilne względem modeli, przyniosły metody PCA i próg wariancji. Dla PCA współczynnik silhouette wzrósł o ok. 0.14 dla każdego z modeli. Pomimo, że metoda z eliminacją zmiennych o niskiej wariancji dała jeszcze większą poprawę, zdecydowaliśmy się ostatecznie na skorzystanie z PCA, jako bardziej sprawdzonej i uznanej metody.

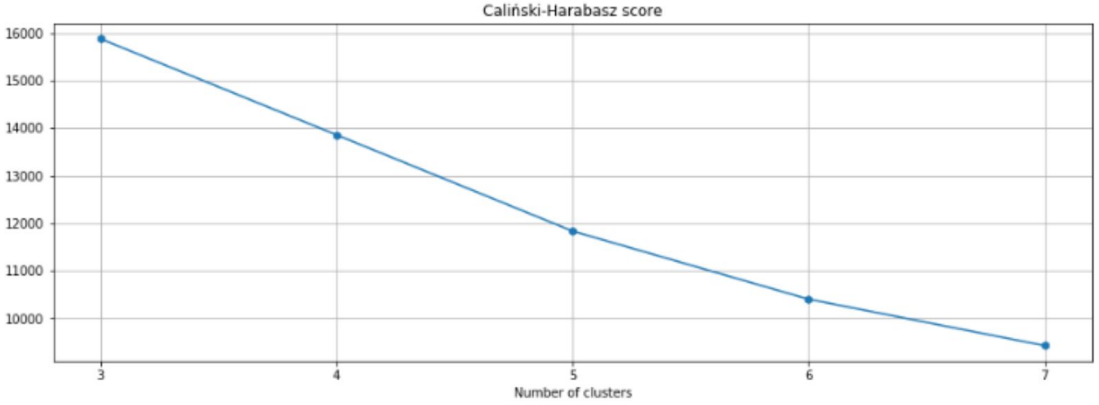
Finalny Model

Bazując na przeprowadzonym porównaniu metod inżynierii zmiennych do ewaluacji ostatecznych modeli użyjemy zbioru danych przekształconego z wykorzystaniem PCA. W celu ostatecznego upewnienia się co do optymalnej liczby klastrów, jeszcze raz skorzystaliśmy z metod jej doboru. Na podstawie metody łokcia oraz analizy współczynników silhouette i Calińskiego-Harabasza wybór nie może być inny niż 3 klastry. Chcemy podkreślić, iż nie dążyliśmy do uzyskania koniecznie liczby skupień równej liczbie oryginalnych kategorii oraz nie korzystaliśmy z tych etykiet aż do momentu skończenia modelowania. Dopiero później wykorzystamy je do porównania z uzyskanym klastrowaniem.

Dobór liczby klastrów względem współczynnika silhouette.



Dobór liczby klastrów względem współczynnika Calińskiego-Harabasza.



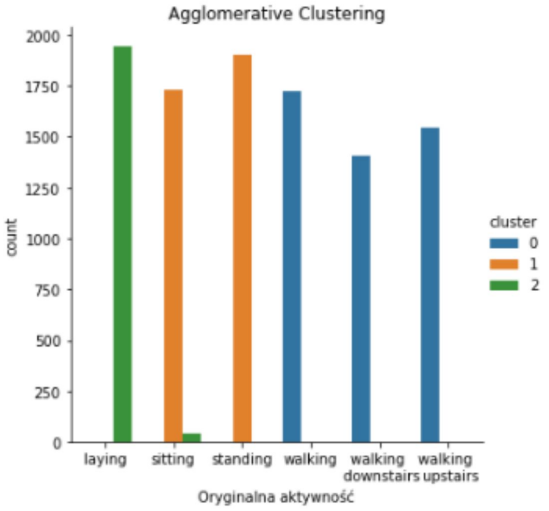
Poza użyciem wcześniej opisywanych modeli postanowiliśmy również sprawdzić działanie algorytmu DBSCAN na naszym zbiorze. Jest to o tyle interesujące, iż nie wymaga on podania wprost liczby klastrów. Wymagał natomiast dostrojenia hiper parametrów co uczyniliśmy według schematów proponowanych w literaturze. Ciekawym wnioskiem z użycia tego algorytmu był fakt, że również znalazł on 3 klastry. Po użyciu wszystkich algorytmów i otrzymaniu ich wyników klasteryzacji, mogliśmy przystąpić do analizy wyników.

Analizując wizualnie otrzymane skupienia dostrzegliśmy podobieństwo wyników DBSCAN i K-Means. Nieco odmienny podział na klastry zaproponował model Agglomerative Clustering. W celu lepszego poznania otrzymanych skupień sprawdziliśmy jak obserwacje rozkładają się według oryginalnych etykiet na klastry zdefiniowane przez modele. Najbardziej klarowny podział, dobrze rozróżniający aktywności, był owocem działania Agglomerative Clustering. Model ten rozdzielił obserwacje na:

- leżenie
- stanie oraz siedzenie
- chodzenie (w tym po schodach w dół i górę)

Pozostałe algorytmy dobrze odnajdywały grupę aktywności złożoną z: siedzenia/stania/leżenia, lecz 2 dodatkowe klastry były mieszanką obserwacji bez sensownej interpretacji. Na podstawie tej analizy za najlepszy model wybraliśmy Agglomerative Clustering.

Rozkład obserwacji na klastry Agglomerative Clustering



Finalnie dokonaliśmy analizy charakterystyki klastrów dla wybranego modelu. Wykorzystaliśmy podstawową metodę analizy średnich ze zmiennej względem klastrów oraz zaawansowane podejście polegające na użyciu modelu klasyfikacji XGBoost na zmiennej celu zdefiniowanej jako ‘klaster vs others’. Pozwoliło to na skorzystanie z features importance dla modelu XGB i interpretacji tego wyniku jako zmiennych wyróżniających dany klaster. Z powodu specyfiki zmiennych w zbiorze, a mianowicie ich trudnej interpretacji w sposób dosłowny, przedstawimy ogólniejsze podsumowanie.
Klaster ‘leżenie’: ujemne wartości zmiennych z grupy GravityAcc.
Klaster ‘siedzenie/stanie’: mocno wyróżniająca się zmienna tGravityAcc-max()-X.
Klaster ‘chodzenie’: dodatnie wartości zmiennych z grupy BodyAcc.

Podsumowanie
Raport w każdej części zawiera bieżące komentarze i wnioski, z tego powodu poniżej zamieszczamy najważniejsze informacje dotyczące efektów pracy nad zadaniem klasteryzacji zbioru Human Activity:

Wybrany model: Agglomerative Clustering
Liczba klastrów: 3
Interpretacja klastrów: laying - sitting/standing - walking/upstairs/downstairs