PROJEKT SZTUCZNA INTELIGENCJA

EMILIA PIEŚNIKOWSKA
180112
GRZEGORZ FIEDORUK
179981

Temat 27: Grupowanie wyników wyszukiwania.

1. Opis problemu:

Problem grupowania, czy też klasteryzacji, danych jest często poruszany w przypadku sztucznej inteligencji i uczenia maszynowego. Przez lata powstało wiele różnych algorytmów grupowania, które wraz z biegiem czasu zaczęły rozdzielać się na 2 kategorie: fuzzy clustering algorithms i crisp clustering algorithms. Pierwszy z tej dwójki charakteryzuje się miękkim przydziałem danych do grup. Punkty danych nie są twardo przypisane do jednej z grup, lecz mają one prawdopodobieństwo z jakim będą przynależeć do 2 lub więcej powstałych w wyniku pracy algorytmu grup. Drugi gatunek algorytmów grupowania jest wręcz jego przeciwnością i polega na twardym przydziale danych do grup, bazując na tym czego nauczył się algorytm w trakcie pracy. Algortymy grupowania są to przykłady uczenia nienadzorowanego, czyli takiego, w którym użytkownik nie ma wpływu na to, co dzieję się wewnątrz algorytmu. Program sam musi nauczyć się potrzebnych przydziałów w celu odpowiedniej selekcji danych. Metody te pozwalają na automatyczne identyfikowanie wzorów i struktur wewnątrz baz danych. Przykładem użycia takich algorytmów mogłyby być jakiekolwiek rodzaju sklepy, które kategoryzowałyby użytkowników swoich kart członkowskich bazując na ich najczęstszych zakupach. Innym zastosowaniem, który pokażemy w tym projekcie, jest sprawdzanie zużycia i poniekąd opłacalności pojazdów kategoryzując je na podstawie ich wykorzystania (liczby przejechanych kilometrów) i wieku auta.

2. Teoretyczny opis użytej metody

Wybrana przez nas metoda klasteryzacji to K-Means. Jest to jeden z najpopularniejszych i zarazem najprostszych algorytmów grupowania. Celem tego algorytmu jest grupowanie podobnych punktów danych razem w celu podzielenia całego zestawu danych na z góry określoną liczbę klastrów. Pierwszym krokiem do implemenacji takiego algorytmu jest wybranie liczby, która będzie reprezentować liczbę grup na jakie dzielimy dane.

```
number_of_clusters = 3
```

Analogicznie liczba ta reprezentuję również liczbę środków klastrów, na podstawie których działa algorytm. Głównym działaniem w całym algorytmie jest przypisywanie punktów, do tych klastrów, których środki znajdują się najbliżej aktualnie sprawdzanego punktu. Człon „means” w nazwię odpowiada za uśrednianie danych, czyli innymi słowy znajdywanie środków klastrów. W calu przetworzenia danych, na których algorytm uczy się przydzielania algorytm zaczyna z grupą losowo wygenerowanych środków (rozmiar grupy predefiniowany przez

programistę), i dostosowuje środek wraz z kolejnymi iteracjami algorytmu, żeby było one jak najbardziej optymalnie rozlokowane wśród ogółu danych.

Skrócony opis przebiegu działania:
1. Wybierz funkcję d(x,y), której zadaniem będzie obliczanie odległości między punktami danych.\
2. Zdefiniuj liczbę klastrów.
3. Wylosuj tyle punktów, ile jest klastrów w celu określenia pierwotnych środków grup.
4. Oblicz odległość wszystkich punktów od wyznaczonych punktów.

```
def kmeans(values, k):
    clusters = []
    for _ in range(len(values)):
        clusters.append(0)

    centers_of_clusters = sample(list(values), k)

    while True:
        for j, value in enumerate(values):
            min_dist = float('inf')
            for cluster_number, centroid in enumerate(centers_of_clusters):
                dist = sqrt((centroid[0] - value[0]) ** 2 + (centroid[1] - value[1]) ** 2)
                if dist < min_dist:
                    min_dist = dist
                    clusters[j] = cluster_number
        new_centroids = DataFrame(values).groupby(by=clusters).mean().values
        if k == 2 or not count_nonzero(centers_of_clusters - new_centroids):
            break
        else:
            centers_of_clusters = new_centroids
    return centers_of_clusters, clusters
```

5. Na podstawie wyników obliczeń przydziel punkty do odpowiednich klastrów.
6. Wyznacz nowe środki klastrów korzystając ze średniej arytmetycznej dystansów między danymi z danej grupy.
7. Powtórz od punktu 4. tak długo jak centra klastrów będą się zmieniać

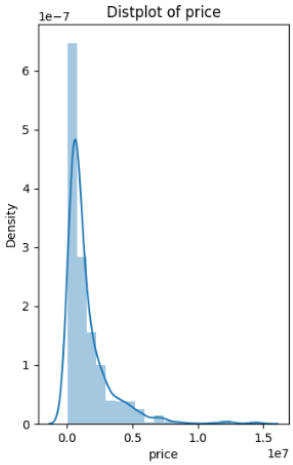
3. Opis realizacji zadania

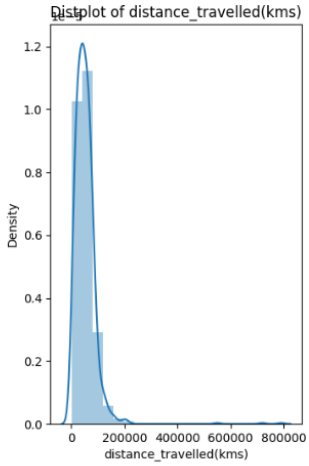
Tak jak wspomniane w punkcie drugim, metodą klasteryzacji jest K-means. Jako źródło danych posłużyła nam baza danych dotycząca samochodów (źródło: kaggle.com). Baza danych zawierała wiele atrybutów, ale do realizacji projektu użyliśmy atrybutów: price (cena samochodu), car age (wiek samochodu), brand (marka auta), distance travelled (przebieg samochodu). Dane przechowywane są w pliku .csv, co ułatwia łatwy do nich dostęp z poziomu kodu w celu przetworzenia przez algorytm. Głównymi znacznikami, według których dokonywaliśmy podziału na klastry były wiek samochodu, przebieg i jego cena. Dzięki użyciu tych danych mogliśmy zdefiniować 3 główne przedziały eksploatacji i zarazem opłacalności poszczególnych aut.

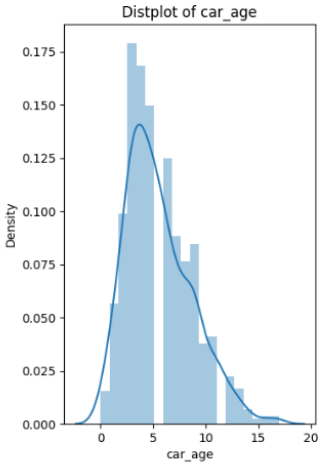
4. Przedstawienie wyników

W celu zobrazowania wyników grupowania danych sporządziliśmy odpowiednie wykresy.

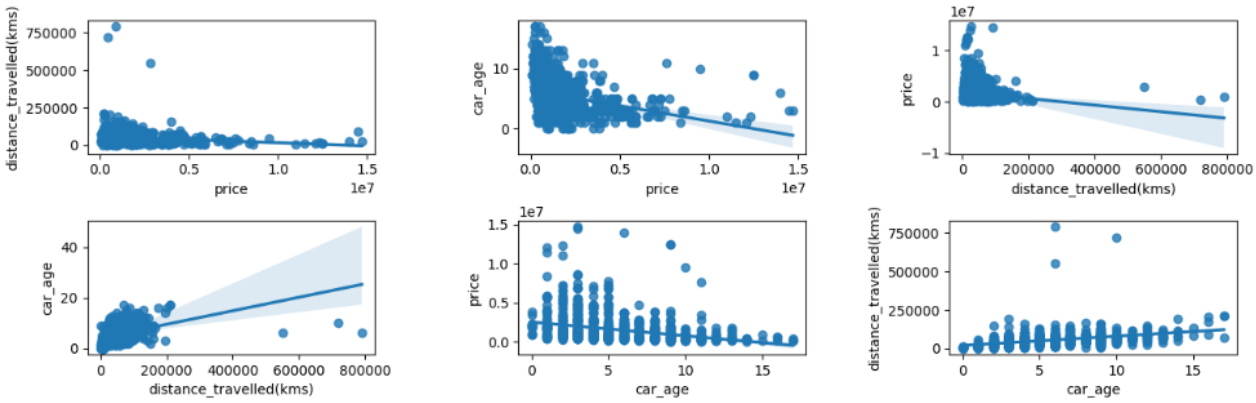
1) Rozkład gęstości danych.

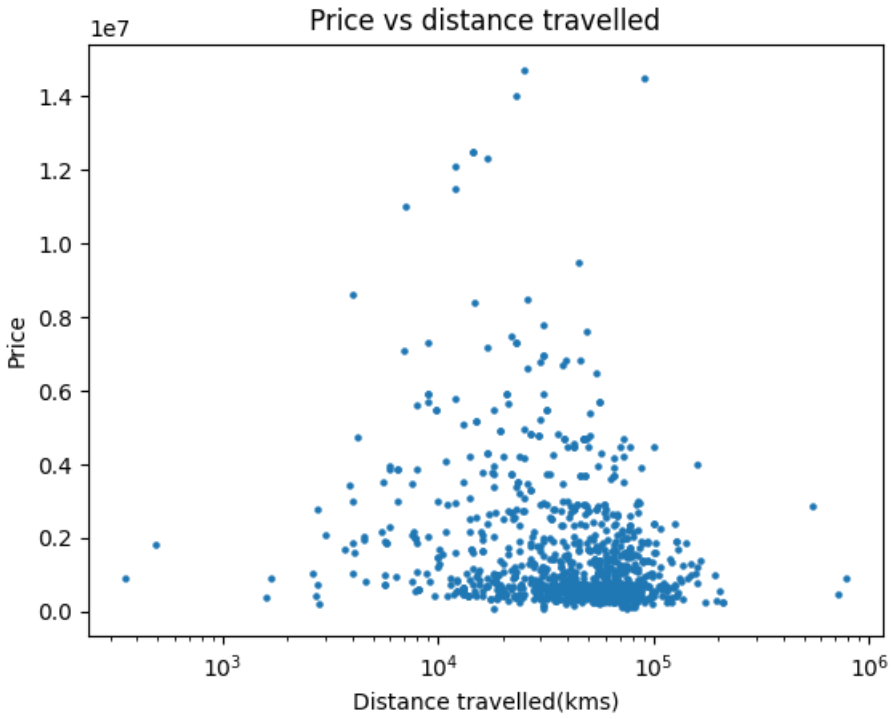


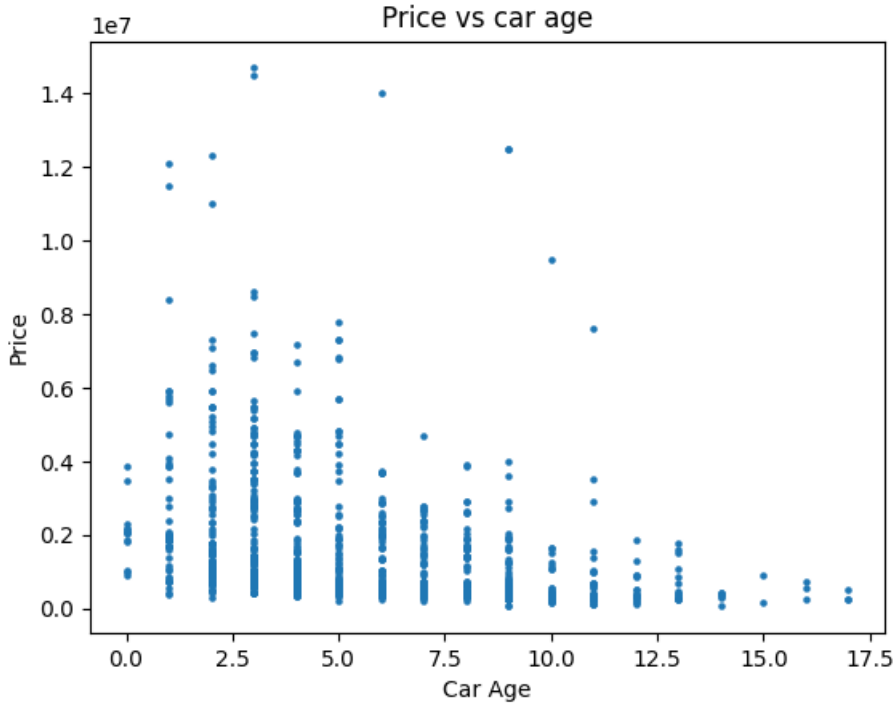




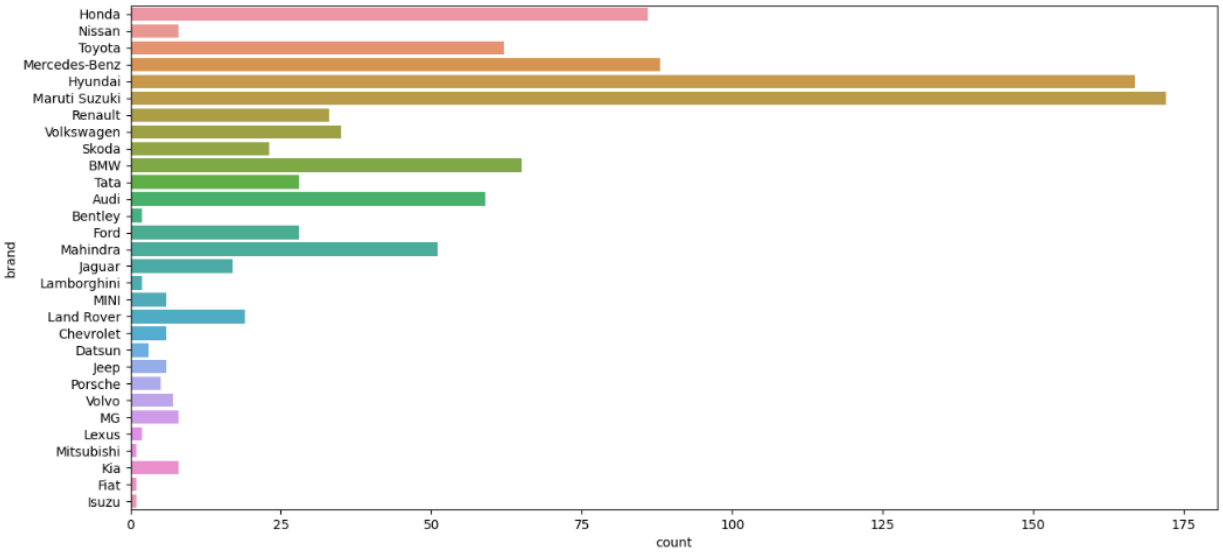
2) Wykresy zależności wybranych atrybutów:



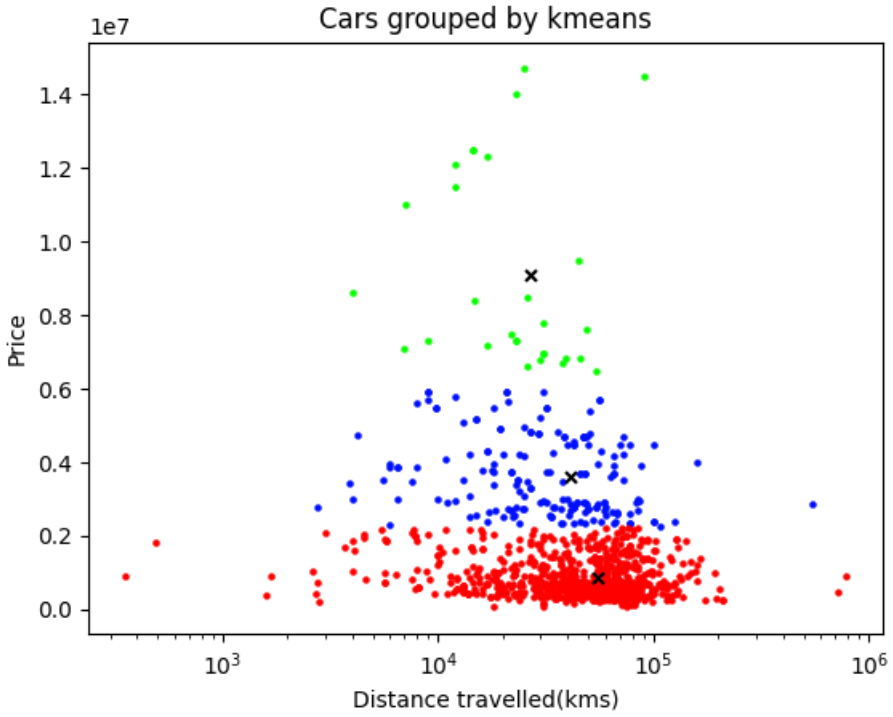

Price vs distance travelled
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Price
10^3
10^4
10^5
10^6
Distance travelled(kms)

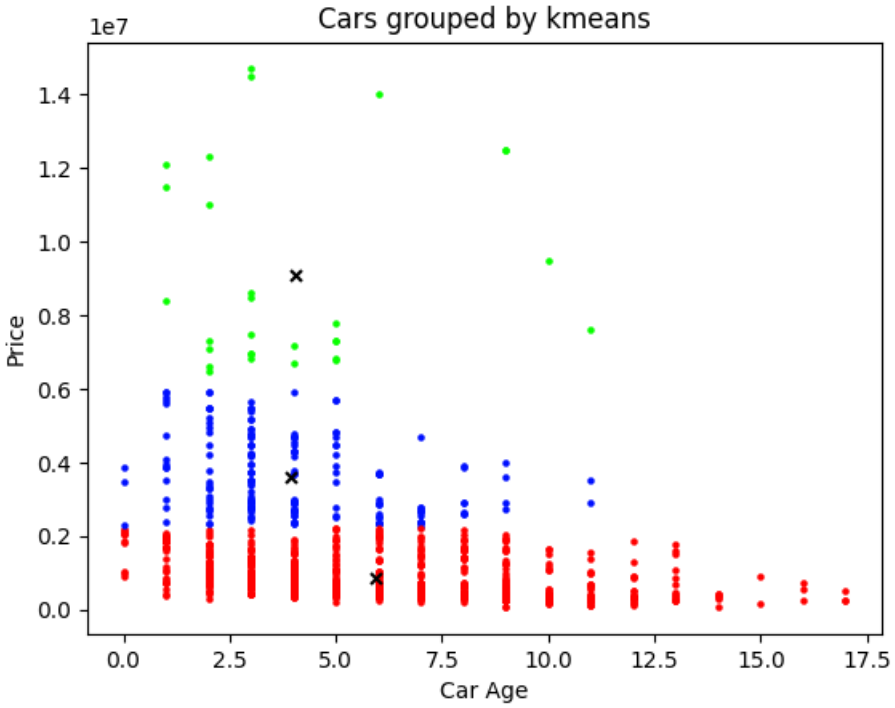

Price vs car age
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Price
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Car Age

3) Rozkład marek samochodów w bazie danych



4) Wykresy obrazujące klastryzację





5. Wnioski

Wnioski dotyczące wykresów:

Wykres pierwszy przedstawia rozkład gęstości wartości wykorzystywanych atrybutów w celu zobrazowania jak może to wpływać na dalsze wykresy. Widać, że wszystkie z tych wartości skupiają się raczej dookoła niższych wartości, jednak z odstępstawami w stronę większych wartości. Najwięcej takowych różnic można zaobserwować w wieku samochodów, a najmniej w przebiegu.

Wykres drugi służy do prostego zobrazowania zależności między atrybutami w celu wyboru tych atrybutów, które najbardziej nadają się do poddania ich dalszej analizie. Z przedstawionych zestawień wynikło, że zestawy ceny i przebiegu i ceny i wieku auta są najbardziej odpowiednie.

Wykresy trzeci i czwarty, pozwalają przyjrzeć się bliżej wybranym zależnościom z wykresu 2. Zasotosowana została tam również skala logarytmiczna w celu lepszego przedstawienia rozproszenia danych.

Wykres piąty jest wykresem pokazującym rozkładam marek samochodów w bazie danych. Jest to wykres arybutu, który nie jest brany pod uwagę w trakcie grupowania z użyciem algorytmu K-means, jednak ma on ogromny wpływ na ewentualne błędy przy tym grupowaniu, jako że oczywistym jest, że niektóre marki samochodowe są droższe od innych. W przypadku takich samochodów ceny będą wywindowane w góre pomimo równego poziomu eksploatacji co może wprowadzać w błąd proces klastryzacji.

Wykresy 6 i 7 to wykresy końcowego wyniku grupowania przez algorytm K-means. Dzielą one zbiór danych na wyraźne 3 grupy. Uzyskany podział można interpretować jako podział na opłacalność danych aut. Mamy grupę najbardziej opłacalną (grupa niebieska), grupę średnio opłacalną(grupa czerwona) i grupę nieopłacalną (grupa zielona). Centra klastrów zaznaczone są czarnymi krzyżykami. Tak jak wspomniałem przy wykresie piątym, na uzyskane wyniki trzeba brać poprawkę na anomalie nie brane pod uwagę przez algorytm. Mogą to być: droga marka samochodu lub duży wiek co robi z niego gratkę dla kolekcjonerów. Obie te cechy sztucznie windują cenę do góry co może nie współgrać ze schematem wyoczonym przez algorytm.

Końcowe wnioski ogólne:

Z przeprowadzonej analizy wynika, że K-means jest prostą metodą grupowania, jednak spelnia swoje zadanie. Dla przedstawionych danych wejściowych poradził on sobie dobrze z grupowaniem ich względem zadanych parametrów. Wyniki podziału jest logiczny i przejrzysty, jednak ma on naszym zdaniem swoje wady. Jako, że jest to algorytm czysto matematyczny bazujący tylko i wyłącznie na wartościach liczbowych to nie uwzględnia on czynników spoza swojego zasięgu co ma wpływ na nieodpowiednie przypisanie do klastrów. Przykładem tego, że

być zabytkowy samochód, który jest bardzo stary i ma bardzo wysoką cenę, co automatycznie przyporządkowało by go do grupy najmniej opłacalnych samochodów, co niekoniecznie musi być prawdą.

6. Bibliografia:

- wykłady dr. Szymańskiego
- „K-means clustering: Algorithm, Applications, Evaluation Methods and Drawbacks”
https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks- aa03e644b48a
-kaggle.com – źródło skąd wzięliśmy bazę danych