Problem: Jakie kluby lubią studenci?

W próbie odpowiedzenia na to pytanie przeprowadzono ankietę, następnie uzyskane odpowiedzi poddano analizie metodami uczenia maszynowego.
Grupą docelową ankiety byli studenci trzeciego roku Bioinformatyki. Rozwiązanie przedstawionego problemu pozwoliłoby lepiej zrozumieć społeczność studencką tego kierunku, ich preferencje i zainteresowania. Odpowiedź na pytanie "Jakie kluby lubią studenci?" mogłaby być bardzo wartościowa na przykład przy wyborze miejsca na imprezę urodzinową, przyszłe zgromadzenia absolwentów czy po prostu w celu dobrego spędzenia czasu.

Opis i cel ankiety:

Ankieta składała się z 12 pytań zamkniętych (w późniejszym etapie dodano ostatnie pytanie).
Pytania ankiety zostały sformułowane w taki sposób, by odpowiedź można było przedstawić na skali lub jako wartości prawda/fałsz.
Ułatwiło to przekształcenie odpowiedzi do wartości numerycznych w celu dalszej analizy.

Opis prac:

Preprocessing danych:

Pierwszym krokiem po zebraniu odpowiedzi było przygotowanie danych.
Korzystając z programu WEKA dokonaliśmy selekcji atrybutów.
Ograniczyliśmy listę atrybutów do sześciu najistotniejszych:

1. Płeć
2. W jaki licznym gronie przeważnie wychodzisz do klubu?
3. Jak ważne są dla Ciebie wydarzenia i imprezy okolicznościowe?
4. Jak ważna jest dla Ciebie strefa dla palących?
5. Czy spożywasz alkohol?
6. Jak ważna jest dla Ciebie bogata oferta alkoholi na barze?

Następnie uproszczono nazwy atrybutów i dokonano konwersji ich wartości na wartości numeryczne.

Zawartość pliku przed konwersją:

```
Płeć,W jaki licznym gronie przeważnie wychodzisz do klubu?,"Jak ważne są dla Ciebie wydarzenia i imprezy okolicznościowe? (np. silent disco, wieczory karaoke, koncerty, spotkania z celebrytami)",Jak ważna jest dla Ciebie strefa dla palących?,Czy spożywasz alkohol?,Jak ważna jest dla Ciebie bogata oferta alkoholi na barze?
Mężczyzna,W większym gronie,1,1,Tak,1
Mężczyzna,Z osobą towarzyszącą,1,4,Nie,1
Kobieta,W większym gronie,1,1,Tak,3
Kobieta,W większym gronie,2,1,Tak,3
Kobieta,W grupie 3 - 4 osób,3,2,Tak,3
Kobieta,W grupie 3 - 4 osób,2,1,Tak,2
Kobieta,W grupie 3 - 4 osób,4,1,Tak,4
Mężczyzna,W grupie 3 - 4 osób,3,1,Tak,3
Kobieta,W grupie 3 - 4 osób,4,4,Tak,4
Mężczyzna,W grupie 3 - 4 osób,3,1,Tak,5
```

Zawartość pliku po konwersji:

```
sex,team_size,events,smoke_room,drinker,alcohol_choice
0,4,1,1,1,1
0,2,1,4,0,1
1,4,1,1,1,3
1,4,2,1,1,3
1,3,3,2,1,3
1,3,2,1,1,2
1,3,4,1,1,4
0,3,3,1,1,3
1,3,4,4,1,4
0,3,3,1,1,5
```

Zastosowane algorytmy uczenia maszynowego:

- scikit-learn.KMeans
- scikit-learn.AgglomerativeClustering

Algorytm k-means jest jednym z podstawowych algorytmów uczenia nienadzorowanego. Należy do algorytmów analizy skupień (inaczej grupowania, klasteryzacji) i pozwala na podzielenie elementów na określoną ilość klas ze względu na podobieństwo.

Grupowanie hierarchiczne (agglomerative clustering) jest odmienną grupą metod grupowania danych w stosunku do metod bazujących na minimalizacji skalarnego współczynnika jakości jak np. algorytm k-means. Metoda ta bazuje na budowie grafu w postaci drzewa. Algorytm ten jest typem algorytmów z dołu do góry „bottom – top" gdzie zakłada się, że każdy wektor stanowi oddzielny klaster, a następnie łączy się małe klastry w coraz to większe. Proces łączenia realizowany jest na zasadzie poszukiwania klasterów leżących najbliżej siebie i zastępowania ich nowym większym klastrem, stanowiącym połączenie dwóch poprzednich. Proces ten stopniowo postępuje aż do chwili, w której zostanie osiągnięta właściwa liczba klastrów (określona przez użytkownika) lub do momentu gdy wszystkie wektory znajdują się w jednym klastrze.

Wyniki:

Zbiór uczący:

K-Means:

	A	B	C	D	E	F	G	H
1		sex	team_size	events	smoke_room	drinker	alcohol_choice	Clusters
2	4	1	3	3	2	1	3	0
3	6	1	3	4	1	1	4	0
4	7	0	3	3	1	1	3	0
5	9	0	3	3	1	1	5	0
6	10	1	4	4	1	1	4	0
7	18	0	4	4	3	1	5	0
8	20	0	3	5	2	1	2	0
9	21	1	4	4	1	1	4	0
10	26	1	1	2	1	1	5	0
11	8	1	3	4	4	1	4	1
12	16	0	2	1	4	1	4	1
13	19	1	3	2	4	1	5	1
14	22	1	4	2	5	1	3	1
15	23	1	4	2	4	1	3	1
16	25	0	4	1	4	1	4	1
17	28	0	3	3	5	1	5	1
18	29	0	3	1	5	1	3	1
19	0	0	4	1	1	1	1	2
20	2	1	4	1	1	1	3	2
21	3	1	4	2	1	1	3	2
22	5	1	3	2	1	1	2	2
23	11	0	4	2	2	1	3	2
24	12	0	3	1	1	1	4	2
25	13	1	3	1	1	1	3	2
26	15	0	4	1	1	1	1	2
27	27	0	4	2	1	1	2	2
28	30	0	3	2	1	1	3	2
29	1	0	2	1	4	0	1	3
30	14	0	4	2	5	0	1	3
31	17	0	3	3	5	1	1	3
32	24	1	3	1	4	1	1	3

Agglomerative Clustering:

	A	B	C	D	E	F	G	H
1		sex	team_size	events	smoke_room	drinker	alcohol_choice	Clusters
2	18	0	4	4	3	1	5	0
3	8	1	3	4	4	1	4	0
4	16	0	2	1	4	1	4	0
5	19	1	3	2	4	1	5	0
6	23	1	4	2	4	1	3	0
7	25	0	4	1	4	1	4	0
8	22	1	4	2	5	1	3	0
9	28	0	3	3	5	1	5	0
10	29	0	3	1	5	1	3	0
11	6	1	3	4	1	1	4	1
12	9	0	3	3	1	1	5	1
13	10	1	4	4	1	1	4	1
14	21	1	4	4	1	1	4	1
15	26	1	1	2	1	1	5	1
16	20	0	3	5	2	1	2	1
17	1	0	2	1	4	0	1	2
18	24	1	3	1	4	1	1	2
19	14	0	4	2	5	0	1	2
20	17	0	3	3	5	1	1	2
21	0	0	4	1	1	1	1	3
22	2	1	4	1	1	1	3	3
23	3	1	4	2	1	1	3	3
24	5	1	3	2	1	1	2	3
25	7	0	3	3	1	1	3	3
26	12	0	3	1	1	1	4	3
27	13	1	3	1	1	1	3	3
28	15	0	4	1	1	1	1	3
29	27	0	4	2	1	1	2	3
30	30	0	3	2	1	1	3	3
31	4	1	3	3	2	1	3	3
32	11	0	4	2	2	1	3	3

Nazwy (numery) otrzymanych klastrów różnią się od tych otrzymanych metodą K-Means ale rozmiary klastrów są bardzo podobne.

Zbiór testowy:

	A	B	C	D	E	F	G	H	I	J
1		sex	team_size	events	smoke_room	drinker	alcohol_choice	cluster z ankiety	wyniki_kmeans	wyniki_hierarchical
2	0	1	3	4	1	1	4	0	0	2
3	1	1	4	3	1	1	2	2	2	1
4	2	1	3	2	1	1	3	2	2	1
5	3	0	3	2	4	1	5	1	3	0
6	4	1	4	2	4	1	4	1	3	0
7	5	0	3	1	1	1	2	2	2	1
8	6	0	4	1	1	1	3	0	2	1
9	7	0	4	1	4	1	1	3	1	3
10	8	0	3	5	1	1	5	0	0	2
11	9	0	3	4	4	1	3	3	1	0
12										

Wyniki uzyskane metodami uczenia maszynowego nie zawsze pokrywają się z tym który klaster wybrał student odpowiadający na ankietę testową. Możliwe, że błąd wystąpił po stronie użytkownika, ale próba 10 testów jest zbyt mała by to określić. Pomijając możliwość błędu osoby ankietowanej, duża część przewidywanych wyników została poprawnie zakwalifikowana.

Widoczne są również pary klastrów które zostały zamienione miejscami. Może to wynikać z nienajlepiej dobranych pytań lub zbyt małego zbioru uczącego.

Oba algorytmy podzieliły dane na cztery bardzo podobne do siebie klastry, które wyłoniły następujące grupy studentów:

- grupa, dla której obecność możliwych eventów organizowanych w klubie ma znaczenie oraz dla której istotna jest bogata oferta alkoholi na barze, nieistotna jest natomiast obecność palarni,
- grupa, która nie jest zainteresowana eventami, dla której istotna jest i obecność palarni i bogata oferta alkoholowa,
- grupa, która nie jest zainteresowana eventami, dla której ważna jest obecność palarni, natomiast nieistotny jest duży wybór alkoholi na barze,
- grupa, która nie jest zainteresowana eventami, dla której nie jest ważna ani obecność palarni, ani bogata oferta alkoholowa.

W każdej grupie praktycznie każda osoba uznała się za osobę pijącą, nie wpłynęło to jednak na to, czy istotny jest dla kogoś duży wybór alkoholi oferowanych w danym lokalu. Można zatem uznać, że ten atrybut nie grał żadnej roli w formowaniu klastrów.

Podobne odpowiedzi między klastrami można było zauważyć dla atrybutu *team_size*, oznaczającego rozmiar grupy, w której ktoś wybiera się do klubu. Zdecydowanie przeważają wartości 3 oraz 4, co oznacza, że większość studentów preferuje chodzenie do tego typu lokali w grupach co najmniej trzyosobowych. Sporadycznie zdarzały się odpowiedzi, które sugerowałyby osoby chodzące do klubów z osobą towarzyszącą lub nawet samotnie. Można dzięki temu wywnioskować, że ten atrybut nie miał zbyt dużego znaczenia przy klastrowaniu.

Nie zaobserwowaliśmy, by w jakiejkolwiek grupie była przewaga żadnej z płci, uznaliśmy zatem, że płeć nie wypłynęła na formowanie klastrów.

Początkowo wydawało się, że atrybut *events* - oznaczający ewentualną obecność jakichś wydarzeń w klubie - będzie wpływał na charakterystykę

klastrów, ale po głębszej analizie okazało się, że wartości tego atrybutu nie są bardzo zróżnicowane pomiędzy grupami. W trzech na cztery grupy obecność eventów nie grała znaczącej roli przy wybraniu klubu, do którego ktoś chciałby się udać.

Pozostały zatem dwa atrybuty, których wartości między klastrami różniły się na tyle, że pozwoliły one zdefiniować każdą grupę. Były to atrybuty:

- *smoke_room* - jak istotna jest obecność palarni/strefy dla palących w danym klubie
- *alcohol_choice* - jak istotna jest bogata oferta alkoholi na barze.

Można zatem uznać, że w głównej mierze na podstawie tych dwóch atrybutów formowane były klastry. Dzięki temu udało nam się wywnioskować, że dla studentów biorących udział w naszej ankiecie najważniejszymi aspektami przy wyborze klubu były obecność palarni i bogata oferta alkoholi na barze.

Założeniem zadania było zebranie ok. 30 odpowiedzi na ankietę. Taki zbiór danych okazał się jednak nie do końca wystarczający, by algorytm nauczył się na zbiorze uczącym i mógł potem odpowiednio zakwalifikować odpowiedzi ze zbioru testowego do odpowiednich klastrów tak, aby pokrywały się one z grupą podaną w ostatnim pytaniu przez osobę wypełniającą ankietę.

Pisząc program używający algorytmów nauczania maszynowego i analizując powstałe klastry doszliśmy do wniosku, że niektóre pytania mogłyby być zmienione lub całkowicie inaczej sformułowane, inne z kolei zaczęły wydawać się mało istotne w analizie preferencji i przy próbie odpowiedzi na postawione na początku pytanie. Podczas analizowania odpowiedzi po ich sklastrowaniu zaczęliśmy się też zastanawiać, czy program WEKA rzeczywiście dokonał dobrej selekcji atrybutów i czy nie pominął takich, które mogłyby okazać się przydatne przy formowaniu klastrów, co z kolei mogłoby zaważyć na ich ostatecznym wyglądzie. Te atrybuty, które program pozostawił, też nie były wystarczająco zróżnicowane, jak np. rozmiar grupy, w której ktoś się wybiera lub czy dana osoba pije. Oznacza to zatem, że przy wyborze innych atrybutów

moglibyśmy uzyskać inne wyniki i inaczej zorganizowane grupy, nie oznacza to jednak, że wyniki algorytmu byłyby lepsze czy dokładniejsze.