Dokumentacja projektowa

Analiza wyników studentów i czynników na nie wpływających

Projekt - Algorytmy Eksploracji Danych

Kierunek: Informatyka

Członkowie zespołu:
Michał Tarnawa
Wojciech Grzywocz
Michał Żarłok

Gliwice, 2025/2026

Spis treści

0.1 Geneza zbioru i jego struktura

Analiza zbioru danych z serwisu Kaggle: Students Grading Dataset.

Zbiór ten zawiera informacje o wynikach studentów oraz czynnikach, które mogą mieć na nie wpływ. Są to:

- **Student_ID**: ID Studenta - Unikalny identyfikator każdego studenta.
- **First_Name**: Imię - Imię studenta.
- **Last_Name**: Nazwisko - Nazwisko studenta.
- **Email**: Email - Adres e-mail kontaktowy.
- **Gender**: Płeć - Płeć: Mężczyzna (Male), Kobieta (Female), Inna (Other).
- **Age**: Wiek - Wiek studenta.
- **Department**: Wydział/Kierunek - Kierunek studiów studenta.
- **Attendance (%)**: Frekwencja (%) - Procentowa frekwencja na zajęciach (0–100%).
- **Midterm_Score**: Wynik z egzaminu śródsemestralnego (0–100).
- **Final_Score**: Wynik z egzaminu końcowego (0–100).
- **Assignments_Avg**: Średnia z zadań domowych -(0–100).
- **Quizzes_Avg**: Średnia z kartkówek/kolowiów (0–100).
- **Participation_Score**: Ocena z aktywności (0–10).
- **Projects_Score**: Ocena z projektów - (0–100).
- **Total_Score**: Łączny wynik -
- **Grade**: Ocena końcowa - Ocena literowa (A, B, C, D, F).
- **Study_Hours_per_Week**: Średnia liczba godzin nauki w tygodniu.
- **Extracurricular_Activities**: Udział w zajęciach pozalekcyjnych (Tak/Nie).
- **Internet_Access_at_Home**: Dostęp do internetu w domu (Tak/Nie).
- **Parent_Education_Level**:Najwyższy poziom wykształcenia rodziców (Brak, Szkoła średnia, Licencjat, Magister, Doktorat/PhD).

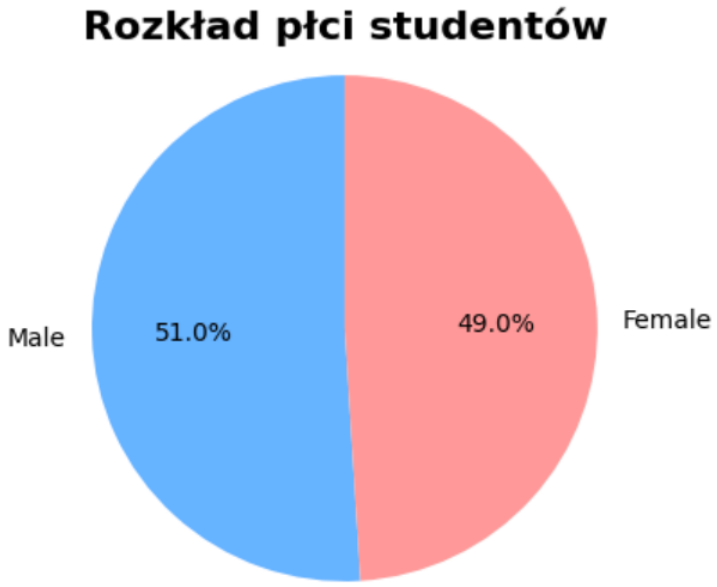
- **Family_Income_Level**:Poziom dochodów rodziny (Niski, Średni, Wysoki).
- **Stress_Level (1–10)**: Samoocena poziomu stresu (1: niski – 10: wysoki).
- **Sleep_Hours_per_Night**: Średnia liczba godzin snu na dobę.

Baza danych posiada 5000 rekordów i nie zawiera brakujących wartości (null).

W analizie naszego zbioru usunięto następujące kolumny: **Student_ID**, **First_Name**, **Last_Name**, **Email**, gdyż zawierają one wartości tekstowe, których nie można sensownie przekształcić do dalszej analizy (nie wnoszą wkładu w modele matematyczne).

0.2 Analiza danych i ich wplyw

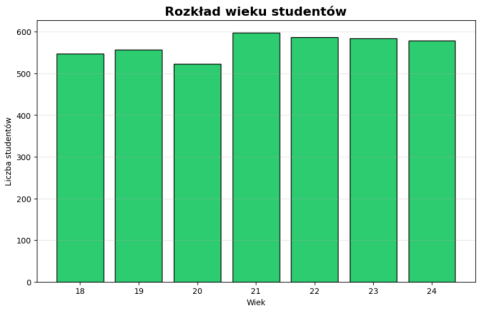
0.2.1 Analiza struktury płci



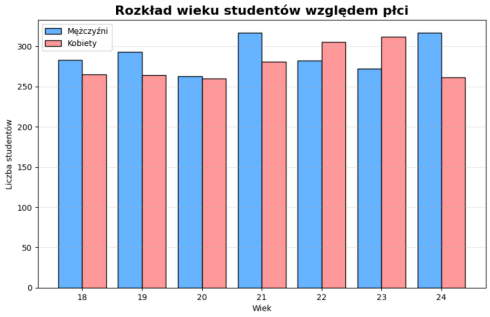
Rysunek 1: Wykres kołowy ukazując rozkład płci w analizowanym zbiorze

Analizując rysunek 1 widzimy, że mamy prawie równomierny podział, gdzie mężczyźni stanowią 51% badanej populacji, a kobiety 49%, co jest wartością podobną, do propocji przy urodzeniu.

0.2.2 Analiza struktury wieku



Rysunek 2: Rozkład wieku studentów

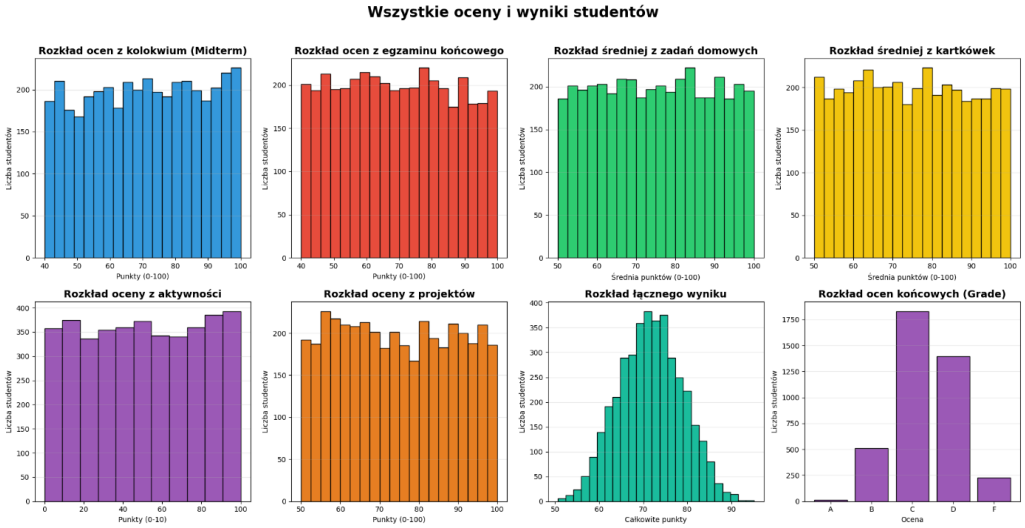


Rysunek 3: Rozkład wieku względem płci

Analizując Rysunek 2, widzimy, że dane obejmują osoby w wieku od 18 do 24 lat. Rozkład jest wygląda na równomierny, a liczebność w każdej grupie wiekowej jest podobna i zamyka się w granicach 500–600 osób. Rozkład przypomina rozkład jednostajny (nie jest to teza, tylko hipoteza).

Łącząc to z analizą Rysunku 3 widzimy, mężczyźni dominują w ilościach pod względem wieku. Wyjątkie są osoby w wieku 22 i 23, gdzie jest więcej kobiet. Największa przepaść jest wśród osób w wieku 24 lat.

0.2.3 Analiza wyników i ocen



Rysunek 4: Zbiorcze zestawienie rozkładów ocen i wyników studentów

Jak widzimy na Rysunku 4, rozkłady punktów z kategorii: **Midterm_Score**, **Final_Score**, **Assignments_Avg** oraz **Quizzes_Avg** posiadają bardzo podobny kształt. Są one stosunkowo równomierne, z niewielkim odchyleniem w stronę wyższych ocen (60–100 pkt). Oznacza to, że studenci są dobrze przygotowani.

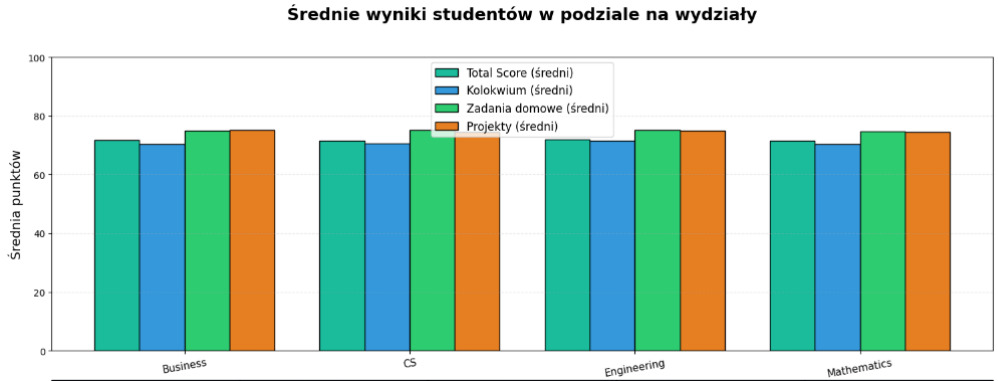
Natomiast wykres oceny z projektów wyróżnia się na tle pozostałych – posiada wyraźny szczyt w przedziale 80–100 punktów, co sugeruje, że z tą formą zaliczenia studenci radzili sobie najlepiej.

Wykres oceny z aktywności charakteryzuje się dość równomiernym rozkładem w zakresie 0–10, z lekkim nagromadzeniem przy wyższych wartościach. Wskazuje to na fakt, że większość studentów wykazywała umiarkowaną lub wysoką aktywność na zajęciach.

Wykres łącznego wyniku przypomina rozkład normalny. Szczyt wartości przypada na okolice 70–75 punktów, co zgadza się z wcześniej wyliczoną średnią wynoszącą $71,6$. Potwierdza to teorię, że wiele zmiennych losowych w przyrodzie bazuje na tym rozkładzie. Wykres ocen końcowych jest odwzorowaniem wykresu łącznego wyniku, zrzutowanym na 5 kategorii (ocen literowych).

Podsumowując, można przypuszczać, że osoby osiągające bardzo dobre wyniki (np. z egzaminów) wcale nie musiały otrzymać równie wysokich not z innych kategorii, takich jak aktywność czy projekty.

0.2.4 Analiza wyników w podziale na wydziały



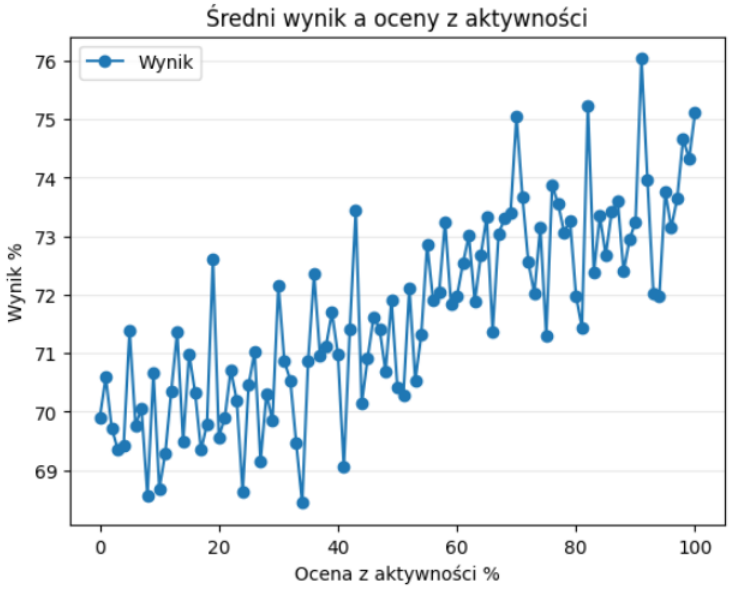
Wydział	Total Score (średnia) (min-max)	Kolokwium (średnia) (min-max)	Zadania domowe (średnia) (min-max)	Projekty (średnia) (min-max)
Business	71.8 (55–93)	70.4 (40–100)	75.0 (50–100)	75.1 (50–100)
CS	71.5 (51–95)	70.6 (40–100)	75.2 (50–100)	74.4 (50–100)
Engineering	71.9 (52–93)	71.4 (40–100)	75.1 (50–100)	75.0 (50–100)
Mathematics	71.4 (51–92)	70.4 (40–100)	74.8 (50–100)	74.4 (50–100)

Rysunek 5: Średnie wyniki studentów na poszczególnych wydziałach

Jak widać na Rysunku 5, niezależnie od wydziału wynikie są podobne. Studenci najlepiej radzą sobie z zadaniami domowymi oraz projektami. Najsłabiej wypada kolokwium, natomiast wynik łączny jest nieco wyższy od niego, co może świadczyć o potencjalnej mobilizacji w okresie egzaminów końcowych.

Analizując tabelę, nie zaobserwowane znaczących dysproporcji między wydziałami - różnice oscylują w granicach jednego punktu procentowego (p.p.). Należy jednak zauważyć, że najlepiej prezentują się wyniki studentów z wydziału Inżynierii, którzy osiągnęli najwyższe średnie w 3 z 4 analizowanych przez nas kategorii.

0.2.5 Wpływ aktywności na wynik końcowy

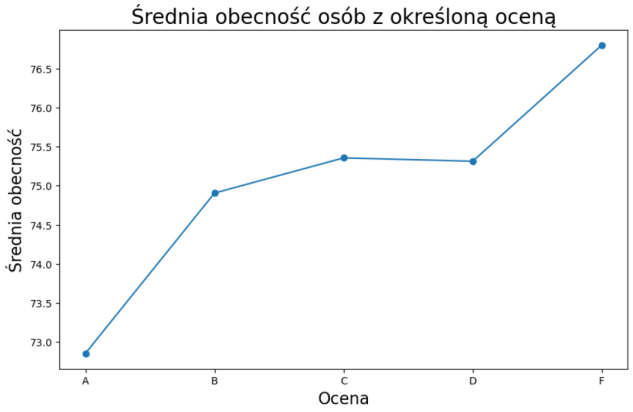


Rysunek 6: Średni wynik końcowy a ocena z aktywności

Wykres przedstawiony na Rysunku 6 nie jest regularny. Wartości rosną i maleją. Można jednak zauważyć zależność, iż im wyższa ocena z aktywności, tym lepsza ocena finalna. Gdyby poddać te dane regresji liniowej(pomijając fakt, wielkiego rozrzutu danych), współczynnik kierunkowy prostej byłby dodatni ($a > 0$), co świadczy o trendzie wzrostowym.

Rysunek 7: Średni wynik końcowy osób z określonym poziomem obecności

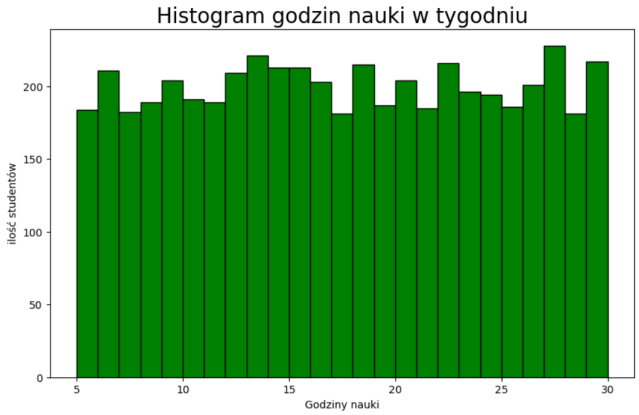
Na Wykresie 7 możemy zauważyć delikatny trend spadkowy, aby jednak dokładniej zbadać wpływ obecności na ocenę możemy odwrócić podejście i sprawdzić jak prezentuję się średnia obecność dla osób osiągających daną ocenę końcową.



Rysunek 8: Średni obecność dla osób z daną oceną końcową

Na Wykresie 8 możemy zauważyć, że osoby, które uzyskały najlepsze wyniki najmniej uczęszczały na zajęcia. Może to świadczyć o tym, że osoby zdolniejsze wolały uczyć się indywidualnie, a mając więcej czasu na naukę indywidualną, przełożyło się to na lepsze wyniki końcowe.

0.2.6 Analiza czasu poświęconego na naukę w tygodniu

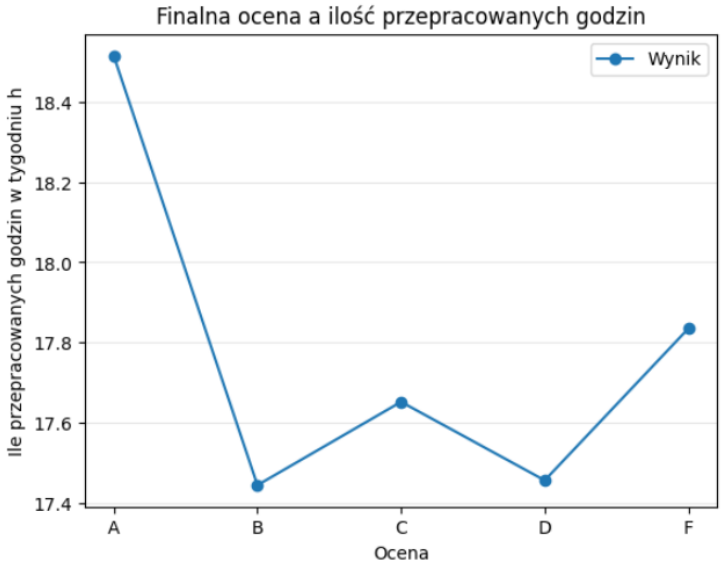


Rysunek 9: Histogram średniego czasu poświęconego na naukę w tygodniu

Analizując Rysunek 9, możemy zauważyć, że średni czas poświęcony na naukę w tygodniu ma rozkład dosyć równomierny. Możemy wyróżnić jednak

sporą grupę studentów, którzy na naukę poświęcają między 12 a 17 godzin. Najwyższy słupek na histogramie to z kolei osoby, które poświęcały między 27 a 28 godzin na naukę w tygodniu.
Dodatkowo, obliczony średni czas nauki w tygodniu (z uwzględnieniem wszystkich 5000 studentów) wyniósł: 17.52 godziny w tygodniu.
Mediana to natomiast: 17.4 godziny w tygodniu.

0.2.7 Zależność oceny końcowej od czasu nauki

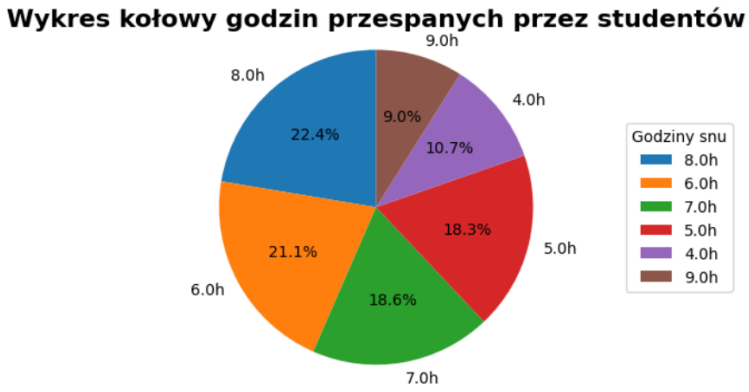


Rysunek 10: Finalna ocena a średnia ilość godzin nauki w tygodniu

Analizując Rysunek 10, można zauważyć, że osoby z najwyższą oceną (A) poświęcały na naukę najwięcej czasu. Co ciekawe, studenci z najgorszą oceną (F) uplasowali się na drugim miejscu pod względem liczby przepracowanych godzin.

Taka anomalia może sugerować dwie hipotezy: albo ankietowani w tej grupie podali nieprawdziwe dane, albo poświęcili dużo czasu na przyswojenie materiału, którego nie byli w stanie zrozumieć (nieefektywna nauka). Z kolei najmniej czasu na naukę poświęciły osoby, które uzyskały ocenę B.

0.2.8 Analiza długości snu



Rysunek 11: Wykres kołowy godzin przespanych przez studentów

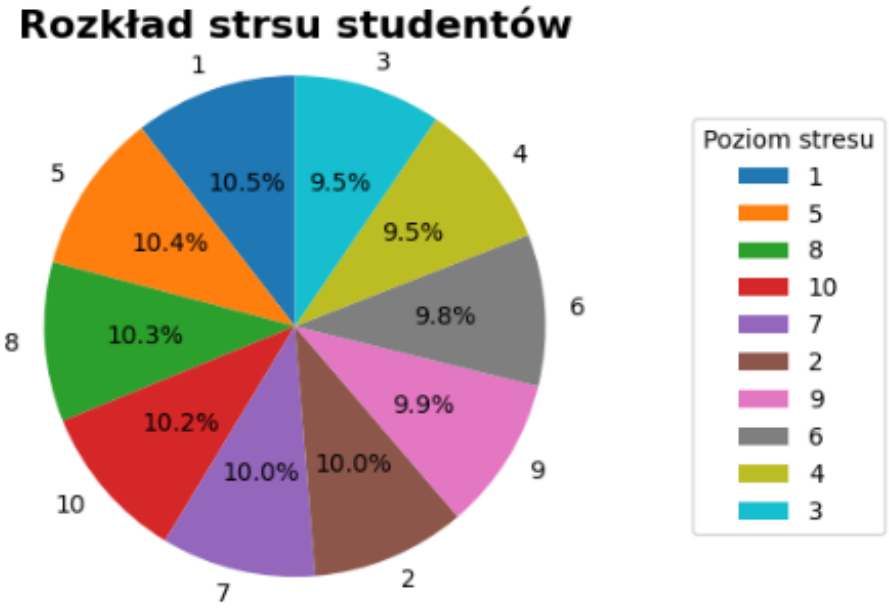
Analizując Rysunek 11, widzimy, że aż $68,6\%$ studentów śpi mniej niż wymagane 8 godzin. Co więcej, 29% badanych deklaruje, że przesypia jedynie 4 lub 5 godzin na dobę.

0.2.9 Wpływ długości snu na wyniki

Rysunek 12: Zależność między średnim wynikiem a liczbą godzin snu

Jak widać na Rysunku 12, najlepsze wyniki osiągają osoby, które śpią 8 godzin dziennie, co jest zgodne z ogólnymi zaleceniami. Najgorzej wypadają studenci, którzy śpią jedynie 4 godziny. Co istotne, dłuższy sen (powyżej 8 godzin) wcale nie przekłada się na polepszenie wyników.

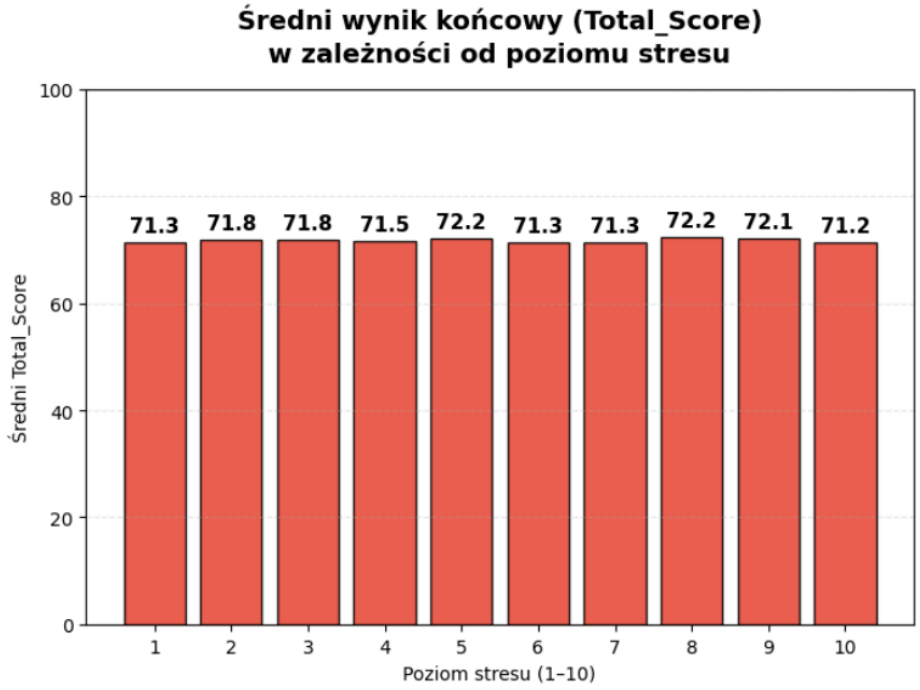
0.2.10 Analiza poziomu stresu



Rysunek 13: Rozkład poziomu stresu wśród studentów

Analizując Rysunek 13 (poniżej), widzimy, że każdy poziom stresu rozkłada się równomiernie. Udział każdej kategorii wynosi w przybliżeniu około 10%, co wskazuje na brak znaczących odstępstw w rozkładzie tej zmiennej. Należy podkreślić fakt, iż przedstawione dane, to samoocena poziomu stresu.

0.2.11 Wpływ stresu na wynik końcowy



Rysunek 14: Średni wynik końcowy w zależności od poziomu stresu

Jak widzimy na Rysunku 14, stres nie zawsze wpływa korzystnie na wyniki. Osoby najbardziej zestresowane osiągają najgorsze rezultaty, które są jednak tylko nieznacznie niższe od wyników osób nieodczuwających stresu.

Co ciekawe, najlepsze wyniki uzyskują studenci, których poziom stresu mieści się w przedziale 8-9 - w tej grupie średni wynik przekracza 72%. Wniosek może być taki, że lekki stres pobudza do działania.

1 Analiza PCA

1.1 Opis modelu matematycznego PCA

Niech $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ będzie wektorem p **standaryzowanych** zmiennych losowych. Celem analizy PCA jest znalezienie nowych, sztucznych zmiennych (głównych składowych) $GS_1, GS_2, \ldots, GS_p$ takich, że:

$$GS_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kp}X_p \qquad k = 1, 2, \ldots, p$$

gdzie współczynniki a_{kj} spełniają warunek:

- $\sum_{j=1}^{p} a_{kj}^2 = 1,$

Rozwiązanie tego problemu sprowadza się do wyznaczenia wartości własnych i wektorów własnych macierzy korelacji $\mathbf{R}$ (bądź – w ogólnym przypadku – macierzy kowariancji $\mathbf{S}$) standaryzowanych danych:

$$\mathbf{R}\mathbf{a}_k = \lambda_k \mathbf{a}_k \qquad k = 1, 2, \ldots, p$$

gdzie:

- $\mathbf{R}$ – macierz korelacji zmiennych X_j,
- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ – wartości własne (wariancje kolejnych składowych głównych),
- $\mathbf{a}_k = (a_{k1}, a_{k2}, \ldots, a_{kp})^T$ – k-ty wektor własny (wektor współczynników obciążenia).

Wynikiem jest zbiór p nowych, wzajemnie nieskorelowanych zmiennych $GS_1, GS_2, \ldots, GS_p$, których wariancje są równe kolejnym wartościom własnym:

$$Var(GS_k) = \lambda_k, \qquad \sum_{k=1}^{p} \lambda_k = p \quad (= \acute{s}lad(\mathbf{R}))$$

Po przeprowadzeniu analizy PCA spełnione są następujące własności:

- Główne składowe $GS_1, \ldots, GS_p$ są wzajemnie nieskorelowanymi zmiennymi sztucznymi będącymi liniowymi kombinacjami oryginalnych zmiennych $X_1, \ldots, X_p$.

- Każda kolejna składowa maksymalizuje wariancję niewyjaśnioną przez wszystkie poprzednie składowe.
- Wektory współczynników (wektory własne) są ortogonalne, co zapewnia nieskorelowanie składowych głównych.
- Wariancja k-tej składowej głównej GS_k jest równa k-tej co do wielkości wartości własnej λ_k macierzy korelacji (lub kowariancji) danych wejściowych.

1.2 Testy KMO oraz sferyczności Bartletta

Przed przeprowadzeniem analizy głównych składowych sprawdzono, czy macierz korelacji danych nadaje się do redukcji wymiarów. W tym celu wykonano dwa testy: Kaisera–Meyera–Olkina (KMO) oraz test sferyczności Bartletta.

1. Miara adekwatności próby Kaisera–Meyera–Olkina (KMO)

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2}$$

gdzie:

- r_{ij} – współczynnik korelacji Pearsona między zmiennymi i i j,
- p_{ij} – korelacja cząstkowa między zmiennymi i i j (przy kontrolowaniu wpływu pozostałych zmiennych).

Interpretacja wartości KMO (według Kaisera):

- $> 0{,}9$ – rewelacyjna,
- 0,8–0,9 – bardzo dobra,
- 0,7–0,8 – dobra,
- 0,6–0,7 – średnia,
- 0,5–0,6 – słaba,
- $< 0{,}5$ – nieakceptowalna (analiza głównych składowych nie jest zalecana).

2. Test sferyczności Bartletta

Testuje hipotezę zerową H_0: macierz korelacji populacyjnej jest macierzą jednostkową (wszystkie zmienne są nieskorelowane). Statystyka testowa ma przybliżony rozkład χ^2:

$$\chi^2 = -\left(n - \frac{2p+5}{6}\right)(n-1)\ln|\mathbf{R}|$$

gdzie:

- n – liczba obserwacji,
- p – liczba zmiennych,
- $|\mathbf{R}|$ – wyznacznik macierzy korelacji.

Przyjmujemy poziom istotności $\alpha = 0{,}05$. Jeżeli $p < 0{,}05$, odrzucamy hipotezę H_0 – zmienne są istotnie skorelowane, co przemawia za zasadnością przeprowadzenia analizy PCA.

W naszym przypadku uzyskano następujące wyniki:

- KMO = 0,41 – wartość poniżej 0,5, co wskazuje na nieakceptowalną adekwatność próby i sugeruje, że analiza głównych składowych nie jest zalecana,
- Test sferyczności Bartletta: $p = 0{,}000 < 0{,}05$ – odrzucamy hipotezę o jednostkowej macierzy korelacji, co przemawia za wykonaniem PCA.

Mimo sprzecznych wskazań testów (niski KMO i istotny test Bartletta), które nie zalecają wykonania PCA zdecydowano się na przeprowadzenie analizy PCA w celu badawczej redukcji liczby zmiennych i uproszczenia modelu.

1.3 Ustalenie liczby głównych składowych

Przed przystąpieniem do interpretacji wyników analizy PCA należy zdecydować, ile głównych składowych zachować do dalszych analiz i modelu.

1.3.1 Wykres osypiska

Wykres osypiska pozwala na wizualną oceną, ile składowych głównych warto zachować. Szuka się na nim osypiska miejsca wyraźnego załamania krzywej, po którym kolejne wartości własne maleją powoli.

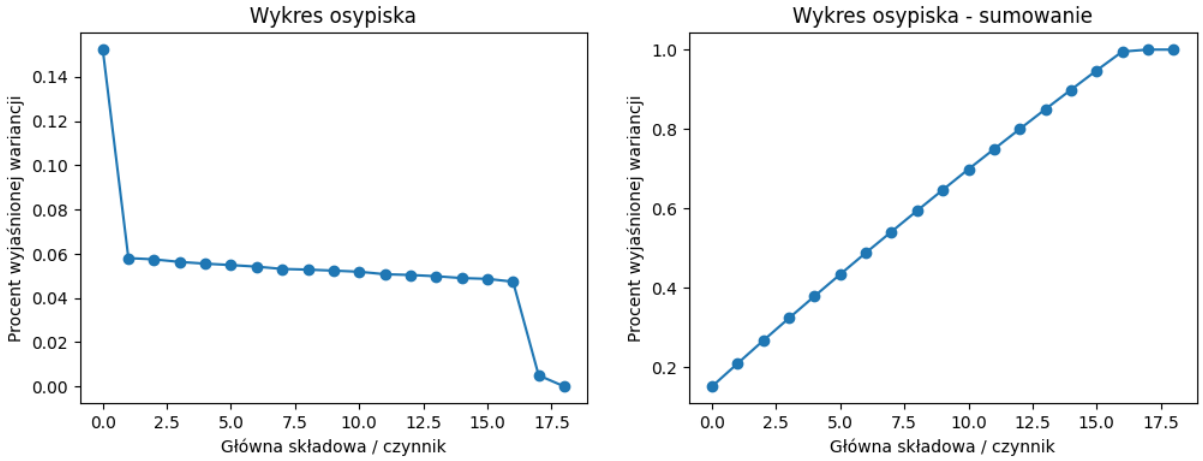
Metoda ta jest subiektywna – różne osoby mogą inaczej interpretować położenie osypiska – dlatego traktowana jest jako metoda wspomagająca.

Do stworzenia wykresu wykorzystuje się wartości własne λ_i macierzy korelacji oraz odpowiadające im udziały wyjaśnionej wariancji (procentowo i skumulowanie).

Tabela 1: Proporcja oraz skumulowany procent wyjaśnionej wariancji przez kolejne główne składowe

Główna składowa	Wartość własna λ_i	Wyjaśniona wariancja (%)	Skumulowana wariancja (%)
GS_1	2,8945	15,23	15,23
GS_2	1,1028	5,80	21,03
GS_3	1,0921	5,75	26,78
GS_4	1,0700	5,63	32,41
GS_5	1,0552	5,55	37,96
GS_6	1,0432	5,49	43,45
GS_7	1,0295	5,42	48,87
GS_8	1,0098	5,31	54,18
GS_9	1,0040	5,28	59,47
GS_{10}	0,9953	5,24	64,71
GS_{11}	0,9859	5,19	69,89
GS_{12}	0,9645	5,08	74,97
GS_{13}	0,9579	5,04	80,01
GS_{14}	0,9471	4,98	84,99
GS_{15}	0,9315	4,90	89,89
GS_{16}	0,9246	4,87	94,76
GS_{17}	0,9003	4,74	99,50
GS_{18}	0,0956	0,50	100,00
GS_{19}	0,0000	0,00	100,00
Razem	**19,0000**	**100,00**	

Ostatnia wartość własna jest bliska zeru (zaokrąglona do 0). Suma wszystkich wartości własnych wynosi dokładnie $\sum_{i=1}^{19} \lambda_i = 19{,}0000$, co jest równe liczbie zmiennych wejściowych (ślad macierzy korelacji = $p = 19$). Potwierdza to poprawność standaryzacji danych i wykonania analizy PCA – cała wariancja zbioru została w pełni wyjaśniona przez 19 ortogonalnych składowych.



Rysunek 15: Wykres osypiska (po lewej) oraz wykres skumulowanej wyjaśnionej wariancji (po prawej)

Na rysunku 15 możliwy uskok pojawia się pomiędzy 2. a 3. składową lub dopiero pod koniec wykresu (przy bardzo małych wartościach własnych). Ze względu na brak wyraźnego załamania na początku i bardzo powolny spadek wariancji zdecydowano się na zastosowanie dodatkowych kryteriów.

1.3.2 Kryterium Kaisera

Kryterium Kaisera (Kaiser–Guttman) polega na zachowaniu wszystkich składowych głównych, których wartość własna $\lambda_i > 1$.

Zgodnie z tym kryterium zachowano 9 pierwszych składowych głównych ($\lambda_i > 1$):

- GS_1 - 2,8945
- GS_2 - 1,1028
- GS_3 - 1,0921
- GS_4 - 1,0700

- GS_5 - 1,0552
- GS_6 - 1,0432
- GS_7 - 1,0295
- GS_8 - 1,0098
- GS_9 - 1,0040

Te 9 składowych wyjaśnia łącznie 59,47% całkowitej wariancji, redukując wymiarowość z 19 do 9 zmiennych (redukcja o ponad 52%), przy utracie ok. 40,53% informacji. Duża liczba składowych spełniających kryterium Kaisera (9 z 19) przy relatywnie niskiej wyjaśnionej wariancji (59,47%) potwierdza, że analizowany zbiór danych posiada duże rozproszeniem danych i brakuje w nim dominujących czynników

1.3.3 Kryterium procentowe

Kryterium to polega na wybraniu takiej liczby składowych, aby wyjaśnić założony, wysoki procent całkowitej wariancji (najczęściej 70–90% lub więcej).

Z tabeli 1 wynika, że: - zachowanie 15 składowych głównych daje 89,89% wyjaśnionej wariancji, - zachowanie tylko 3 składowych (sugerowane przez wykres osypiska) wyjaśnia zaledwie 26,78% wariancji – wartość tą uznano za zbyt niską.

1.3.4 Wniosek

Ze względu na trudności z wyznaczenia wartości skłądowych na wykresie osypiska oraz bardzo rozłożoną wariancję, zdecydowano się przyjąć kryterium Kaisera. Pozwala ono zachować 9 składowych głównych, które wyjaśniają prawie 60% wariancji przy znacznej redukcji liczby zmiennych (z 19 do 9). Jest to rozsądny kompromis między zachowaniem informacji a uproszczeniem modelu, zwłaszcza że wiele zmiennych wejściowych mogło zawierać błędy pomiaru.

1.4 Interpretacja głównych składowych

Ostatnim etapem analizy jest interpretacja dziewięciu wybranych głównych składowych na podstawie macierzy ładunków (współczynników wektorów własnych) – tabela 2.

Im większa wartość bezwzględna $|a_{ij}|$, tym silniejszy wpływ danej cechy na daną składową główną.

1.4.1 Macierz wartości wektorów własnych głównych składowych

Tabela 2: Wektory własne dziewięciu głównych składowych po PCA - pełna macierz

Cecha	GS_1	GS_2	GS_3	GS_4	GS_5	GS_6	GS_7	GS_8	GS_9
X_1	0.0000	-0.3637	-0.0650	0.3832	-0.2752	0.3035	0.0434	0.0652	0.0650
X_2	0.0000	0.0439	0.1277	-0.1665	0.2431	0.3137	0.4461	-0.2717	-0.2534
X_3	0.0000	-0.2162	-0.1332	0.4903	0.1417	-0.1127	-0.0138	0.1127	-0.1043
X_4	-0.0148	-0.1285	0.0614	-0.4046	-0.2324	-0.2953	0.4272	-0.1674	0.2443
X_5	0.1846	-0.2251	0.4644	-0.0872	-0.2830	-0.1009	-0.2356	0.2049	-0.0617
X_6	0.3439	0.0000	0.1502	0.1888	0.0427	-0.0885	-0.1472	-0.3633	0.4581
X_7	0.1941	0.0105	-0.3838	0.0418	-0.1873	0.2187	0.2563	0.0000	0.0631
X_8	0.1198	-0.1421	-0.1481	-0.2799	0.4306	0.2320	-0.3673	-0.1645	-0.1318
X_9	0.1092	0.2649	-0.1147	-0.3995	-0.2358	0.3138	-0.2019	0.2665	0.0774
X_{10}	0.3463	0.0848	-0.1300	0.0666	0.1538	-0.1723	0.3417	0.2067	-0.4248
X_{11}	0.5830	0.0000	0.0112	0.0000	0.0000	-0.0182	0.0000	0.0000	0.0000
X_{12}	0.5683	0.0000	0.0118	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X_{13}	0.0000	0.0329	0.3224	0.0789	-0.3241	-0.0991	0.1651	-0.2320	-0.3095
X_{14}	0.0000	0.4270	0.3219	0.0830	0.0741	0.0979	0.0324	0.0739	0.1415
X_{15}	0.0000	0.0000	0.4741	0.0921	0.3145	0.1059	-0.0580	-0.1005	-0.1465
X_{16}	0.0000	0.3525	-0.0252	0.1565	-0.4207	0.2303	-0.2202	-0.1972	-0.3975
X_{17}	0.0000	0.3533	0.1450	0.2425	0.1314	0.2467	0.2738	0.3332	0.3436
X_{18}	0.0125	-0.3494	0.2388	-0.1707	0.0132	0.1867	0.0841	0.5224	-0.0952
X_{19}	0.0168	-0.3156	0.0958	0.0000	-0.0741	0.5310	0.1371	-0.2723	0.1305

UWAGA!!!, wartości własne 0.000 zostały zaokrąglone przez interpreter Pythona. Analizą 2 zajmiemy się w podsumowaniu.

1.5 Podsumowanie i interpretacja wyników analizy PCA

Analiza głównych składowych (PCA) została przeprowadzona w celu redukcji wymiarowości zbioru danych składającego się z 19 zmiennych. Poniżej przedstawiono podsumowanie uzyskanych wyników oraz najważniejsze wnioski.

1.5.1 Ocena przydatności danych do analizy

Test sferyczności Bartletta ($p < 0{,}001$) potwierdził występowanie istotnych powiązań między zmiennymi. Z kolei bardzo niska wartość miary KMO (0,41) wskazuje, że korelacje te są słabe i mają charakter raczej parami niż w całym zbiorze. Mimo niekorzystnych wskazań testu KMO, w celach eksploracyjnych i redukcyjnych zdecydowano się na wykonanie analizy PCA.

1.5.2 Efektywność redukcji wymiarowości

Na podstawie kryterium Kaisera wybrano **9 głównych składowych**, co pozwoliło:

- zredukować liczbę zmiennych wejściowych z 19 do 9,
- zachować **59,47%** całkowitej wariancji oryginalnego zbioru danych.

Porównanie z innymi kryteriami:

- Wykres osypiska sugerował ewentualne załamanie w okolicy 3 składnika – ograniczenie do 3 wymiarów dałoby jednak zaledwie ok. 26-27% wyjaśnionej wariancji,
- Kryterium procentowe (np. zachowanie 90% wariancji) wymagałoby pozostawienia 13-15 składowych, co oznaczałoby bardzo słabą redukcję wymiarowości i niewielkie uproszczenie modelu.

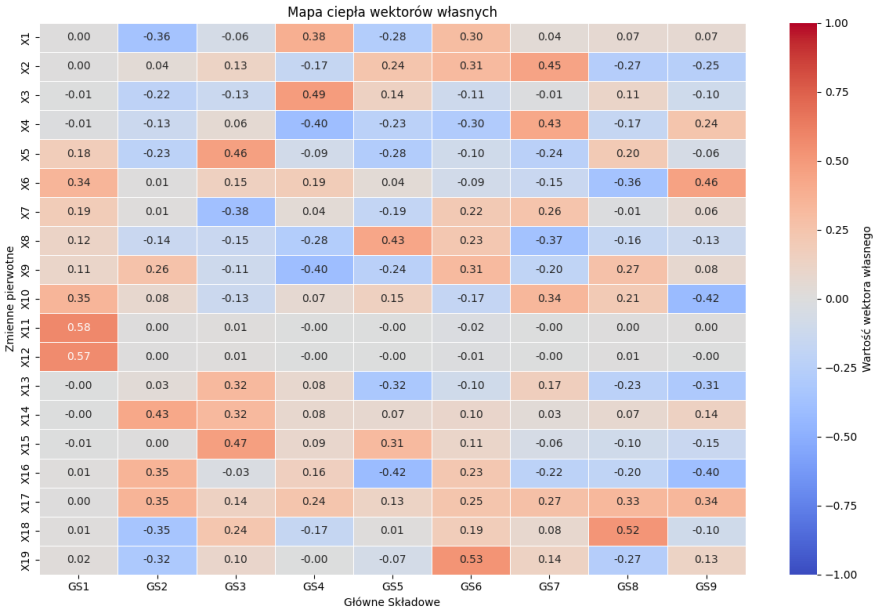
Kryterium Kaisera okazało się zatem rozsądnym kompromisem między stopniem redukcji a zachowaniem istotnej części informacji.

1.5.3 Struktura powiązań – analiza współczynników

Do interpretacji zawartości nowo powstałych zmiennych wykorzystano macierz wektorów własnych w postaci mapy ciepła (rys. 16) oraz w formie tabelarycznej (tabela 2).

Na podstawie analizy macierzy ładunków wyciągnięto następujące wnioski:

1. **GS_1 – wyraźny klaster dwóch zmiennych.** Pierwsza i najsilniejsza składowa jest zdominowana przez zmienne $\mathbf{X}_{11}$ i $\mathbf{X}_{12}$ (ładunki $> 0{,}56$). Brak istotnych powiązań tej pary z pozostałymi składowymi wskazuje na jej odrębność w grupie danych.
2. **GS_2** Składowa ta charakteryzuje się wysokimi ładunkami dodatnimi dla zmiennych $\mathbf{X}_{14}$ i $\mathbf{X}_{17}$ oraz ujemnymi dla $\mathbf{X}_1$ i $\mathbf{X}_{18}$, co sugeruje kontrast między tymi dwiema zmiennymi.
3. **Rozproszenie i słaba grupowość w wyższych składowych.** W składowych od GS_3 do GS_9 większość zmiennych ma istotne ładunki tylko w jednej składowej (przykłady: GS_6 zdominowana przez X_{19}, GS_8 przez X_{18}). Wynik ten potwierdza wcześniejsze wskazania niskiej wartości testu KMO – zmienne w analizowanym zbiorze grupują się w niewiele grup, a większość powiązań ma charakter słaby lub parami.



Rysunek 16: Mapa ciepła współczynników składowych głównych

Wniosek ogólny: Zastosowanie PCA umożliwiło wydobycie najsilniejszej zależności w danych (klaster X_{11}–X_{12}) oraz pewnych kontrastów (GS_2). Jednocześnie zachowano jedynie 60% wariancji przy redukcji do 9 wymiarów, co wskazuje na duże nieuporządkowanie danych, ich słabą grupowość i wysoki poziom tzw. szumu.

2 Klasyfikacja: Drzewo decyzyjne (Decision Tree)

2.1 Opis zastosowanej metody

Cel: Celem tego etapu analizy jest automatyczna predykcja oceny studenta (*Grade*) wyrażonej w skali 1-5 (odpowiedniki ocen F, D, C, B, A) na podstawie szerokiego spektrum cech wejściowych.

Charakterystyka metody: Do realizacji zadania wybrano algorytm **drzewa decyzyjnego**. Jest to jedna z najbardziej intuicyjnych metod uczenia nadzorowanego, która dokonuje podziału zbioru danych na coraz bardziej jednorodne podgrupy (liście). Kluczowe dla algorytmu jest pojęcie *zysku informacyjnego*, oparte na miarach takich jak:

- **Entropia** – mierzy stopień nieuporządkowania informacji w węźle. Al-

gorytm dąży do jej minimalizacji.

- **Wskaźnik Giniego** – określa prawdopodobieństwo błędnej klasyfikacji losowego elementu ze zbioru, jeśli zostałby on sklasyfikowany zgodnie z rozkładem klas w danym węźle.

2.2 Implementacja i optymalizacja modelu

Proces implementacji w środowisku Python z wykorzystaniem biblioteki `scikit-learn` obejmował następujące kroki:

1. **Podział danych:** Zbiór został podzielony na część treningową i testową w proporcji 80:20, co pozwala na rzetelną ocenę zdolności generalizacyjnych modelu.
2. **Tuning hiperparametrów:** Wykorzystano metodę `GridSearchCV` z 5-krotną walidacją krzyżową (`cv=5`). Przeszukiwano przestrzeń parametrów:
   - `criterion`: {`'gini'`, `'entropy'`}
   - `max_depth`: {5, 10, 12, 15, 20}
   - `min_samples_split`: {2, 5, 10, 20}
   - `min_samples_leaf`: {1, 2, 5}
3. **Finalny model:** Algorytm wyłonił optymalną konfigurację: kryterium entropii, maksymalna głębokość 15, minimalna liczba próbek w liściu równa 5.

2.3 Wyniki i interpretacja

Model oparty na pełnym zbiorze cech uzyskał **dokładność (accuracy) na poziomie 0,7650 (76,5%)**.

Analiza raportu klasyfikacji:

- Dla klas 2.0 (D) oraz 3.0 (C) model osiąga najwyższe wartości *F1-score* (odpowiednio 0,76 i 0,80), co świadczy o dobrym rozpoznawaniu średnich wyników.
- Klasa 5.0 (ocena A) posiada bardzo niską czułość (*recall* = 0,33), co wynika z małej liczności tej grupy w zbiorze danych (*support* = 3).

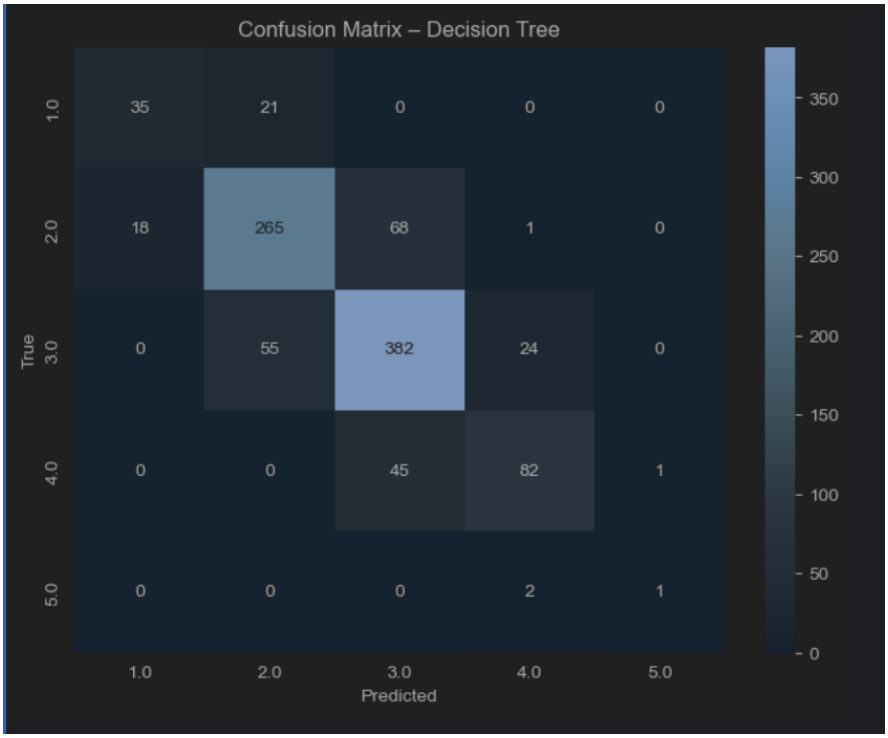
```
Accuracy: 0.7650

Classification report:
              precision    recall  f1-score   support

         1.0       0.66      0.62      0.64        56
         2.0       0.78      0.75      0.76       352
         3.0       0.77      0.83      0.80       461
         4.0       0.75      0.64      0.69       128
         5.0       0.50      0.33      0.40         3

    accuracy                           0.77      1000
   macro avg       0.69      0.64      0.66      1000
weighted avg       0.76      0.77      0.76      1000
```

Rysunek 17: Raport klasyfikacji



Rysunek 18: Macierz pomyłek

2.4 Szczegółowa analiza macierzy pomyłek

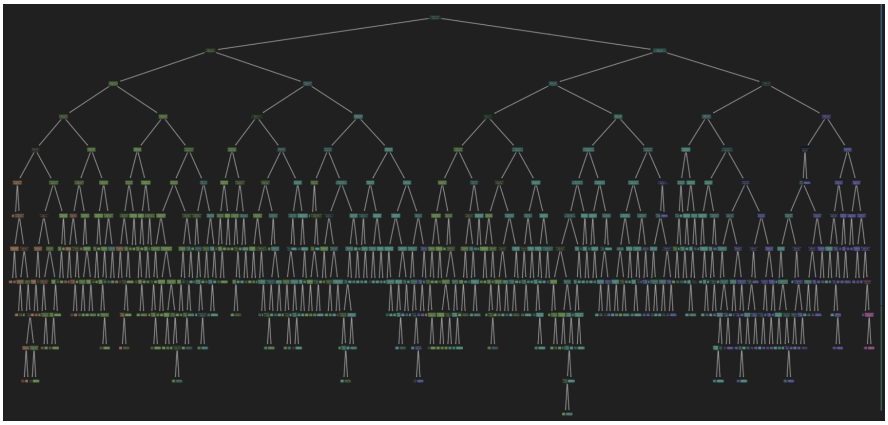
Macierz pomyłek, przedstawiona na rysunku 18, pozwala na analizę struktury błędów popełnianych przez klasyfikator. Wiersze macierzy reprezentują klasy rzeczywiste, podczas gdy kolumny odpowiadają decyzjom podjętym przez model.

Analiza pozwala na sformułowanie następujących wniosków:

- **Dominacja przekątnej:** Najwyższe wartości liczbowe znajdują się na głównej przekątnej macierzy (382 dla klasy 3 oraz 265 dla klasy 2), co świadczy o tym, że model poprawnie klasyfikuje **większość** przypadków.
- **Błędy sąsiedztwa:** Zdecydowana większość błędnych klasyfikacji dotyczy klas będących obok siebie. Przykładowo:
  - Dla klasy **4** (B), aż 45 przypadków zostało błędnie zaklasyfikowanych jako **3** (C), ale żaden nie został pomylony z 1 czy 2.
  - Dla klasy **1** (F), 21 przypadków zostało zaklasyfikowanych jako **2** (D), a nie jako ocena wyższa.

  Świadczy to o tym, że model, nawet gdy popełnia błąd, – myli się zazwyczaj „o jedną ocenę".
- **Przechylenie w stronę klas dominujących:** Wyraźnie widoczna jest klasyfkiacja do klas najliczniejszych (2 i 3).
  - Klasa **4** jest częściej mylona z niższą oceną 3 (45 przypadków) niż poprawnie rozpoznawana jako 5 (1 przypadek).
  - Klasa **5** (najmniej liczna, 3 obserwacje) została poprawnie rozpoznana tylko raz, a w dwóch przypadkach jako 4.

Wnioski: Macierz pomyłek ujawnia, że model klasyfikuje ostrożnie – unika predykcji ocen 1 i 5, częściej przepisuje predykcje niepewne do bezpiecznych, liczniejszych klas.



Rysunek 19: Wizualizacja najlepszego drzewa decyzyjnego. Całe dostępne w załączniku

2.5 Podsumowanie

Wnioski: Uzyskany wynik accuracy nie jest bardzo wysoki, jednak jest on zadowalający, biorąc pod uwagę dużą liczbę rekordów oraz cech w bazie danych.

Porównanie z PCA: Przeprowadzona próba klasyfikacji na danych po redukcji wymiarowości metodą PCA wykazała spadek dokładności do ok. 60%. Sugeruje to, że nieliniowe zależności w oryginalnych cechach są zbyt złożone, by proste składowe główne mogły je w pełni oddać bez utraty kluczowej informacji.

3 Regresja wieloraka (Multiple Regression)

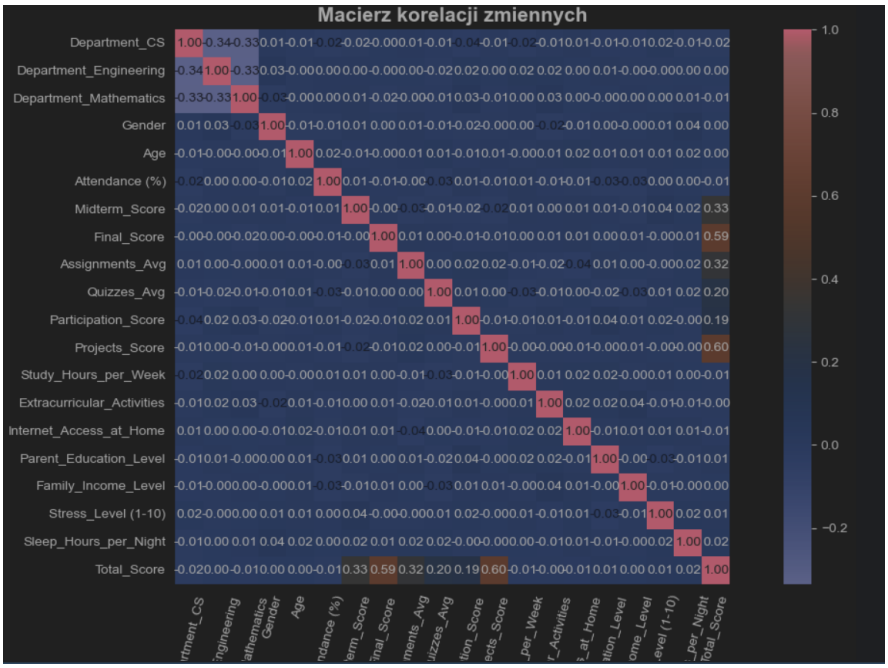
3.1 Opis zastosowanej metody

Cel: Celem regresji jest oszacowanie wartości zmiennej ciągłej `Total_Score` (wynik punktowy w skali 0–100) na podstawie zmiennych objaśniających.

Charakterystyka metody: Regresja liniowa zakłada istnienie zależności postaci $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon$. Metoda ta pozwala nie tylko na predykcję, ale przede wszystkim na analizę istotności poszczególnych zmiennych (tzw. wag modelu), co daje wgląd w to, które aspekty życia studenta najsilniej korelują z jego wynikiem.

3.2 Implementacja i przygotowanie danych

- **Kodowanie kategorialne:** Zmienna `Department` została przetworzona metodą *one-hot encoding*, co eliminuje problem sztucznej hierarchii wydziałów.
- **Standaryzacja:** Zastosowano `StandardScaler` dla wszystkich zmiennych niezależnych. Jest to kluczowe w regresji liniowej, ponieważ umożliwia bezpośrednie porównywanie wartości bezwzględnych współczynników (β) – większa wartość bezwzględna oznacza silniejszy wpływ danej cechy.
- **Metryki oceny:** Wykorzystano współczynnik determinacji (R^2) oraz średni błąd kwadratowy (MSE).



Rysunek 20: Macierz korelacji cech

3.3 Wyniki i analiza współczynników

Model uzyskał nietypowo wysokie wyniki: $R^2 = 1{,}0$ oraz błąd **MSE** $\approx 1{,}29 \times 10^{-28}$.

Istotność zmiennych (najwyższe współczynniki):

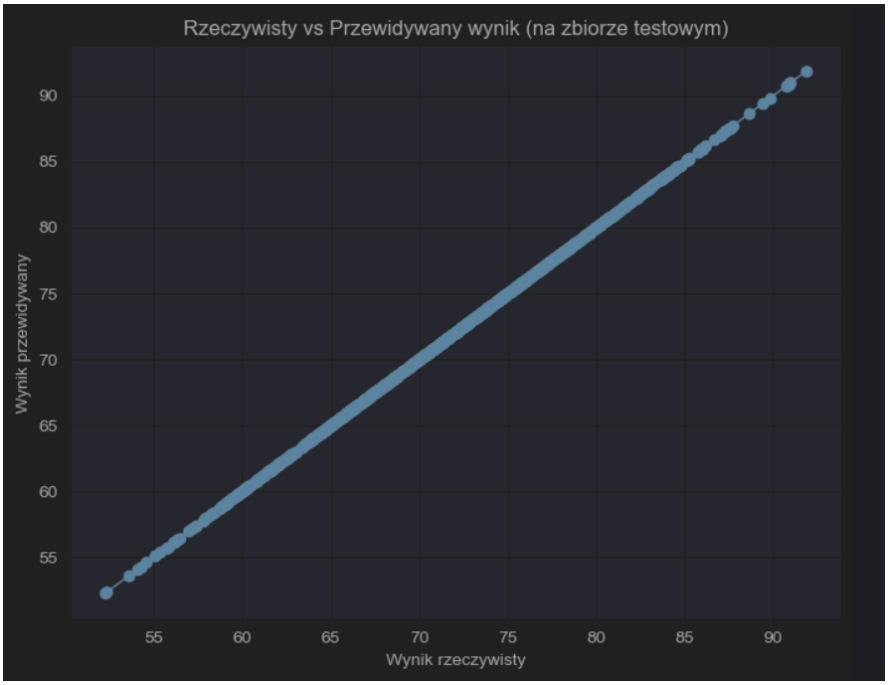
Cecha	**Współczynnik (waga)**
`Projects_Score`	4,3352
`Final_Score`	4,2508
`Midterm_Score`	2,6172
`Assignments_Avg`	2,1683
`Quizes_Avg`	1,45
`Participation_Score`	1,44
`Gender, Age, Stress_Level`	$\approx 10^{-15}$

Tabela 3: Współczynniki modelu regresji liniowej – wybrane najważniejsze cechy

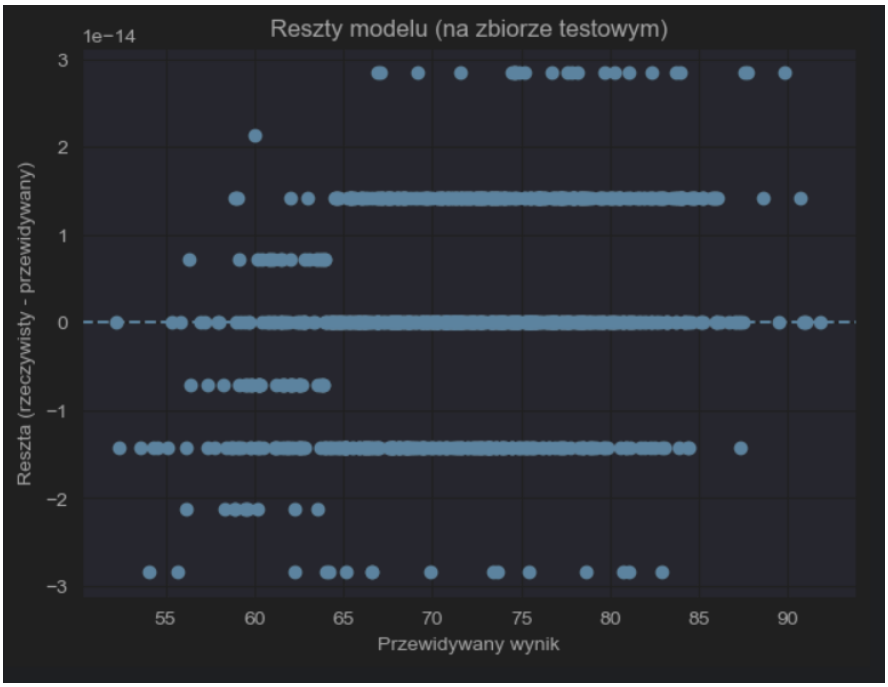
Wnioski końcowe:

1. **Charakter danych:** Wynik $R^2 = 1{,}0$ jednoznacznie wskazuje, że zmienna `Total_Score` jest w tym zbiorze danych obliczana jako liniowa kombinacja ocen cząstkowych (*Final, Midterm, Projects, Assignments*). Model regresji bezbłędnie odtworzył ten matematyczny wzór.

2. **Wpływ czynników zewnętrznych:** Bardzo niskie wartości bezwzględne współczynników dla zmiennych takich jak płeć, wiek, poziom stresu czy wykształcenie rodziców (rzędu 10^{-15}) wskazują, że w obliczu twardych wyników testów czynniki demograficzne i socjodemograficzne nie mają praktycznie żadnego bezpośredniego przełożenia na ostateczną punktację w tym konkretnym zbiorze danych.



Rysunek 21: Wykres przewidywanych vs rzeczywistych wartości na zbiorze testowym



Rysunek 22: Wykres reszt modelu na zbiorze testowym

4 Reguły asocjacyjne

4.1 Opis metody

Reguły asocjacyjne stosuje się w celu wykrywania zależności między poszczególnymi elementami zbiorów transakcyjnych. W przypadku tej metody kolejność elementów w transakcji nie ma znaczenia. Służą one m.in. do budowy systemów rekomendacyjnych.

Trzy podstawowe miary skuteczności reguł asocjacyjnych to:

- **Support** – określa, w jakiej części transakcji występuje dany zbiór elementów (częstość występowania),
- **Confidence** – wskazuje, jak często reguła jest prawdziwa (prawdopodobieństwo, że jeśli występuje lewa strona, to wystąpi także prawa strona),
- **Lift** – pokazuje, czy elementy reguły występują razem częściej, niż wynikałoby to z przypadku (wartość > 1 oznacza pozytywną zależność, < 1 – negatywną, $= 1$ – brak zależności).

4.2 Implementacja

Do implementacji algorytmu wykorzystano bibliotekę `mlxtend`, a konkretnie funkcje `apriori` (do wyszukiwania zbiorów częstych według zadanego progu support) oraz `association_rules` (do generowania reguł asocjacyjnych spełniających wybrane kryteria).

Ustawiono następujące progi:

- minimalny **support** = 0,2 (zbiory częste występujące w co najmniej 20% transakcji),
- minimalny **confidence** = początkowo 0,7 – nie znaleziono żadnych reguł, dlatego obniżono próg do 0,5.

4.3 Wyniki

zbiory częste:	support	
0	0.2040	(Bachelor's)
1	0.2528	(Business)
2	0.4614	(C)
3	0.2478	(CS)
4	0.3520	(D)
5	0.2548	(Engineering)
6	0.4898	(Female)
7	0.3278	(High)
8	0.3374	(Low)
9	0.5102	(Male)
10	0.2000	(Master's)
11	0.2446	(Mathematics)
12	0.3348	(Medium)
13	0.4976	(Noextra)
14	0.4960	(Nointernet)
15	0.2050	(None)
16	0.2024	(PhD)
17	0.5024	(Yesextra)
18	0.5040	(Yesinternet)
19	0.2258	(Female, C)
20	0.2356	(C, Male)
21	0.2354	(C, Noextra)
22	0.2240	(C, Nointernet)
23	0.2260	(C, Yesextra)
24	0.2374	(C, Yesinternet)
25	0.2380	(Female, Noextra)
26	0.2402	(Female, Nointernet)
27	0.2518	(Female, Yesextra)
28	0.2496	(Female, Yesinternet)
29	0.2596	(Male, Noextra)
30	0.2558	(Male, Nointernet)
31	0.2506	(Yesextra, Male)
32	0.2544	(Male, Yesinternet)
33	0.2520	(Noextra, Nointernet)
34	0.2456	(Noextra, Yesinternet)
35	0.2440	(Yesextra, Nointernet)
36	0.2584	(Yesextra, Yesinternet)

Rysunek 23: Zbiory częste w danych

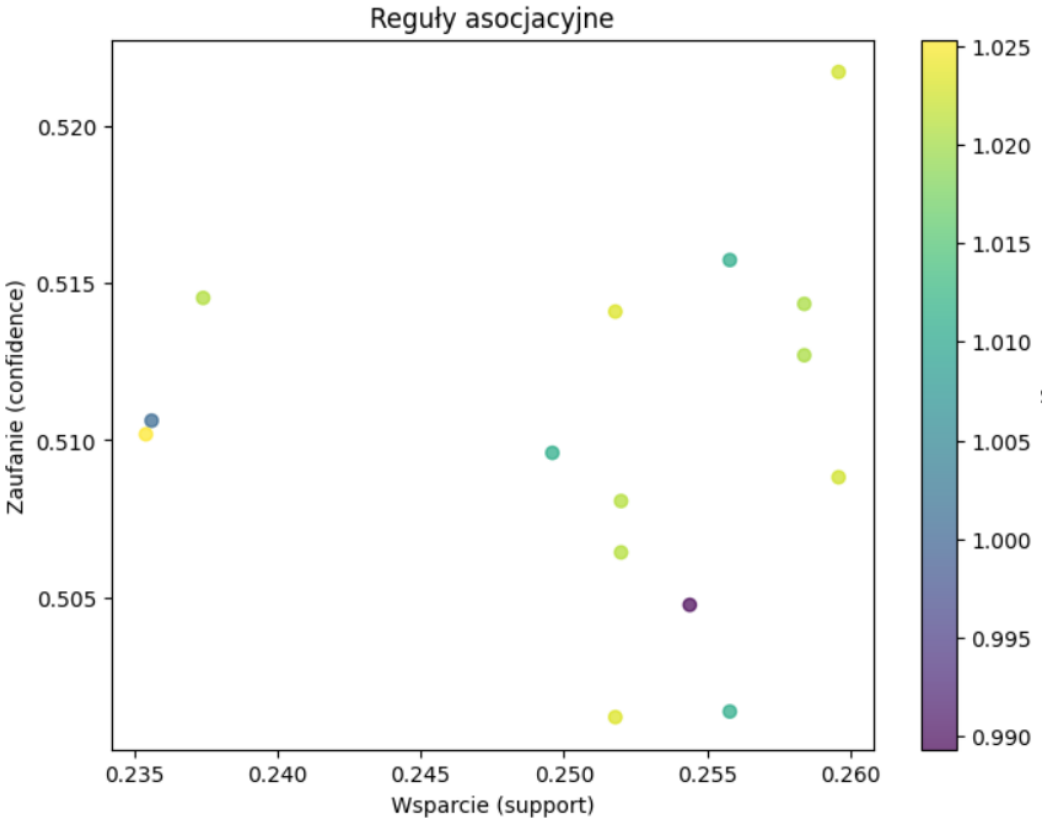
Jak widać na rysunku 23, w danych występuje duża liczba zbiorów częstych.

Powstałe z nich reguły prezentują się następująco

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(C)	(Male)	0.4614	0.5102	0.2356	0.510620	1.000823
1	(C)	(Noextra)	0.4614	0.4976	0.2354	0.510186	1.025294
2	(C)	(Yesinternet)	0.4614	0.5040	0.2374	0.514521	1.020875
3	(Female)	(Yesextra)	0.4898	0.5024	0.2518	0.514087	1.023263
4	(Yesextra)	(Female)	0.5024	0.4898	0.2518	0.501194	1.023263
5	(Female)	(Yesinternet)	0.4898	0.5040	0.2496	0.509596	1.011103
6	(Male)	(Noextra)	0.5102	0.4976	0.2596	0.508820	1.022548
7	(Noextra)	(Male)	0.4976	0.5102	0.2596	0.521704	1.022548
8	(Male)	(Nointernet)	0.5102	0.4960	0.2558	0.501372	1.010831
9	(Nointernet)	(Male)	0.4960	0.5102	0.2558	0.515726	1.010831
10	(Yesinternet)	(Male)	0.5040	0.5102	0.2544	0.504762	0.989341
11	(Noextra)	(Nointernet)	0.4976	0.4960	0.2520	0.506431	1.021030
12	(Nointernet)	(Noextra)	0.4960	0.4976	0.2520	0.508065	1.021030
13	(Yesextra)	(Yesinternet)	0.5024	0.5040	0.2584	0.514331	1.020498
14	(Yesinternet)	(Yesextra)	0.5040	0.5024	0.2584	0.512698	1.020498

Rysunek 24: Reguły asocjacyjne

Można zauważyć, że wszystkie wygenerowane reguły asocjacyjne mają wartość confidence nieco wyższą od 0,5, co nie jest szczególnie dobrym wynikiem. Pozytywnym aspektem jest natomiast wartość lift większa od 1 we prawie wszystkich przypadkach – oznacza to, że współwystępowanie elementów nie jest przypadkowe.



Rysunek 25: Wykres najważniejszych informacji dotyczących reguł asocjacyjnych

Wykres punktowy (Rysunek 25) ilustruje rozkład wygenerowanych reguł w przestrzeni parametrów: wsparcie (oś X), zaufanie (oś Y) oraz lift.

Analiza pozwala na odkrycie następujących wniosków:

- **Wysoka koncentracja reguł:** Wszystkie punkty na wykresie są skupione w bardzo wąskim obszarze wartości (wsparcie $0{,}235 - 0{,}260$ oraz zaufanie $0{,}500 - 0{,}525$). Brak punktów odstających od tego zakresu.
- **Rozrzut**Punkty są rozrzucone chaotycznie, świadczy to o niskiej korelacji między wspraciem a zaufaniem.
- **Niska wartość informacyjna (Lift):** Wartości wskaźnika Lift mieszczą się w przedziale $0{,}99 - 1{,}025$. Dominacja kolorów pochodzących ze skali ukazuje fakt, że analizowane wykryte reguły mają charakter trywialny.

4.4 Podsumowanie

Z uzyskanych wyników wynika, że w analizowanym zbiorze danych nie występują zbyt pewne (silne) reguły asocjacyjne – trudno jest jednoznacznie powiązać jedną zmienną z drugą. Prawdopodobnie wynika to z faktu, że oceny (które teoretycznie miały największą szansę na utworzenie reguł) są zależne od wielu czynników numerycznych, co znacząco osłabia wpływ zmiennych nienumerycznych.

5 Wnioski

- W naszym zbiorze danych nie ma reguł asocjacyjnych o dużej pewności. Prawdopodobnie wynika to z tego że oceny które miały największą szansę stworzyć regułę asocjacyjną są zależne od zbyt wielu czynników numerycznych co zmniejsza wpływ nie numerycznycych.
- Analiza PCA pokazała słabe korelacje między zmiennymi. Większość zmiennych sztucznych zależy od par prawdziwych zmiennych
- Klasyfikacja ukazała, że wybór model PCA, który zachował około 60% wariancji zbioru oryginalnego znacząco go osłabił zmiejszając precyzję klasyfikatora z ok. 75% do 60%

6 Bibliografia

1. Internet