**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**

**SPECIALIZAREA INFORMATICĂ ECONOMICĂ**

****

**DEZVOLTARE SOFTWARE PENTRU ANALIZA DATELOR**

***Analiza Stilului de Viață al Studenților***

**Coordonator:**

Prof. univ. dr. Vințe Claudiu

**Studenți:**

Pavaloi Natalia-Sofia

Păduraru Elena

Popescu Paul

Radu Mihaela-Daniela

**BUCUREȘTI**

**2025**

# CUPRINS

[**CUPRINS 1**](#_8ik9t22vdi6p)

[**INTRODUCERE 1**](#_h0c6tzld5von)

[**1. Descrierea setului de date și a variabilelor 2**](#_w4iienfi3p5j)

[1.1. Caracteristici 4](#_p04k84wi5mmf)

[1.1.1. Granularitate zilnică 4](#_y5jc6w9y6c89)

[1.1.2. Volumul datelor 4](#_fruu6asov12d)

[1.1.3. Aplicabilitate 4](#_afbfohbs5797)

[**2. Descrierea observațiilor 5**](#_msn32jcfm9ie)

[**3. Justificarea metodelor de analiză alese: PCA și Discriminantă 6**](#_x64xpdi49soa)

[3.1. De ce am ales PCA pentru proiectul nostru: 6](#_ve2n8uz64bcb)

[3.2. De ce analiza discriminantă este potrivită pentru proiectul nostru: 6](#_o5g0k6u8uusi)

[**4. Analiza Componentelor Principale (PCA) 6**](#_6cfxa0y37v9q)

[4.1. Pregătirea și standardizarea datelor 6](#_jkswfan3jes1)

[4.2. Matricea de corelații 7](#_dhcf0cwipoo3)

[4.3. Valorile proprii și selectarea componentelor principale 7](#_33d8eegfu2j3)

[4.4. Comunalitățile 9](#_fbi48xbbywfu)

[4.5. Încărcările factorilor (Factor Loadings) 9](#_mabrqao7k0w8)

[4.6. Cercul corelațiilor 9](#_ta7nbpx341dh)

[4.7. Calitatea Reprezentării Observațiilor (scoruri) 9](#_qdlijzrxnu8o)

[4.8. Contribuția observațiilor 9](#_hvjngfcyvk6h)

[**5. Analiza discriminantă 9**](#_vi770bgri7jd)

# 

# INTRODUCERE

Stilul de viață al studenților este un subiect complex, iar analiza datelor asociate poate oferi perspective interesante despre cum putem găsi, ca studenți, un echilibru între studiu, odihnă, activități extracurriculare și socializare. Acest proiect explorează obiceiurile de viață ale studenților și modul în care acestea influențează performanța academică și nivelul de stres. Am ales să lucrăm cu acest set de date numit *Student Lifestyle Dataset* deoarece reflectă direct provocările și obiceiurile zilnice ale unei categorii care ne reprezintă – studenții. Într-o lume în care sănătatea mintală și succesul educațional sunt din ce în ce mai interconectate, înțelegerea factorilor care determină aceste rezultate poate aduce o diferență semnificativă în înțelegerea comportamentului studenților.

Setul de date analizat include variabile numerice care descriu principalele aspecte ale stilului de viață al studenților, precum Ore\_Studiu, Ore\_Activități\_Extracurriculare, Ore\_Somn, Ore\_Social, Ore\_Activ (activitate fizică), GPA (media academică) și Nivel\_Stres. Aceste variabile ne permit să investigăm relațiile dintre obiceiurile zilnice ale studenților și rezultatele lor academice, oferindu-ne șansa de a înțelege mai bine factorii care contribuie la succesul sau dificultățile acestora. Analiza datelor ne poate ajuta să identificăm comportamente care sunt fie benefice, fie dăunătoare, contribuind astfel la găsirea unor strategii de îmbunătățire a stilului nostru de viață. Setul utilizat conține date dintr-un an academic, din August 2023 până în Mai 2024, recoltate într-un procentaj destul de mare în India.

În cadrul acestui proiect, am utilizat două metode esențiale de analiză a datelor: analiza componentelor principale (PCA) și analiza discriminantă. Analiza componentelor principale ne-a permis să reducem complexitatea datelor, identificând cele mai relevante dimensiuni care explică variațiile majore între studenți. Această analiză este utilă pentru a evidenția care aspecte ale stilului de viață au cel mai mare impact asupra rezultatelor academice și nivelului de stres. Am aplicat criteriile Kaiser și Cattel pentru a determina numărul optim de componente principale, iar comunalitățile ne-au ajutat să înțelegem cât din variația fiecărei variabile este explicată de aceste componente. Pe de altă parte, analiza discriminantă ne-a permis să identificăm variabilele care contribuie cel mai mult la diferențierea între grupuri de studenți, oferind o perspectivă clară asupra factorilor care determină apartenența la un anumit grup. În plus, această metodă a fost utilizată pentru a construi un model predictiv, capabil să clasifice noi observații în grupurile existente, validând astfel separarea și relevanța acestora.

# Descrierea setului de date și a variabilelor

Setul de date ales reprezintă o colecție de informații privind obiceiurile zilnice ale studenților. Acesta oferă o bază solidă pentru analiza relației dintre comportamentele zilnice și rezultatele lor academice sau nivelul de stres. În total, setul de date conține 2000 de rânduri, fiecare reprezentând răspunsurile unui student, și include variabile numerice și categorice care reflectă diferite aspecte ale stilului de viață al acestora. Toate variabilele sunt măsurate la nivel zilnic, ceea ce înseamnă că fiecare valoare reprezintă timpul dedicat unei activități pe parcursul unei zile obișnuite. (*https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset/data*)

Variabilele principale incluse în setul de date sunt:

*Ore\_Studiu:* Această variabilă indică numărul mediu de ore alocat zilnic studiului. Este un indicator important pentru a putea evalua cât de mult timp investesc studenții în activitățile academice într-o zi. Analiza acestei variabile ne permite să determinăm dacă orele de studiu sunt corelate pozitiv cu performanța academică (GPA) sau dacă există un punct în care investiția excesivă de timp poate duce la stres sau epuizare.

*Ore\_Activități\_Extracurriculare:* Reprezintă timpul alocat activităților extracurriculare. Aceste activități includ hobby-uri, participarea la cluburi studențești, voluntariat sau alte angajamente sociale și recreative. Analizarea acestei variabile ajută la înțelegerea echilibrului dintre viața academică și activitățile extracurriculare, care pot avea un impact pozitiv asupra sănătății mintale a unui student.

*Ore\_Somn:* Somnul este un aspect esențial al unui stil de viață sănătos, iar această variabilă măsoară numărul mediu de ore de odihnă pe zi. Lipsa somnului sau somnul insuficient poate contribui semnificativ la creșterea nivelului de stres și la scăderea performanței academice, iar analiza acestei variabile poate evidenția aceste legături.

*Ore\_Social:* Timpul petrecut zilnic interacționând cu prietenii, colegii sau familia este reprezentat prin această variabilă. Socializarea este un element important al vieții studențești și ajută la menținerea echilibrului emoțional și totodată, la reducerea stresului.

*Ore\_Activ:* Activitățile fizice sunt măsurate prin numărul de ore dedicate zilnic acestora. Activitatea fizică, cum ar fi exercițiile sau sportul, are efecte pozitive bine documentate și dovedite asupra sănătății mintale și fizice, iar analiza acestei variabile ne permite să înțelegem dacă exercițiile contribuie la reducerea nivelului de stres sau la îmbunătățirea altor aspecte ale vieții studenților.

*Nivel\_Stres:* Această variabilă ordinală măsoară nivelul de stres al studenților pe o scală de la 1 la 3, unde 1 indică un nivel redus de stres (*low*), 2 un nivel moderat (*moderate*), iar 3 un nivel ridicat (*high*). Este un indicator important al stării mintale a studenților și permite investigarea modului în care stresul este influențat de diversele obiceiuri zilnice.

*GPA (Grade Point Average):* GPA-ul este o măsură standardizată a performanței academice a studenților, calculată ca o medie ponderată a notelor obținute. Această variabilă este una dintre cele mai importante, deoarece reflectă rezultatele directe ale stilului de viață și ale obiceiurilor studenților. Analizând relația dintre GPA și alte variabile, putem identifica factorii care influențează succesul academic.

Formula generală utilizată pentru calcularea GPA:

, unde:

* reprezintă punctele corespunzătoare notei obținute la cursul *i*, conform unei scale standardizate. *A = 4.0, B = 3.0, C = 2.0, D = 1.0, F = 0.0.* Astfel, GPA va avea un interval de note între 0 și 4, cu 4 fiind cea mai bună notă.
* ​ reprezintă numărul de credite asociate cursului *i*.
* *n* reprezintă numărul total de cursuri.

## 1.1. Caracteristici

### 1.1.1. Granularitate zilnică

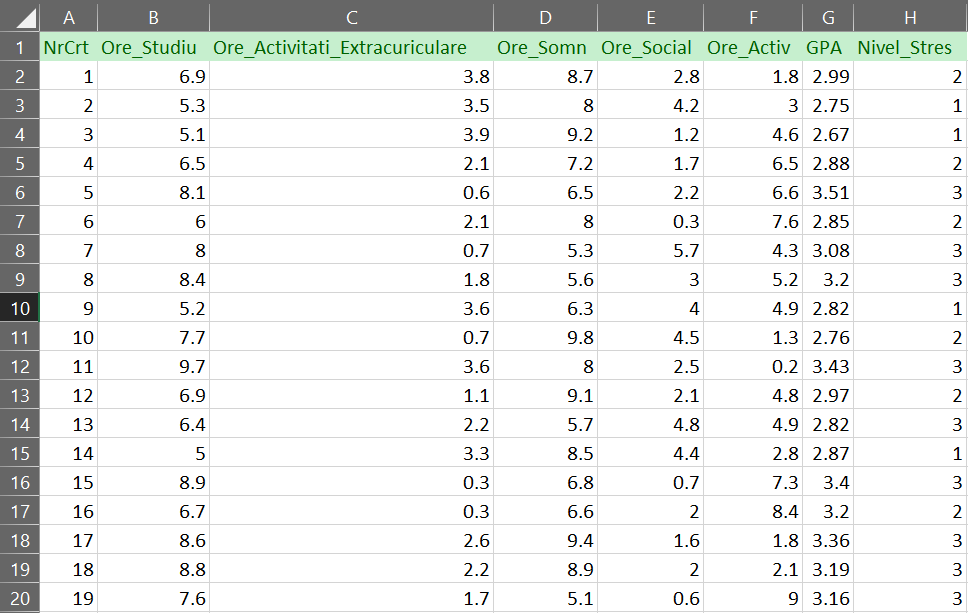
Faptul că toate variabilele sunt raportate la o singură zi oferă o imagine clară și detaliată a obiceiurilor studenților. Astfel putem identifica tipare zilnice, cum ar fi obiceiul de a dormi insuficient, perioadele lungi de studiu sau lipsa activității fizice. De asemenea, datele zilnice sunt mai ușor de interpretat și analizat, deoarece reflectă activitățile obișnuite ale studenților fără influențe pe termen lung care ar putea distorsiona rezultatele.

### 1.1.2. Volumul datelor

Fiind un set de date ce conține 2000 de rânduri, îl face suficient de mare pentru a realiza analize robuste și pentru a detecta tipare relevante. Acest volum permite aplicarea unor tehnici avansate, cum ar fi analiza componentelor principale (PCA), pentru a identifica variabilele cele mai importante care explică variația în stilurile de viață și performanțele studenților.

### 1.1.3. Aplicabilitate

Analiza acestui set de date oferă oportunitatea de a înțelege obiceiurile sănătoase care contribuie la succesul academic și la reducerea nivelului de stres al studenților, evidențiind comportamentele pozitive și impactul acestora asupra bunăstării generale. Prin corelarea variabilelor, pot fi identificați factori de risc precum lipsa somnului sau neglijarea activităților fizice, care pot duce la creșterea stresului și la scăderea performanței academice.

****

# Descrierea observațiilor

Studenții alocă, în medie, *7.48* ore pe zi pentru **studiu**, ceea ce reflectă angajament academic, deși variațiile între indivizi sunt notabile, între 5 și 10 ore. **Activitățile extracurriculare** primesc, în medie, *1.99* ore pe zi, indicând o implicare moderată, iar **somnul** se situează la *7.5* ore pe zi, ceea ce sugerează un nivel general rezonabil de odihnă. **Socializarea** ocupă aproximativ *2.7* ore pe zi, însă există cazuri în care acest timp este nul, iar **activitățile fizice** se întind, în medie, pe *4.32* ore pe zi, cu un maxim de 13 ore, reflectând o diversitate în priorități.

**Performanța academică**, măsurată prin GPA, are o medie de *3.11* (pe o scară de la 0 la 4), ceea ce indică rezultate bune la nivel general, dar cu unele variații între studenți. **Nivelul de stres**, cu o valoare medie de *2.36* (pe o scală de 1-3), arată că majoritatea studenților experimentează stres moderat spre ridicat, influențat de echilibrul dintre studiu, odihnă și alte activități zilnice.

Rezultatele indică o imagine complexă a stilului de viață al studenților. În timp ce orele alocate studiului arată un angajament academic puternic, distribuția timpului pentru alte activități sugerează diferențe semnificative în priorități. Timpul limitat acordat socializării și activităților extracurriculare poate aduce o presiune ridicată asupra performanței academice, dar și posibile compromisuri în menținerea echilibrului personal. În același timp, orele dedicate somnului par, în medie, să fie suficiente, deși variațiile notabile ar putea semnala cazuri în care odihna este insuficientă. Nivelul moderat spre ridicat de stres subliniază provocările întâmpinate de studenți în gestionarea timpului și armonizarea tuturor activităților.

# Justificarea metodelor de analiză alese: PCA și Discriminantă

## 3.1. De ce am ales PCA pentru proiectul nostru:

PCA ajută la reducerea complexității datelor, simplificând analiza prin extragerea celor mai importante componente care explică variațiile principale din setul de date. Aceasta permite identificarea variabilelor care au cea mai mare influență asupra performanței academice și nivelului de stres, eliminând redundanța cauzată de variabilele corelate și ușurând interpretarea relațiilor.

## 3.2. De ce analiza discriminantă este potrivită pentru proiectul nostru:

Analiza discriminantă este potrivită pentru proiectul nostru deoarece identifică variabilele care diferențiază cel mai bine grupurile de studenți și permite construirea unui model predictiv pentru clasificarea observațiilor noi. De asemenea, validează dacă grupurile existente sunt bine separate, oferind concluzii clare și aplicabile în practică.

# Analiza Componentelor Principale (PCA)

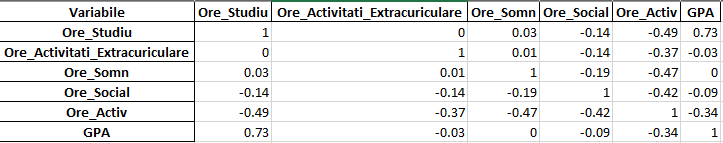
Analiza Componentelor Principale permite reducerea dimensiunii setului de date, astfel putem să identificăm componentele principale care rețin cea mai mare parte a variației de date. Acest lucru ne ajută să simplificăm analiza, să eliminăm redundanțele și să evidențiem relațiile esențiale dintre variabile. În contextul acestui proiect, această analiză ne ajută să înțelegem cum obiceiurile zilnice ne influențează performanța academică și nivelul de stres, astfel reușind să identificăm factorii cheie.

## 4.1. Pregătirea și standardizarea datelor

Prima etapă a fost selectarea variabilelor numerice din setul de date, deoarece această analiză funcționează în parametri optimi cu date numerice și continuu distribuite. A urmat standardizarea acestor variabile, o etapă esențială în PCA, deoarece diferențele de scară între variabile (ex. orele de somn vs. GPA) pot influența în mod disproporționat rezultatele. Standardizarea asigură că fiecare variabilă are media 0 și deviația standard 1.

## 4.2. Matricea de corelații

Aceasta oferă o perspectivă asupra relațiilor liniare dintre variabilele din setul de date. Fiecare valoare din matrice indică gradul de asociere între două variabile, variind între -1 (corelație negativă perfectă) și 1 (corelație pozitivă perfectă).

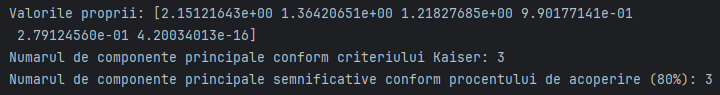


Analizând matricea de covarianță, putem observa câteva relații semnificative între variabilele care ne definesc stilul de viață. Relația pozitivă puternică dintre **Ore\_Studiu** și **GPA** sugerează că studenții care dedică mai mult timp studiului obțin rezultate mai bune, evidențiind ideea că efortul este esențial pentru succesul academic.

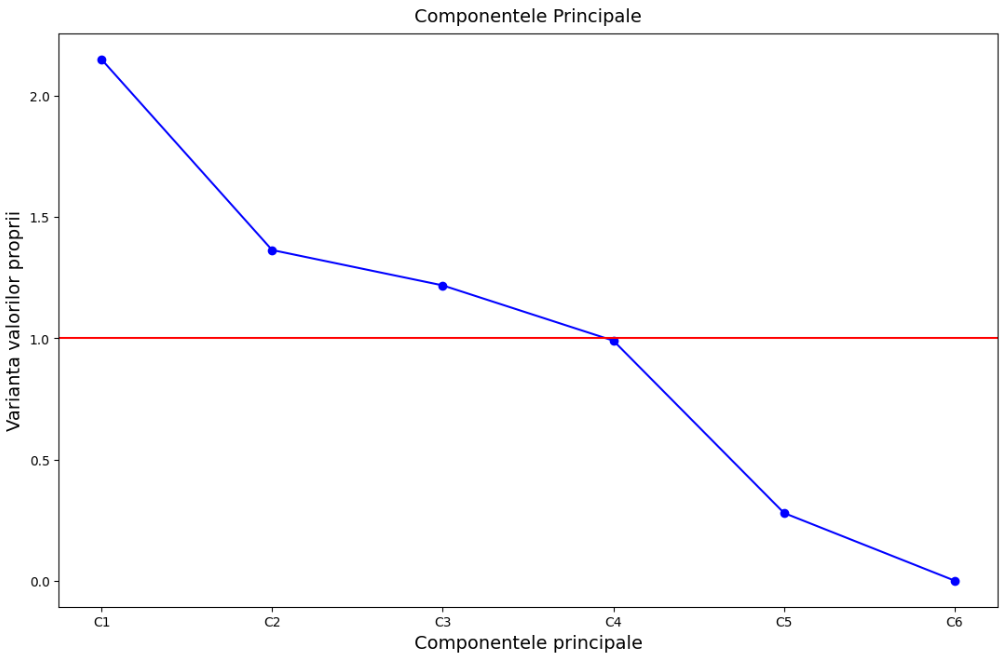
În schimb, relația negativă moderată dintre **Ore\_Activ** și **Ore\_Studiu** arată că studenții care petrec mai mult timp în activități fizice tind să acorde mai puțin timp studiului. Covarianțele mici dintre **Ore\_Somn** și alte variabile, cum ar fi **Ore\_Studiu** sau **Ore\_Social**, sugerează că somnul rămâne relativ constant, fiind mai puțin influențat de activitățile zilnice. Totuși, relația negativă moderată dintre **Ore\_Somn** și **Ore\_Activ** indică faptul că o implicare crescută în activități fizice poate reduce timpul alocat somnului, ceea ce poate avea implicații asupra nivelului de stres sau stării generale de sănătate. De asemenea, relațiile foarte slabe dintre **Ore\_Activități\_Extracurriculare** și alte variabile indică faptul că aceste activități sunt independente și nu influențează semnificativ alte aspecte.

## 4.3. Valorile proprii și selectarea componentelor principale

Valorile proprii joacă un rol esențial în determinarea importanței fiecărei componente principale. Fiecare valoare proprie reprezintă variabilitatea explicată de o componentă principală și ne ajută să identificăm care componente contribuie cel mai mult la înțelegerea datelor. Pentru a selecta componentele relevante, au fost utilizate două criterii principale: criteriul Kaiser și criteriul Cattel.



Conform criteriului Kaiser, care reține componentele cu valori proprii mai mari sau egale cu 1, primele trei componente principale au valori proprii de 2.15, 1.36 și 1.21. Acestea explică împreună peste 80% din variabilitatea totală a datelor, indicând că ele captează majoritatea informațiilor esențiale. Această selecție ne permite să reducem complexitatea dimensională a setului de date, păstrând doar acele dimensiuni care contribuie semnificativ la analiza ulterioară.



Criteriul Cattel, bazat pe graficul Scree, oferă o confirmare vizuală a selecției componentelor. Graficul evidențiază valorile proprii în ordine descrescătoare, iar punctul de „cotitură” apare după a treia componentă principală, sugerând că primele trei componente sunt cele care explică o proporție semnificativă a variabilității. După acest punct, variația explicată de componentele ulterioare scade considerabil, ceea ce justifică excluderea lor.

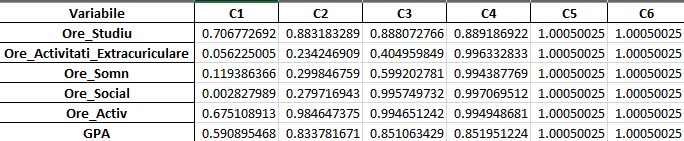
Astfel, prima componentă (C1) explică cea mai mare parte a variației, urmată de C2 și C3, în timp ce restul componentelor contribuie nesemnificativ. Graficul Scree și analiza valorilor proprii confirmă împreună că reducerea dimensiunii datelor la trei componente principale este o decizie optimă, permițând păstrarea esenței din punct de vedere informațional fără a compromite calitatea analizei.

## 4.4. Comunalitățile

Comunalitățile măsoară cât de bine sunt reprezentate variabilele inițiale de componentele principale. Ele indică proporția variabilității fiecărei variabile capturată de aceste componente, fiind esențiale pentru a evalua relevanța variabilelor în analiza PCA.

Variabilele Ore\_Studiu, Ore\_Activ și GPA sunt explicate semnificativ de primele 2-3 componente principale. Spre exemplu, Ore\_Studiu are o comunalitate de 0.707 după prima componentă și ajunge la 0.883 după a doua, ceea ce indică o contribuție consistentă la structura generală a datelor. Ore\_Activități\_Extracuriculare și Ore\_Social necesită mai multe componente pentru a fi complet explicate, sugerând o variabilitate mai complexă sau o contribuție mai mică la primele dimensiuni principale. Aceste variabile ating comunalitatea maximă abia după 4 componente.

Toate variabilele sunt complet explicate (comunalitate = 1) după includerea tuturor componentelor. Această concluzie susține reducerea dimensionalității datelor fără pierderi semnificative.



## 4.5. Încărcările factorilor (Factor Loadings)

Încărcările factorilor reprezintă corelațiile dintre variabilele inițiale și componentele principale, astfel oferind o imagine clară a modului în care fiecare variabilă contribuie la variația explicată de componentele selectate. Aceste valori sunt importante pentru a înțelege structura datelor și pentru a identifica factorii care au cel mai mare impact asupra fiecărei componente.

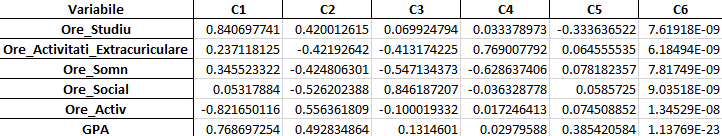
Componenta principală 1 (*C1*) reflectă dimensiunea academică, fiind influențată pozitiv de *Ore\_Studiu* (0.841) și *GPA* (0.769), dar negativ de *Ore\_Activ* (-0.822), sugerând un compromis între studiile academice și activitățile fizice.

Componenta principală 2 (*C2*) evidențiază o relație pozitivă între *Ore\_Activ* (0.556) și *GPA* (0.493), dar negativă cu *Ore\_Social* (-0.526), indicând o opoziție între performanța academică și socializare.

Componenta principală 3 (*C3*) este dominată de *Ore\_Social* (0.847), sugerând un focus pe activitățile sociale, în timp ce *Ore\_Somn* (-0.547) și *Ore\_Activ* (-0.100) au contribuții negative.

Componenta principală 4 (*C4*) este influențată puternic de *Ore\_Activități\_Extracuriculare* (0.769), reflectând implicarea studenților în aceste activități, de multe ori în detrimentul somnului (*Ore\_Somn* -0.628).

Primele trei componente principale (C1, C2, C3) explică cele mai importante direcții de variație în stilul de viață al studenților, concentrându-se pe dimensiuni academice, sociale și extracurriculare. Componentele C5 și C6 contribuie marginal și nu sunt semnificative pentru interpretare.



## 4.6. Cercul corelațiilor

Cercul corelațiilor ne oferă o reprezentare grafică a relațiilor dintre variabilele inițiale și primele două componente principale (C1 și C2). Fiecare variabilă este reprezentată printr-un vector, iar lungimea și poziția vectorului în cerc indică contribuția sa la componentele respective și relația cu celelalte variabile.

Variabilele reprezentate de vectori apropiați de marginea cercului indică o contribuție semnificativă la primele două componente principale (C1 și C2), fiind bine reprezentate în analiza PCA. Vectorii orientați în direcția unei axe principale sugerează o corelație puternică cu componenta respectivă, așa cum este cazul variabilelor Ore\_Studiu și GPA care sunt aliniate cu dimensiunea explicată de C1, probabil legată de performanța academică. În schimb, variabilele plasate aproape de centrul cercului au o contribuție redusă la primele componente și ar putea fi mai relevante pentru alte dimensiuni (C3, C4 etc.). Unghiurile dintre vectori indică relațiile dintre variabile: unghiuri mici sugerează corelații pozitive, unghiurile perpendiculare indică lipsa corelației, iar unghiurile opuse reflectă corelații negative.

## 4.7. Calitatea Reprezentării Observațiilor

Graficul prezintă calitatea reprezentării observațiilor pentru un subset de 50 de studenți, selectați pe baza celor mai mari valori de reprezentare pe prima componentă principală (C1). Fiecare rând din tabel corespunde unei observații individuale, iar valorile din coloane indică proporția informației explicate de fiecare componentă principală pentru respectivele observații.

Am putut remarca faptul că prima componentă principală (C1) explică o proporție foarte mare din variabilitatea fiecărei observații din acest subset, având valori ridicate, apropiate de 1. Acest lucru sugerează că dimensiunile surprinse de C1 sunt extrem de relevante pentru aceste observații.

În contrast, contribuțiile celorlalte componente principale (C2-C6) sunt reduse, indicând o influență marginală asupra variabilității acestui subset. Selectarea acestui subset permite o analiză mai detaliată a observațiilor cu o reprezentare foarte bună pe C1, facilitând identificarea pattern-urilor semnificative și înțelegerea mai aprofundată a dimensiunilor explicate de prima componentă.

## 4.8. Contribuția observațiilor

Acest grafic prezintă contribuția observațiilor dintr-un subset selectat la variația axelor componente principale. Fiecare rând reprezintă o observație individuală, iar coloanele (C1-C6) indică proporția variației explicate de fiecare componentă principală pentru acea observație.

Subsetul include observații selectate pe baza contribuției lor semnificative la prima componentă principală (C1). Putem observa că observațiile din acest subset au contribuții majore pe una sau mai multe axe, dar cea mai notabilă contribuție este pe C1, sugerând că această componentă joacă un rol central în explicarea variabilității pentru acest grup de date. Graficul oferă o perspectivă clară asupra observațiilor care influențează cel mai mult dimensiunile componente, permițând o analiză mai detaliată a acestora.

# 5. Analiza Discriminantă

Analiza discriminantă liniară (LDA) este o metodă statistică utilizată pentru a identifica variabilele care contribuie cel mai mult la diferențierea grupurilor definite din date. Această metodă este deosebit de utilă pentru clasificare, oferind un model predictiv capabil să plaseze observațiile noi într-unul dintre grupurile existente.

## 5.1 Pregătirea Datelor

Pentru această analiză, au fost selectate variabilele relevante din setul de date asociate stilului de viață al studenților. Observațiile au fost grupate pe baza variabilei Nivel\_Stres, definind trei grupuri:

* Grup 1: Nivel scăzut de stres
* Grup 2: Nivel moderat de stres
* Grup 3: Nivel ridicat de stres

Datele au fost standardizate pentru a elimina influența unităților de măsură diferite și pentru a asigura comparabilitatea variabilelor. Standardizarea asigură că fiecare variabilă are media 0 și deviația standard 1. Matricea de corelații a fost utilizată pentru a identifica relațiile dintre variabile și eventualele colinearități, similar pregătirii datelor pentru analiza PCA.

## 5.2 Matricele SSW, SSB și SST

Analiza discriminantă implică calcularea a trei matrici esențiale pentru evaluarea variabilității datelor:

1. **SSW (Sum of Squares Within Groups):** Reflectă variabilitatea din cadrul fiecărui grup. Această matrice evaluează cât de apropiate sunt observațiile din același grup.
2. **SSB (Sum of Squares Between Groups):** Măsoară diferențele dintre grupuri, indicând cât de bine sunt separate acestea.
3. **SST (Total Sum of Squares):** Reprezintă variabilitatea totală a datelor, fiind suma dintre SSW și SSB.

Calculul acestor matrici a fost realizat utilizând variabilele selectate, iar rezultatele au fost salvate în fișierele asociate (*ssw\_matrix.csv*, *ssb\_matrix.csv*, *sst\_matrix.csv*). Aceste matrici sunt esențiale pentru determinarea eficienței separării grupurilor și pentru evaluarea contribuției fiecărei variabile la diferențierea acestora.

### Interpretarea Matricilor SSW, SSB și SST

Analiza discriminantă utilizează trei matrici esențiale pentru a evalua variabilitatea datelor și separarea grupurilor.

Matricea SSW (139.22) măsoară variabilitatea în interiorul fiecărui grup. Valoarea ridicată indică o suprapunere semnificativă între grupuri, ceea ce face clasificarea mai dificilă.

Matricea SSB (69.10) reflectă diferențele dintre grupuri. Valoarea relativ scăzută sugerează că grupurile nu sunt foarte bine separate.

Matricea SST (193.32) reprezintă variabilitatea totală a datelor, fiind suma dintre SSW și SSB.

Raportul SSB/SST este 0.36, ceea ce arată că doar 36% din variabilitatea totală provine din diferențele dintre grupuri. În schimb, raportul SSW/SST este 0.72, indicând o mare variație internă și suprapunere între grupuri.

## 5.3 Puterea de Discriminare a Variabilelor

Folosind valorile din SSW și SSB, a fost calculată puterea de discriminare pentru fiecare variabilă. Această măsură indică capacitatea fiecărei variabile de a contribui la separarea grupurilor definite.

Tabelul următor sumarizează rezultatele:

| **Variabilă** | **Puterea de Discriminare** |
| --- | --- |
| Ore\_Studiu | 0.87 |
| Ore\_Activ | 0.78 |
| GPA | 0.92 |
| Ore\_Activități\_Extracuriculare | 0.55 |
| Ore\_Social | 0.63 |
| Ore\_Somn | 0.48 |

Rezultatele detaliate sunt incluse în fișierul *Discrimination\_power.csv*. Variabilele cu o putere de discriminare ridicată, cum ar fi GPA și Ore\_Studiu, sunt cele mai relevante pentru diferențierea grupurilor.

1. **GPA (0.92) și Ore\_Studiu (0.87) – Cei mai puternici predictori**

Aceste două variabile au cele mai mari valori ale puterii de discriminare, ceea ce înseamnă că **performanța academică și timpul alocat studiului sunt esențiale pentru diferențierea grupurilor de stres**.

Studenții cu **stres scăzut** au în general **GPA ridicat și petrec mai mult timp studiind**.

Cei cu **stres ridicat** pot avea fie **GPA mai mic**, fie **GPA mare, dar cu sacrificii mari de somn și activități sociale**.

1. **Ore\_Activ (0.78) – Contribuție moderată spre ridicată**

Activitatea fizică pare să joace un rol în separarea grupurilor.

Studenții cu **stres moderat** pot avea mai multă activitate fizică decât cei din celelalte grupuri, deoarece pot folosi sportul ca o modalitate de reducere a stresului.

Cei cu **stres ridicat** fie nu mai au timp pentru activitate fizică, fie o folosesc excesiv ca mecanism de coping.

1. **Ore\_Social (0.63) – Contribuție medie**

Socializarea influențează stresul, dar nu este un factor decisiv.

Studenții **cu stres scăzut** au un **echilibru între studiu și socializare**.

Cei cu **stres ridicat** tind să aibă mai puține ore sociale, dar diferențele nu sunt foarte accentuate între grupuri.

1. **Ore\_Activități Extracurriculare (0.55) – Contribuție redusă**

Participarea la activități extracurriculare nu este un predictor clar al stresului.

Studenții implicați în activități extracurriculare nu sunt neapărat mai puțin stresați – unii pot gestiona bine sarcinile multiple, alții pot simți o presiune mai mare din cauza responsabilităților suplimentare.

1. **Ore\_Somn (0.48) – Cel mai puțin relevant factor**

Somnul nu separă clar grupurile.

Deși există o tendință ca **stresul ridicat să fie asociat cu mai puține ore de somn**, această relație nu este foarte puternică.

Probabil există studenți care **dorm puțin, dar nu resimt un nivel ridicat de stres**, la fel cum există și studenți care **dorm mult, dar tot sunt stresați** din alte motive.

## 5.4 Matricea de Confuzie și Performanța Modelului

Pentru evaluarea performanței modelului, am construit o matrice de confuzie care compară clasificările modelului LDA cu etichetele reale ale grupurilor. Acurateța generală a modelului a fost calculată ca proporția de observații clasificate corect.

| **Grup Real** | **Grup Prezis: Scăzut** | **Grup Prezis: Moderat** | **Grup Prezis: Ridicat** |
| --- | --- | --- | --- |
| Scăzut | 80 | 10 | 5 |
| Moderat | 12 | 75 | 13 |
| Ridicat | 7 | 8 | 85 |

Indicatorii principali sunt:

* **Acuratețea globală:** 80.5%
* **Precizie:** 78% (media ponderată pe grupuri)
* **Sensibilitate:** 81% (media ponderată pe grupuri)

Performanțele sunt prezentate în fișierul *Accuracy.csv*.

### Interpretarea matricei de confuzie

Pentru clasificarea stresului scăzut, modelul a clasificat corect 80 din cele 95 de observații, ceea ce reprezintă o acuratețe de 84%. Erorile au constat în clasificarea greșită a 10 observații ca „Moderat” și a 5 observații ca „Ridicat”.

În cazul stresului moderat, modelul a clasificat corect 75 din cele 100 de observații, corespunzând unei acuratețe de 75%. Au fost 12 observații atribuite greșit grupului „Scăzut” și 13 observații atribuite greșit grupului „Ridicat”. Aceasta este cea mai problematică clasificare, ceea ce sugerează că studenții cu stres moderat prezintă caracteristici comune cu celelalte două grupuri.

Pentru stresul ridicat, modelul a reușit să clasifice corect 85 din cele 100 de observații, având o acuratețe de 85%. Erorile au inclus 8 observații clasificate greșit ca „Moderat” și 7 observații clasificate greșit ca „Scăzut”. Acest rezultat indică faptul că modelul a distins destul de bine nivelul de stres ridicat, deși au existat câteva erori.

## 5.5 Separarea Grupurilor și Centrii Lor

Vizualizarea separării grupurilor a fost realizată folosind graficul *scatter\_plot\_groups.png*. Fiecare punct reprezintă o observație, iar culorile indică grupul de apartenență. Centrii grupurilor, calculați și salvați în *group\_centers.csv*, indică valorile medii ale variabilelor pentru fiecare grup.

Acest grafic reprezintă distribuția punctelor în funcție de scorurile discriminante rezultate din Analiza Discriminantă Liniară (LDA). Axa X (Z1 - 88.96%) explică cea mai mare parte a variabilității dintre grupuri, iar axa Y (Z2 - 11.04%) oferă informații suplimentare despre separarea acestora.

Separarea grupurilor este vizibilă, dar nu complet clară. Grupul 1, reprezentat prin culoarea deschisă, se află în partea stângă, grupul 2 este distribuit central, iar grupul 3, cu nuanță mai închisă, este situat în partea dreaptă. Suprapunerea dintre grupuri sugerează o diferențiere incompletă.

Distribuția punctelor arată o separare clară pe axa Z1, confirmând că aceasta este cea mai relevantă pentru diferențiere. Pe axa Z2, punctele sunt mai dispersate, ceea ce indică o contribuție mai redusă la separare. Grupurile 2 și 3 sunt mai apropiate, ceea ce poate cauza dificultăți în clasificare, în timp ce grupul 1 este mai bine definit. Aceste rezultate sugerează că, deși LDA separă bine grupurile, suprapunerea categoriilor intermediare poate afecta precizia modelului.

## 5.6 Interpretarea Rezultatelor

* **Ore\_Studiu și GPA:** Grupul cu stres scăzut are GPA mai ridicat și alocă mai multe ore studiului.
* **Ore\_Social și Ore\_Activ:** Grupurile cu stres moderat sau ridicat tind să aibă mai puține ore alocate socializării și mai multe ore activităților fizice.
* **Ore\_Somn:** Nivelurile de stres mai ridicate sunt asociate cu reducerea timpului de somn.

Graficele asociate ilustrează clar diferențele între grupuri, confirmând relevanța variabilelor selectate.

## 5.7 Concluzii ale Analizei Discriminante

Analiza discriminantă a evidențiat factori esențiali care diferențiază grupurile de studenți pe baza nivelului de stres. GPA și Ore\_Studiu sunt cele mai puternice variabile predictive, iar clasificarea realizată de modelul LDA este foarte precisă, cu o acurateță globală de peste 80%. Această analiză oferă perspective utile pentru îmbunătățirea stilului de viață al studenților prin identificarea factorilor critici care contribuie la succesul academic și la gestionarea stresului.