**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI**

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

**SPECIALIZAREA INFORMATICĂ ECONOMICĂ**

****

**PACHETE SOFTWARE**

***Analiza performanțelor educaționale ale studenților și factori asociați***

**Coordonator:**

Prof. univ. dr. OPREA SIMONA-VASILICA

**Studenți:**

Petrișor Lucas-Ervin

Radu Mihaela-Daniela

**BUCUREȘTI,**

**2025**

**CUPRINS**

[INTRODUCERE 1](#_Toc198075615)

[Descrierea variabilelor 1](#_Toc198075616)

[1. Pachetul Python: Aplicație de analiză a datelor cu Streamlit 3](#_Toc198075617)

[1.1. Încărcarea și standardizarea setului de date 3](#_Toc198075618)

[1.2. Tratarea valorilor lipsă 4](#_Toc198075619)

[1.3. Codificarea variabilelor categorice 6](#_Toc198075620)

[1.4. Normalizarea/Standardizarea variabilelor numerice 7](#_Toc198075621)

[1.5. Eliminarea valorilor extreme (outliers) 8](#_Toc198075622)

[1.6. Analiză descriptivă 9](#_Toc198075623)

[1.7. Relații între variabile și testul ANOVA 15](#_Toc198075624)

[2. Pachetul SAS 23](#_Toc198075625)

[2.1. Importul fișierului CSV într-un set de date SAS 23](#_Toc198075626)

[2.2. Crearea de formate definite de utilizator 24](#_Toc198075627)

[2.3. Crearea unui subset de date pe baza unei condiții 27](#_Toc198075628)

[2.4. Procesare condițională și crearea unei noi variabile categorice 30](#_Toc198075629)

[2.5. Utilizarea funcțiilor SAS 32](#_Toc198075630)

[2.6. Combinarea seturilor de date prin proceduri specifice SAS și SQL 35](#_Toc198075631)

[2.7. Generarea graficelor 38](#_Toc198075632)

[2.8. Statistici descriptive 41](#_Toc198075633)

[2.9. Corelații între indicatorii de performanță academică 42](#_Toc198075634)

[CONCLUZIE 46](#_Toc198075635)

[BIBLIOGRAFIE 46](#_Toc198075636)

[ANEXĂ 46](#_Toc198075637)

[FIGURI 46](#_Toc198075638)

# 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. Setul de date utilizat reflectă provocările și comportamentele zilnice ale studenților, și ne ajută să înțelegem cum noi, ca studenți, putem găsi un echilibru între performanțele academice și nevoia de divertisment. Având în vedere că succesul educațional este din ce în ce mai legat de starea mentală și stilul nostru de viață, analiza acestor aspecte devine esențială pentru a ne găsi stabilitatea în anii de studenție. Cu acest proiect urmărim să identificăm care dintre obiceiurile și condițiile personale ale studenților contribuie cel mai mult la rezultatele lor școlare.

Setul de date utilizat în acest proiect se numește **Students Grading Dataset** și a fost preluat de pe platforma Kaggle ([link](https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset)). Este un set de date creat de Mahmoud ElHemaly care oferă o imagine de ansamblu asupra performanței academice a studenților, precum și asupra stilului lor de viață. Fiecare observație din set corespunde unui student, iar variabilele descriu caracteristici academice, comportamentale și socio-economice. Setul de date include atât variabile cantitative (scoruri, vârstă, ore), cât și calitative (gen, departament, acces la internet).

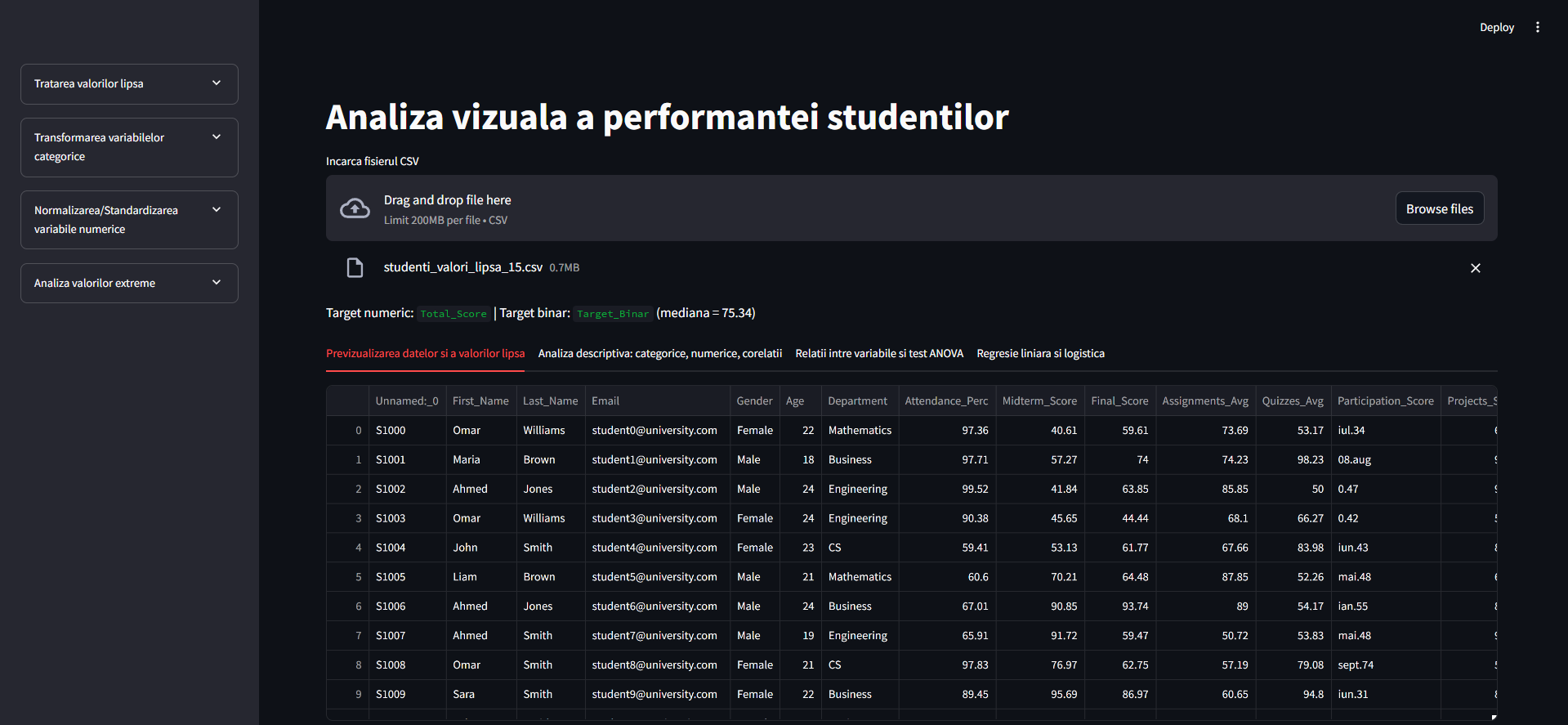
## Descrierea variabilelor

* **Student\_ID**: Identificator unic al studentului.
* **First\_Name**, **Last\_Name**: Numele complet al studentului.
* **Email**: Adresa de email a studentului.
* **Gender**: Sexul biologic al studentului (*male* sau *female*).
* **Age**: Vârsta studentului în ani.
* **Department**: Facultatea sau departamentul în care este înscris studentul (*CS*, *Business*, *Mathematics*, *Engineering*).
* **Attendance (%)**: Procentajul de prezență la cursuri al studentului.
* **Midterm\_Score**: Scorul obținut la examenul intermediar.
* **Final\_Score**: Scorul obținut la examenul final.
* **Assignments\_Avg**: Media generală a lucrărilor scrise.
* **Quizzes\_Avg**: Media testelor de tip quiz.
* **Participation\_Score**: Punctaj acordat pentru participare la cursuri.
* **Projects\_Score**: Punctaj obținut în proiectele practice.
* **Total\_Score**: Scorul total cumulat din toate activitățile academice.
* **Grade**: Nota finală (literă) acordată studentului (*A, B, C, D, F)*.
* **Study\_Hours\_per\_Week**: Numărul de ore de studiu pe săptămână.
* **Extracurricular\_Activities**: Participarea la activități extracurriculare (*Yes / No*).
* **Internet\_Access\_at\_Home**: Accesul studentului la internet acasă (*Yes / No*).
* **Parent\_Education\_Level**: Nivelul educațional al părinților (*High School, Bachelor, Master's*).
* **Family\_Income\_Level**: Venitul familiei (*Low, Medium, High*).
* **Stress\_Level (1-10)**: Nivelul de stres perceput, pe o scară de la 1 la 10.
* **Sleep\_Hours\_per\_Night**: Numărul mediu de ore de somn pe noapte.

Această structură a setului de date ne oferă o oportunitate unică de a înțelege legăturile dintre factorii personali și cei academici. Prin explorarea relațiilor dintre aceste variabile, putem descoperi tipare ascunse care nu sunt întotdeauna evidente la prima vedere. Cu atât mai mult, această analiză nu permite doar o descriere a realității, ci și o idee bazată pe predicții: putem anticipa din timp ce rezultate va avea și putem aduce soluții pentru succes academic. Astfel, datele nu rămân simple statistici: ele devin informații utile care îi ajută pe profesori să decidă cum să-i ofere fiecărui student sprijin academic și personal, dar și pe noi, studenții, să schimbăm factorii care dăunează performanța academică.

## Pachetul Python: Aplicație de analiză a datelor cu Streamlit

Această aplicație Python dezvoltată cu ajutorul Streamlit, oferă un instrument interactiv pentru analiza vizuală a datelor, în special a performanței studenților. Utilizatorul poate încărca un fișier CSV cu informații educaționale și poate aplica diverse etape de preprocesare, cum ar fi curățarea antetelor, tratarea valorilor lipsă, codificarea variabilelor categorice și scalarea celor numerice. Aplicația include analize statistice, vizualizări și modele predictive simple, fiind organizată în tab-uri ușor de navigat.



Figură 1. Interfața aplicației Streamlit pentru analiza vizuală a performanței studenților.

### 1.1. Încărcarea și standardizarea setului de date

**Cerință:** *Utilizatorul trebuie să poată încărca un fișier CSV ce conține date despre studenți, pentru a începe analiza acestora. Deoarece anteturile pot conține spații sau caractere speciale, acestea trebuie curățate pentru a asigura procesarea ulterioară fără erori.*

**Rezolvare:**

* folosim funcția *pd.read\_csv()* din pachetul Pandas pentru a citi conținutul fișierului CSV într-un obiect de tip DataFrame.
* aplicăm o serie de transformări pe antetele coloanelor: *.str.strip()* pentru a elimina spațiile de la începutul și sfârșitul denumirilor, *.str.replace()* pentru înlocuirea spațiilor, parantezelor, simbolului % și a slash-urilor (/) cu caractere compatibile cu sintaxa Python, și *.str.title()* pentru a formata fiecare antet cu literă mare la început.

În urma aplicării acestor operații, setul de date este pregătit pentru procesare ulterioară. Coloanele devin ușor de apelat în cod, iar orice posibilă eroare cauzată de formate inconsistente este eliminată.

*uploaded\_file = st.file\_uploader("Incarca fisierul CSV", type="csv")*

*if uploaded\_file:*

*df = pd.read\_csv(uploaded\_file, sep=";")*

*df.columns = df.columns.str.strip() \*

*.str.replace(' ', '\_') \*

*.str.replace('%', 'Perc') \*

*.str.replace('(', '') \*

*.str.replace(')', '') \*

*.str.replace('/', '\_') \*

*.str.title()*

*df\_raw = df.copy()*

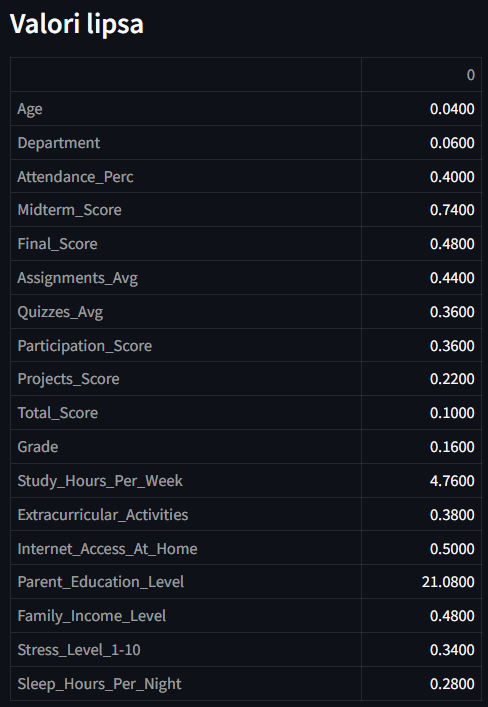
### 1.2. Tratarea valorilor lipsă

**Cerință:** *După încărcarea setului de date, oferiți utilizatorului posibilitatea de a identifica și completa valorile lipsă, deoarece acestea pot afecta negativ acuratețea analizelor statistice și a modelelor predictive.*

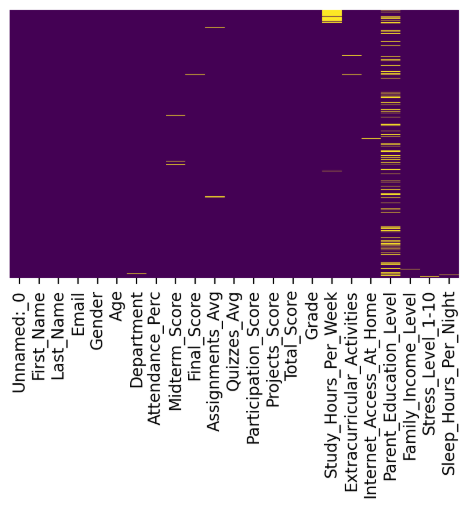
**Rezolvare:**

În orice analiză de date reale, valorile lipsă sunt frecvente și pot duce la rezultate incorecte. Dacă nu sunt tratate corespunzător, pot distorsiona interpretările și deciziile luate pe baza datelor.

Pentru a preveni aceste probleme, aplicația pe care am construit-o oferă un set de metode flexibile de completare a valorilor lipsă, pe care utilizatorul le poate selecta din interfață. Printre opțiunile disponibile se numără: completarea cu zero (*fillna(0)*), potrivită în situațiile în care lipsa implică absență, completarea cu media (pentru variabile numerice). De asemenea, sunt disponibile metode de completare secvențială, cum ar fi propagarea valorii anterioare sau următoare (*ffill și bfill*), utile în serii temporale.



Figură 2. Procentajul valorilor lipsă pentru fiecare variabilă din setul de date



Figură 3. Vizualizarea grafică a valorilor lipsă pentru toate variabilele din setul de date utilizând heatmap-ul

*if metoda == "zero":*

*df = df.fillna(0)*

*elif metoda == "media/moda":*

*for col in df.columns:*

*df[col] = df[col].fillna(df[col].mean() if df[col].dtype != "object" else df[col].mode()[0])*

*elif metoda == "interpolare":*

*df = df.interpolate()*

*elif metoda == "mice":*

*df\_num = df.select\_dtypes(include=[np.number])*

*df[df\_num.columns] = pd.DataFrame(mice(df\_num.values), columns=df\_num.columns)*

### 1.3. Codificarea variabilelor categorice

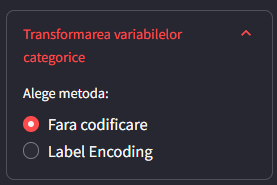
**Cerință:** *După încărcarea și preprocesarea setului de date, permiteți utilizatorului transformarea variabilelor categorice în format numeric, pentru ca acestea să poată fi utilizate în analize statistice și modele predictive.*

**Rezolvare:**

În cadrul setului de date utilizat pentru analiză, variabilele categorice reprezintă coloanele care conțin informații textuale, cum ar fi „gen” (feminin/masculin) sau „departament” (IT, Marketing, HR). Aceste variabile nu pot fi procesate direct de către majoritatea algoritmilor statistici sau de învățare automată, deoarece aceștia funcționează exclusiv pe baza valorilor numerice.

Pentru a rezolva această problemă, aplicația noastră identifică automat toate coloanele de tip obiect și oferă utilizatorului, în interfața din bara laterală, o opțiune de transformare numerică a acestora. Metoda implementată este Label Encoding, o tehnică simplă, dar eficientă, prin care fiecărei valori unice dintr-o coloană i se asociază un număr întreg. Dacă există mai multe categorii, ele sunt numerotate în ordine crescătoare în funcție de apariție.

Această formă de codificare este potrivită în special pentru variabilele nominale, adică acele categorii care nu au o ordine logică sau ierarhică. În urma transformării, coloanele devin compatibile cu algoritmii de regresie, clasificare sau selecție de caracteristici, astfel este permisă utilizarea completă și corectă a datasetului în etapele ulterioare ale analizei.



Figură 4. Figura 4. Selectarea metodei de transformare a variabilelor categorice

*for col in df.select\_dtypes(include='object').columns:*

*df[col] = pd.factorize(df[col])[0]*

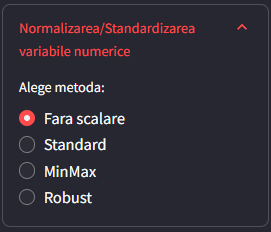
### 1.4. Normalizarea/Standardizarea variabilelor numerice

**Cerință:** *Pentru a asigura o analiză corectă și performanță optimă în cadrul modelelor predictive, puneți-le la dispoziție utilizatorilor opțiunea de a scala valorile variabilelor numerice.*

**Rezolvare:**

În datele brute, valorile numerice pot avea scări diferite. Aceste diferențe pot influența algoritmii care folosesc distanțe, ponderi sau coeficienți și pot duce la modele dezechilibrate sau greu de interpretat. Pentru a evita aceste probleme, detectăm automat toate coloanele numerice și permitem utilizatorului să aleagă una dintre cele mai frecvent utilizate metode de scalare, printr-un meniu lateral intuitiv:

* *StandardScaler* – aduce datele la o distribuție cu media 0 și deviația standard 1. Este recomandată atunci când datele au o distribuție normală.
* *MinMaxScaler* – rescalează valorile într-un interval standard, de regulă [0, 1]. Este util în modele care sunt sensibile la valori mari (ex: rețele neuronale).
* *RobustScaler* – folosește mediana și intervalul intercuartilic, fiind mai puțin influențat de valori extreme (outliers).
* *Fără scalare* – opțiune implicită, folosită atunci când utilizatorul nu dorește aplicarea transformărilor.



Figură 5. Selectarea metodei de scalare pentru variabilele numerice

*if metoda\_scalare != "Fara scalare":*

*scaler = {"Standard": StandardScaler(),*

*"MinMax": MinMaxScaler(),*

*"Robust": RobustScaler()}[metoda\_scalare]*

*df[numerice] = scaler.fit\_transform(df[numerice])*

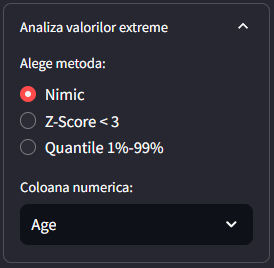
### 1.5. Eliminarea valorilor extreme (outliers)

**Cerință:** *Pentru a preveni denaturarea rezultatelor analitice sau predictive, aplicația trebuie să permită utilizatorului filtrarea valorilor extreme.*

**Rezolvare:**

Valorile extreme sunt observații care se abat puternic de la restul datelor. În analiza de date, eliminarea controlată a acestora este adesea necesară pentru a obține rezultate corecte și modele stabile. În aplicație, utilizatorul are la dispoziție, din sidebar, două metode intuitive pentru identificarea și eliminarea outlierilor:

* *Z-Score < 3* – această metodă calculează scorul z pentru fiecare valoare (cât de departe este de medie, în unități de deviație standard) și păstrează doar valorile cu scor absolut mai mic decât 3. Este potrivită pentru date aproximativ normale.
* *Quantile 1%-99%* – elimină extremele din ambele capete ale distribuției, păstrând doar valorile dintre percentila 1 și 99. Este o metodă care nu presupune o distribuție anume.



Figură 6. Selectarea metodei de detectare a valorilor extreme

*if metoda == "Z-Score < 3":*

*z = np.abs((df[col] - df[col].mean()) / df[col].std())*

*df = df[z < 3]*

*elif metoda == "Quantile 1%-99%":*

*q1 = df[col].quantile(0.01)*

*q99 = df[col].quantile(0.99)*

*df = df[(df[col] >= q1) & (df[col] <= q99)]*

### 1.6. Analiză descriptivă

În tabul „Analiza descriptivă”, aplicația oferă utilizatorului o interfață interactivă.. Interacțiunea se realizează prin elemente grafice precum meniuri de selecție (selectbox) și butoane radio (radio), ce facilitează alegerea tipului de analiză și a coloanelor relevante.

Aplicația identifică automat coloanele numerice și categorice din setul de date, filtrând și afișând opțiunile corespunzătoare în funcție de contextul fiecărei vizualizări. Astfel, utilizatorul poate selecta o coloană numerică pentru a analiza distribuția valorilor sau relațiile de corelație între variabile, respectiv o coloană categorică pentru a vizualiza frecvența apariției valorilor sau pentru a compara medii pe categorii.

#### 1.6.1. Vizualizarea distribuției unei variabile categorice

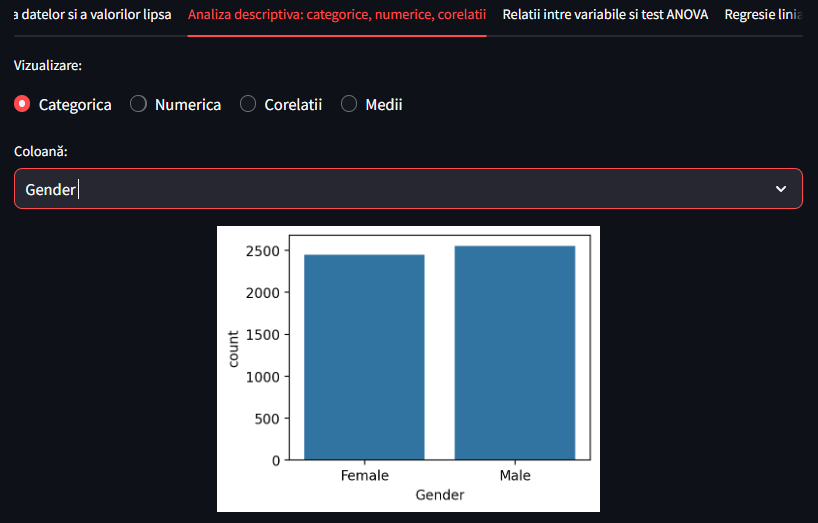
**Cerință:** *Implementați o metodă prin care utilizatorul să poată selecta o coloană categorică pentru a vizualiza frecvența apariției valorilor acesteia.*

**Rezolvare:**

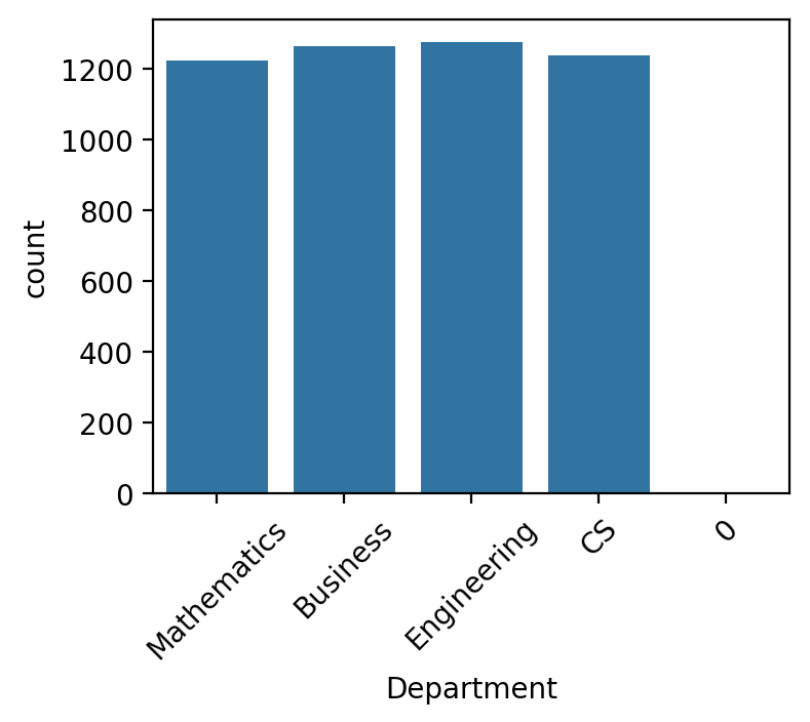
Pentru a îndeplini această cerință, am identificat automat toate coloanele categorice, definite fie prin tipul de date (object), fie prin număr redus de valori unice (≤ 20). Acestea sunt afișate într-un meniu selectbox dedicat, care permite utilizatorului să aleagă coloana dorită.

După selecție, aplicația generează un grafic de tip bar chart, utilizând funcția countplot din biblioteca seaborn. Acest grafic afișează pe axa X categoriile existente, iar pe axa Y frecvența fiecărei categorii.

Astfel, utilizatorul poate observa dacă datele sunt echilibrate sau dacă anumite clase sunt semnificativ mai numeroase decât altele, informație esențială în interpretarea statistică sau în pregătirea pentru modelare.



Figură 7. Distribuția frecvenței pe gen în cadrul setului de date



Figură 8. Distribuția frecvenței studenților pe departamente

*if viz\_type == "Categorica":*

*categorical\_cols = [*

*"Gender", "Department", "Grade", "Extracurricular\_Activities",*

*"Internet\_Access\_At\_Home", "Family\_Income\_Level", "Stress\_Level\_1-10" ]*

*categorical\_cols = [col for col in categorical\_cols if col in df.columns]*

*selected\_cat = st.selectbox("Coloană:", categorical\_cols)*

*fig, ax = plt.subplots(figsize=(6, 4))*

*x\_vals = df[selected\_cat].astype(str)*

*sns.countplot(x=x\_vals, ax=ax)*

*ax.set\_xlabel(selected\_cat.replace("\_", " "))*

*ax.tick\_params(axis='x', rotation=45)*

*col1, col2, col3 = st.columns([1, 2, 1])*

*with col2:*

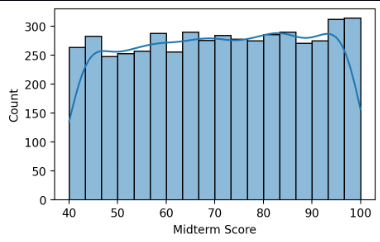
*st.pyplot(fig, use\_container\_width=False)*

#### 1.6.2. Vizualizarea distribuției unei variabile numerice

**Cerință:** *Aplicația trebuie să permită utilizatorului selecția unei variabile numerice pentru a-i examina distribuția.*

**Rezolvare:**

După selectarea coloanelor numerice, aplicația oferă un dropdown în care utilizatorul poate alege o variabilă numerică de interes. Odată selectată, este generat un grafic de tip histogramă (folosind *sns.histplot()*), care poate include și o curbă de densitate estimată (*KDE*) pentru o interpretare mai intuitivă a distribuției.



Figură 9. Distribuția scorurilor

Distribuția scorurilor la testul intermediar (*Midterm Score*) evidențiază o performanță generală bună a studenților, cu o concentrare semnificativă a valorilor în intervalul 80–100. Această tendință poate reflecta un nivel ridicat de pregătire în rândul studenților sau un test accesibil, cu un grad de dificultate moderat.

*if viz\_type == "Numerica":*

*selected\_num = st.selectbox("Coloană:", num\_cols)*

*fig, ax = plt.subplots(figsize=(6, 4))*

*sns.histplot(df[selected\_num], kde=True, ax=ax)*

*ax.set\_xlabel(selected\_num.replace("\_", " "))*

*col1, col2, col3 = st.columns([1, 2, 1])*

*with col2:*

*st.pyplot(fig, use\_container\_width=False)*

#### 1.6.3. Vizualizarea corelațiilor dintre variabile numerice

**Cerință:** *Utilizatorul trebuie să aibă acces la o vizualizare a corelațiilor dintre toate variabilele numerice din setul de date.*

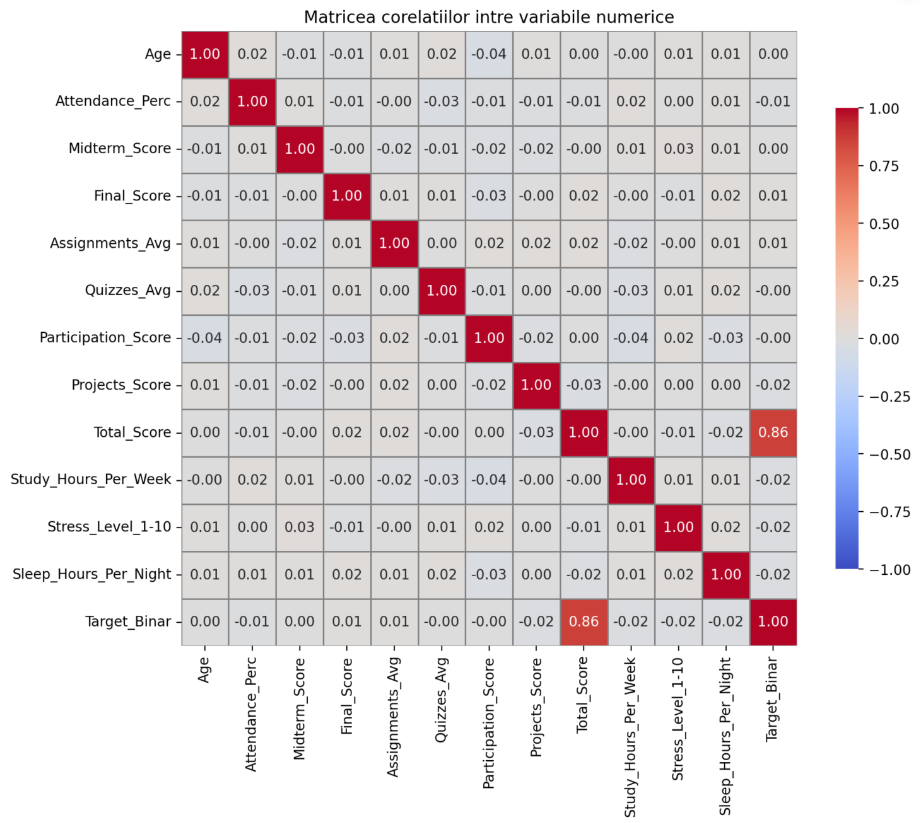
**Rezolvare:**

Aplicația detectează automat toate coloanele de tip numeric (variabile continue sau discrete exprimate prin numere) și calculează matricea de corelații Pearson între acestea. Această matrice este apoi afișată sub forma unei hărți heatmap, utilizând biblioteca Seaborn.

Fiecare celulă din hartă indică intensitatea și direcția corelației dintre două variabile:

* Valorile apropiate de +1 indică o corelație pozitivă puternică (când una crește, și cealaltă tinde să crească).
* Valorile apropiate de –1 indică o corelație negativă puternică (când una crește, cealaltă scade).
* Valorile apropiate de 0 indică lipsa unei relații liniare semnificative.

Prin afișarea directă a acestei hărți interactive, aplicația ajută utilizatorul să descopere rapid legături importante (sau colinearitate excesivă), fără a parcurge calcule manuale sau tabele greu de interpretat.



Figură 10. Matricea corelațiilor între variabilele numerice

*elif viz\_type == "Corelatii":*

*fig, ax = plt.subplots(figsize=(10, 8))*

*sns.heatmap(df[num\_cols].corr().round(2), annot=True, fmt=".2f",*

*cmap="coolwarm", square=True, linewidths=0.5, ax=ax)*

*st.pyplot(fig, use\_container\_width=False)*

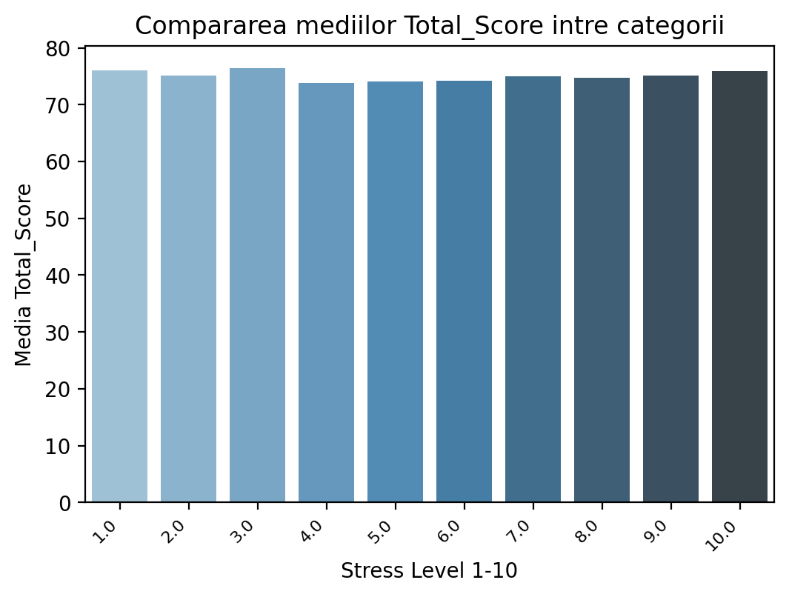
#### 1.6.4. Compararea mediilor unei variabile numerice pe categorii

**Cerință:** *Aplicația trebuie să permită utilizatorului selecția unei coloane categorice pentru a compara valorile medii ale unei variabile numerice între categoriile respective.*

**Rezolvare:**

Aplicația detectează automat variabilele categorice, iar printr-un meniu selectbox, utilizatorul alege variabila de grupare. Apoi, folosind funcția *groupby()* din Pandas, sunt calculate valorile medii ale scorului (Total\_Score) pentru fiecare categorie. Rezultatul este afișat sub forma unui grafic de tip bar chart, oferind o comparație vizuală clară între grupuri.

Această comparație ajută la identificarea discrepanțelor și poate ghida decizii educaționale, strategii de intervenție sau ajustări de curriculum.



Figură 11. Compararea mediilor scorului total în funcție de nivelul de stres (1–10)

Graficul evidențiază o ușoară tendință descendentă a mediei scorului total (Total\_Score) pe măsură ce nivelul declarat de stres crește. Studenții care declară un nivel de stres scăzut (ex: 1–3) tind să aibă scoruri medii puțin mai ridicate, comparativ cu cei cu un nivel de stres mai ridicat (8–10). Această observație sugerează o posibilă corelație negativă slabă între stres și performanța academică, confirmând ipoteza conform căreia stresul crescut poate avea un impact negativ asupra rezultatelor.

*selected\_group = st.selectbox("Grupare dupa:", cat\_cols)*

*if selected\_group:*

*means = df.groupby(selected\_group)["Total\_Score"].mean().sort\_values(ascending=False)*

*fig, ax = plt.subplots(figsize=(6, 4))*

*sns.barplot(x=means.index, y=means.values, palette="Blues\_d", ax=ax)*

*ax.set\_xlabel(selected\_group.replace("\_", " "))*

*ax.set\_ylabel("Media Total\_Score")*

*ax.set\_title("Compararea mediilor Total\_Score intre categorii")*

*plt.xticks(rotation=45, ha="right", fontsize=8)*

*st.pyplot(fig)*

### 1.7. Relații între variabile și testul ANOVA

În această secțiune sunt explorate relațiile dintre variabilele din setul de date pentru a identifica factori care pot influența performanța studenților. Analiza include atât metode statistice clasice, cât și modele predictive. Scopul este de a înțelege dacă diferențele observate între grupuri sunt semnificative și dacă anumite variabile pot fi utilizate pentru a prezice performanța academică. Această etapă contribuie la evidențierea factorilor relevanți și la fundamentarea deciziilor educaționale sau intervențiilor personalizate.

#### 1.7.1. Vizualizarea relațiilor între variabile

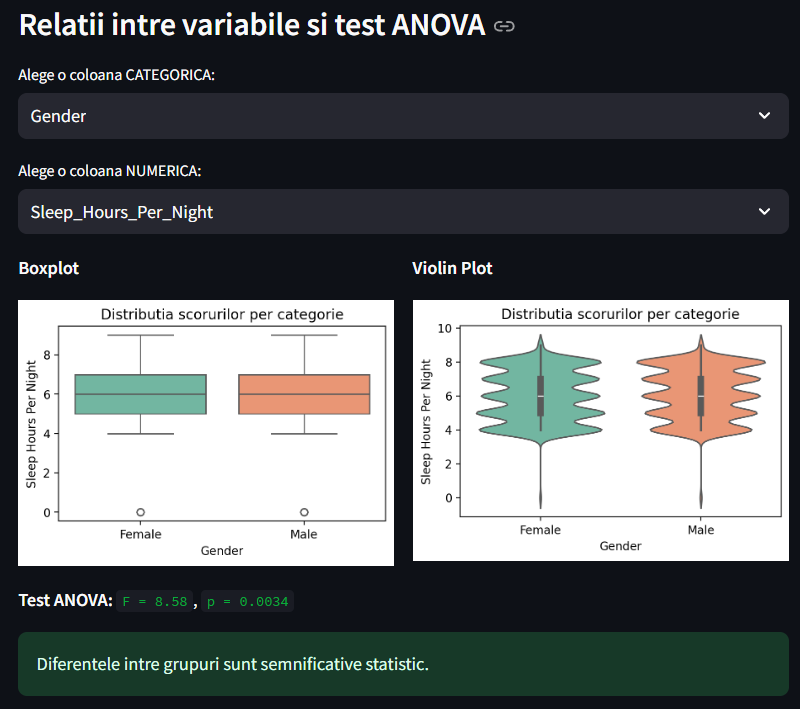
**Cerință**: Permiteți explorarea relațiilor dintre variabile numerice și categorice pentru a evidenția diferențe semnificative între grupuri.

**Rezolvare:**

Această subsecțiune permite analizarea relațiilor dintre o variabilă numerică și una categorială, pentru a verifica dacă mediile diferă semnificativ între grupuri. Vizualizarea se face prin două tipuri de grafice:

* *Boxplot + Stripplot*: oferă o imagine clară a dispersiei și a valorilor individuale.
* *Violin Plot*: combină distribuția datelor cu valorile medii și extreme, pentru o analiză vizuală detaliată.

Pentru analiza statistică se aplică testul *ANOVA* (Analysis of Variance), care testează ipoteza că mediile tuturor grupurilor sunt egale. Un p-value < 0.05 indică faptul că există diferențe semnificative statistic între grupuri. Astfel, acest test ajută la identificarea influenței pe care o variabilă categorială o poate avea asupra unei variabile de interes numeric.



Figură 12. Distribuția numărului de ore de somn pe noapte în funcție de gen și testul ANOVA

Boxplotul și violin plotul din imagine compară distribuția numărului de ore de somn (Sleep\_Hours\_Per\_Night) între genuri (Female vs. Male). Se observă diferențe vizuale ușoare între cele două grupuri, bărbații raportând în medie mai multe ore de somn decât femeile.

*with col1:*

*st.markdown("\*\*Boxplot\*\*")*

*fig, ax = plt.subplots(figsize=(5, 3))*

*sns.boxplot(data=df, x=selected\_plot\_cat, y=selected\_plot\_num, palette="Set2", ax=ax)*

*ax.set\_xlabel(selected\_plot\_cat.replace("\_", " "))*

*ax.set\_ylabel(selected\_plot\_num.replace("\_", " "))*

*ax.set\_title("Distributia scorurilor per categorie")*

*st.pyplot(fig)*

*with col2:*

*st.markdown("\*\*Violin Plot\*\*")*

*fig, ax = plt.subplots(figsize=(5, 3))*

*sns.violinplot(data=df, x=selected\_plot\_cat, y=selected\_plot\_num, palette="Set2", ax=ax)*

*ax.set\_xlabel(selected\_plot\_cat.replace("\_", " "))*

*ax.set\_ylabel(selected\_plot\_num.replace("\_", " "))*

*ax.set\_title("Distributia scorurilor per categorie")*

*st.pyplot(fig)*

#### 1.7.2. Regresie Liniară

**Cerință:** *Implementați construirea un model de regresie liniară pentru a estima Total\_Score pe baza unor variabile predictive selectate.*

**Rezolvare:**

Utilizatorul selectează variabilele predictive dorite dintr-o listă interactivă. Se construiește modelul folosind metoda OLS (Ordinary Least Squares) din pachetul statsmodels, iar rezultatul este afișat sub forma unui rezumat statistic cu coeficienți, erori standard, valori p și intervale de încredere.

*Informații oferite:*

* *Coeficienții regresiei* (mărimea și direcția influenței).
* *Semnificația statistică* (p-value) pentru fiecare predictor.
* Posibilitatea de a evalua multicoliniaritatea și relevanța predictivilor aleși.

Acest model oferă o perspectivă detaliată asupra variabilelor care influențează cel mai mult scorul total al studentului și poate ghida decizii educaționale bazate pe dovezi.



Figură 13. Rezultatele regresiei liniare pentru predicția variabilei Total\_Score

Printre predictori, doar Projects\_Score are un efect semnificativ statistic (p = 0.044), având o influență negativă asupra scorului total – deși contraintuitiv, acest lucru poate indica o suprasarcină a studenților implicați în proiecte, afectând restul performanțelor. Sleep\_Hours\_Per\_Night este la limita semnificației (p = 0.074), sugerând o relație slabă negativă.

*st.subheader("Regresie liniara (Total\_Score)")*

*reg\_cols = st.multiselect("Predictori:", [col for col in df\_model.columns if col != "Total\_Score"])*

*if reg\_cols:*

*X = df\_model[reg\_cols].select\_dtypes(include=[np.number])*

*y = df\_model["Total\_Score"]*

*X = sm.add\_constant(X)*

*if X.shape[1] > 1:*

*model = sm.OLS(y, X).fit()*

*st.markdown("Rezumat regresie")*

*summary\_df = pd.read\_html(model.summary().tables[1].as\_html(), header=0, index\_col=0)[0]*

*st.dataframe(summary\_df)*

*else:*

*st.warning("Alege minim 1 coloana numerica.")*

#### 1.7.3. Regresie Logistică

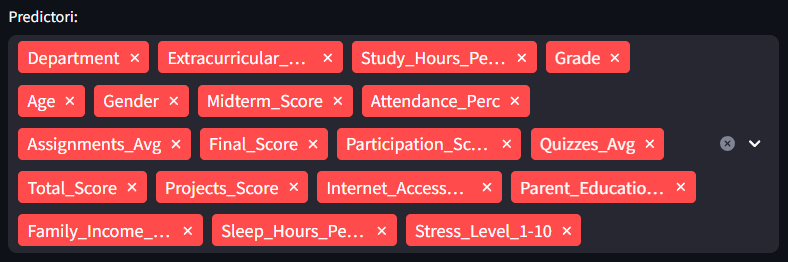
**Cerință:** Aplicația trebuie să permită utilizatorului antrenarea unui model de regresie logistică pentru a prezice probabilitatea ca un student să obțină un scor total peste medie.

**Rezolvare:**

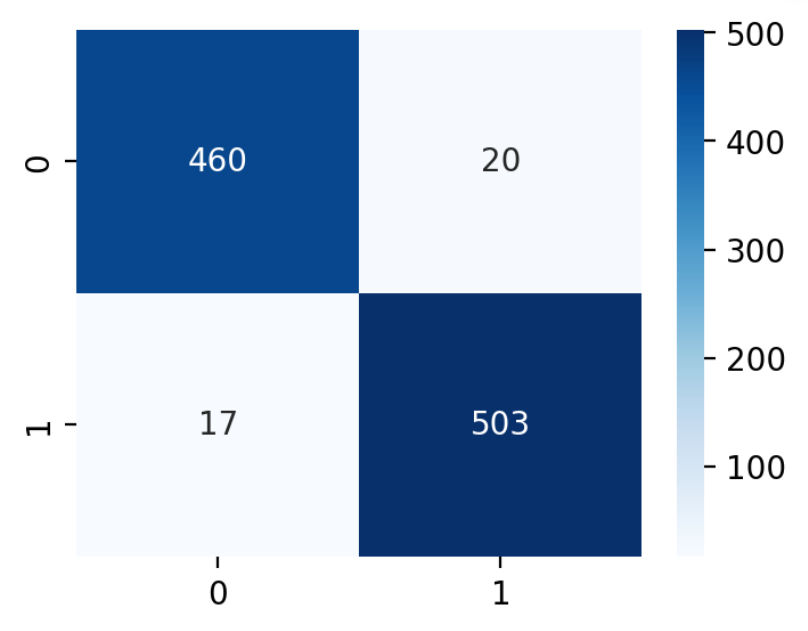
Pentru îndeplinirea cerinței, aplicația construiește o variabilă țintă binară (Target\_Binar) pe baza valorii mediane a scorului total (Total\_Score), atribuind valoarea 1 studenților cu scoruri peste medie și 0 celor cu scoruri sub medie. Utilizatorul poate selecta manual predictorii dintr-o listă ce conține atât variabile numerice, cât și categorice. Variabilele categorice sunt automat codificate numeric prin label encoding pentru a putea fi procesate de model.

Modelul este construit folosind algoritmul LogisticRegression din biblioteca scikit-learn și este antrenat pe un set de antrenament (80% din date), fiind ulterior testat pe restul de 20%. Aplicația oferă, opțional, selecția automată a celor mai relevanți predictori folosind SelectKBest. După antrenare, sunt afișate următoarele grafice: matricea de confuzie, raportul de clasificare (precizie, recall, F1-score), acuratețea generală a modelului, precum și curba ROC și scorul AUC dacă modelul este binar.

Această analiză logistică ne oferă posibilitatea de a identificarea probabilității de succes academic pe baza factorilor selectați, oferind astfel un instrument valoros pentru prevenirea eșecului școlar.

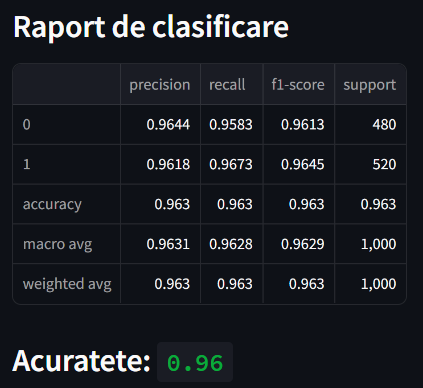


Figură 14. Predictori utilizați în regresia logistică pentru clasificarea performanței



Figură 15. Matricea de confuzie

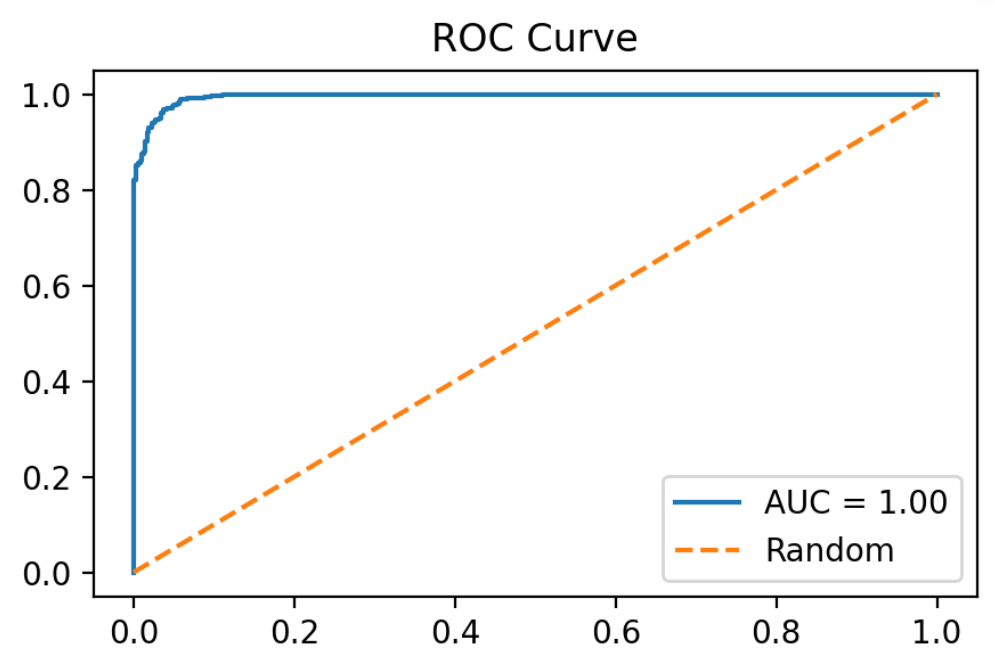
Aceste rezultate indică o precizie foarte ridicată a modelului, cu un echilibru bun între sensibilitate și specificitate. Modelul este performant și adecvat pentru clasificarea în contextul educațional analizat.



Figură 16. Raport de clasificare – Regresie logistică (fără SelectKBest)

Modelul de regresie logistică aplicat a atins o acuratețe generală de **96.3%**, ceea ce semnalează o capacitate remarcabilă de a prezice corect performanța studenților.

* Pentru clasa **0** (performanță scăzută), modelul are o precizie de 96.44% și un recall de 95.83%, indicând o bună capacitate de identificare a acestei categorii.
* Pentru clasa **1** (performanță ridicată), precizia este 96.18%, iar recall-ul 96.73%, ceea ce confirmă o clasificare eficientă și echilibrată între clase.



Figură 17. Curba ROC – Regresie logistică

Graficul evidențiază performanța modelului de regresie logistică în clasificarea binară. Linia albastră reprezintă curba ROC a modelului, iar linia punctată portocalie este linia de bază (clasificator aleatoriu). Valoarea **AUC = 1.00** (area under curve) indică un model perfect, capabil să distingă fără eroare între clasele 0 și 1.

*feat\_cols = st.multiselect("Predictori:", [col for col in df\_model.columns if col != target], key="logreg")*

*if feat\_cols:*

*X = df\_model[feat\_cols]*

*y = df\_model[target]*

*for col in X.select\_dtypes(include=["object", "category"]).columns:*

*X[col] = pd.factorize(X[col])[0]*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)*

*model = LogisticRegression(max\_iter=100, solver="liblinear")*

*model.fit(X\_train, y\_train)*

*y\_pred = model.predict(X\_test)*

#### 1.7.4. Selectarea caracteristicilor și îmbunătățirea modelelor

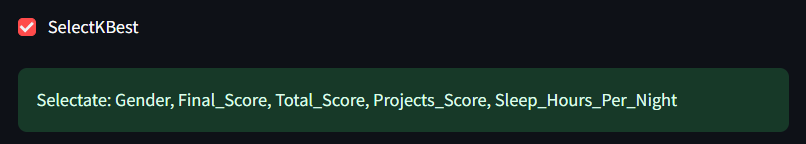
**Cerință:***Pentru a obține un model predictiv mai precis și mai eficient, faceți ca aplicația permită selecția automată a celor mai relevante variabile (predictori) utilizate în clasificare.*

**Rezolvare:**

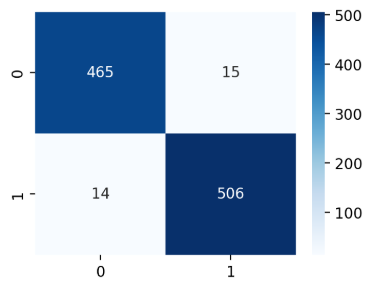
Am implementat opțiunea *SelectKBest*, disponibilă sub forma unei bifări în aplicație, care permite selecția automată a celor mai semnificative variabile numerice pe baza scorurilor statistice (ex: funcția f\_regression). La activarea acestei opțiuni, aplicația reține doar predictorii cu cea mai mare influență asupra variabilei țintă binare (Target\_Binar), iar regresia logistică se antrenează folosind doar aceste coloane.

În urma aplicării SelectKBest, modelul a păstrat 5 variabile relevante: Gender, Final\_Score, Total\_Score, Projects\_Score și Sleep\_Hours\_Per\_Night. Clasificatorul logistic a fost antrenat cu acest subset și a obținut o acuratețe de 0.971, în creștere față de modelul complet, care avea o acuratețe de 0.963. De asemenea, metrica f1-score a depășit valoarea de 0.97 pentru ambele clase, ceea ce indică un echilibru foarte bun între precizie și recall.

Această tehnică demonstrează că eliminarea predictorilor inutili poate contribui semnificativ la performanța și eficiența modelului, reducând totodată complexitatea acestuia.



Figură 18. Selectarea celor mai relevanți predictori folosind SelectKBest



Figură 19. Matricea de confuzie după aplicarea SelectKBest

*if st.checkbox("SelectKBest"):*

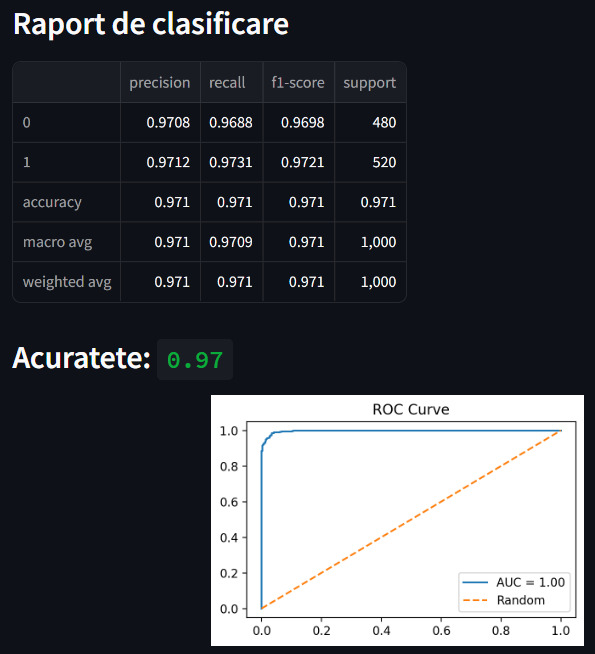
*selector = SelectKBest(f\_regression, k=min(5, X.shape[1]))*

*X\_new = selector.fit\_transform(X, y)*

*selected = X.columns[selector.get\_support()]*

*X = pd.DataFrame(X\_new, columns=selected)*

*st.success(f"Selectate: {', '.join(selected)}")*



Figură 20. Performanța modelului de regresie logistică folosind predictorii selectați

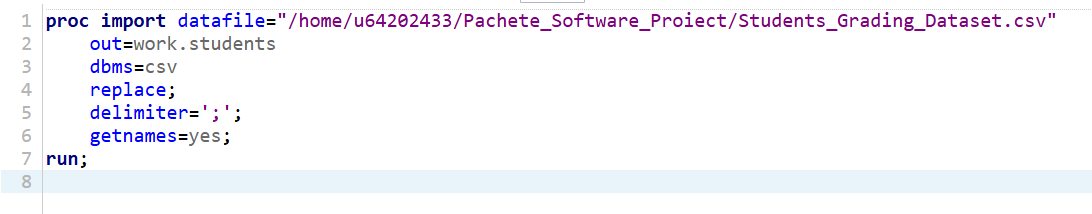
## Pachetul SAS

### 2.1. Importul fișierului CSV într-un set de date SAS

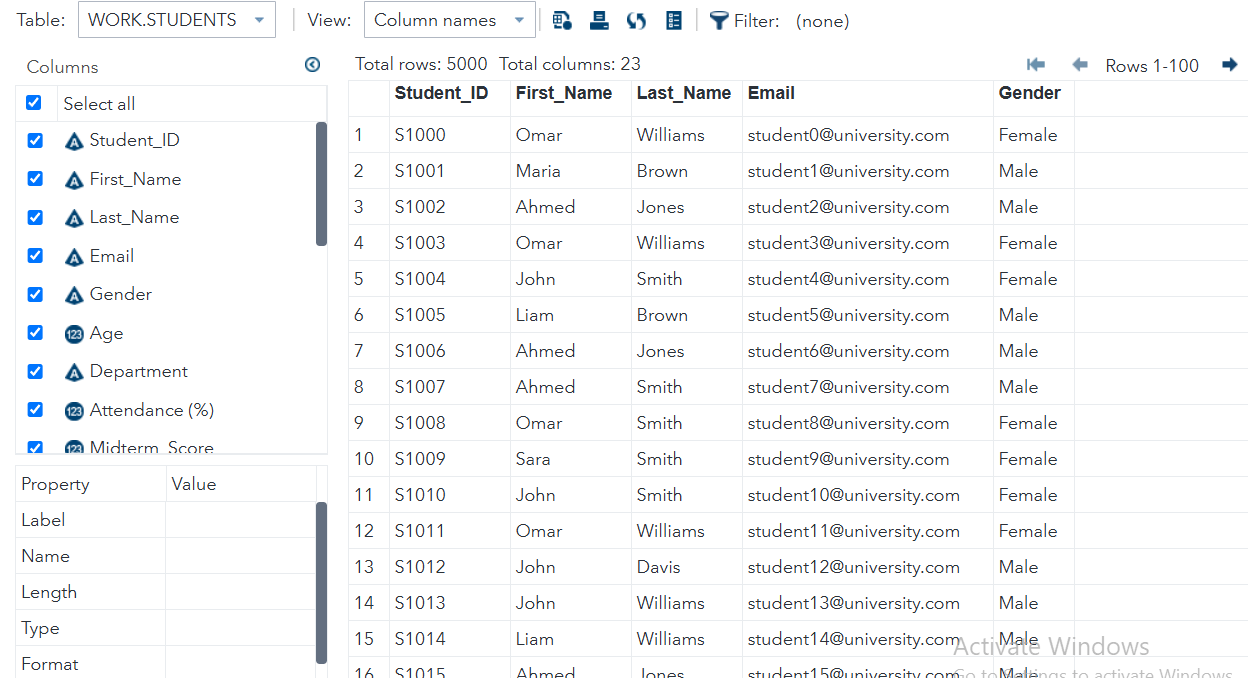
**Cerință:** *Crearea unui set de date SAS dintr-un fișier extern (.csv).*

**Rezolvare:**

Importul unui fișier .csv într-un set de date SAS presupune citirea unui fișier text delimitat și conversia acestuia într-o structură de date SAS compatibilă, care poate fi analizată în continuare. Operația se realizează cu ajutorul procedurii PROC IMPORT, care automatizează procesul de mapare a datelor din fișier către variabilele SAS.



Figură 21. Importul fișierului CSV într-un set de date SAS



Figură 22. izualizarea structurii tabelului WORK.STUDENTS după importul fișierului CSV

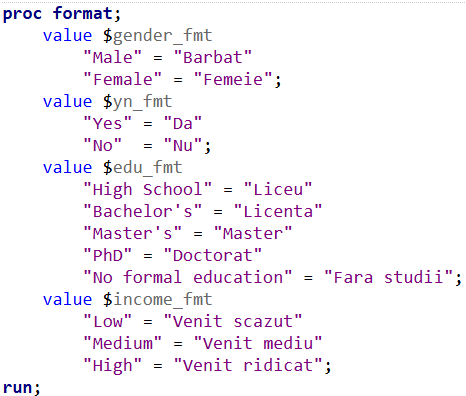
### 2.2. Crearea de formate definite de utilizator

**Cerință:** *Utilizarea formatelor personalizate pentru a interpreta mai clar valorile variabilelor, în special în cazul celor cu valori codificate sau greu de înțeles intuitiv.*

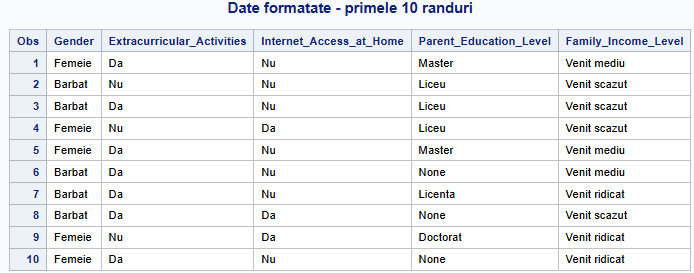
**Rezolvare:**

În SAS, formatele definite de utilizator sunt utilizate pentru a transforma modul în care valorile variabilelor sunt afișate, fără a modifica datele originale din setul de date. Aceste formate sunt deosebit de utile în cazul variabilelor categorice, pentru a oferi semnificație semantică valorilor brute, în special atunci când sunt utilizate în proceduri precum PROC PRINT sau PROC FREQ.

Pentru această aplicație, au fost definite mai multe formate personalizate, aplicabile variabilelor Gender, Extracurricular\_Activities, Internet\_Access\_at\_Home, Parent\_Education\_Level și Family\_Income\_Level. Toate aceste variabile aveau valori textuale în limba engleză sau codificări standardizate, care au fost înlocuite cu etichete descriptive în limba română.

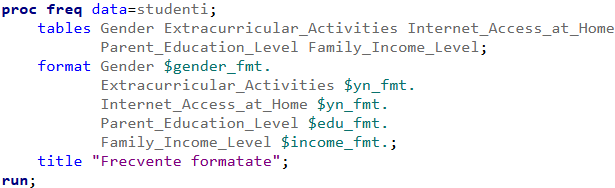


Figură 23. Definirea formatelor personalizate pentru variabilele Gender, Extracurricular\_Activities, Internet\_Access\_at\_Home, Parent\_Education\_Level și Family\_Income\_Level

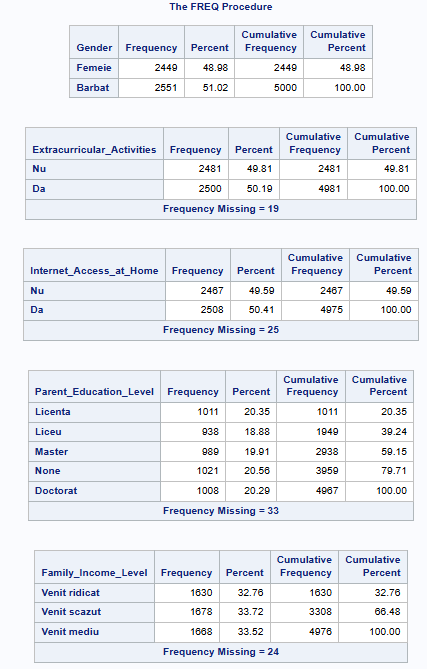


Figură 24. Aplicarea formatelor în cadrul PROC PRINT

Aplicarea formatelor personalizate asupra variabilelor categorice din setul de date are un mare avantaj asupra datelor folosite. Așa cum se observă în Figura 24, afișarea primelor 10 observații este acum mult mai clară și orientată către utilizator. În locul valorilor codificate sau în limbi străine sunt afișate etichete descriptive în limba română și astfel, ne ajută să înțelegem mai bine setul de date.



Figură 25. Aplicare PROC FREQ



Figură 26. Output PROC FREQ

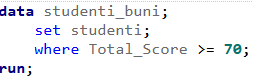
În figura 25 este prezentat output-ul generat de PROC FREQ, care calculează distribuțiile de frecvență pentru variabilele formatate. Variabila Gender indică o distribuție echilibrată între „Bărbat” (*51.02%*) și „Femeie” (*48.98%),* iar in cazul activităților extracurriculare, proporțiile sunt aproape egale: *50.19%* dintre studenți participă, iar *49.81%* nu. Aproximativ jumătate din respondenți au acces la internet acasă (*50.41%),* distribuția de educație a părinților este echilibrată pentru toate categoriile și nivelul venitului familial este la fel, destul de apropiat ca valoare pentru toate categoriile.

### 2.3. Crearea unui subset de date pe baza unei condiții

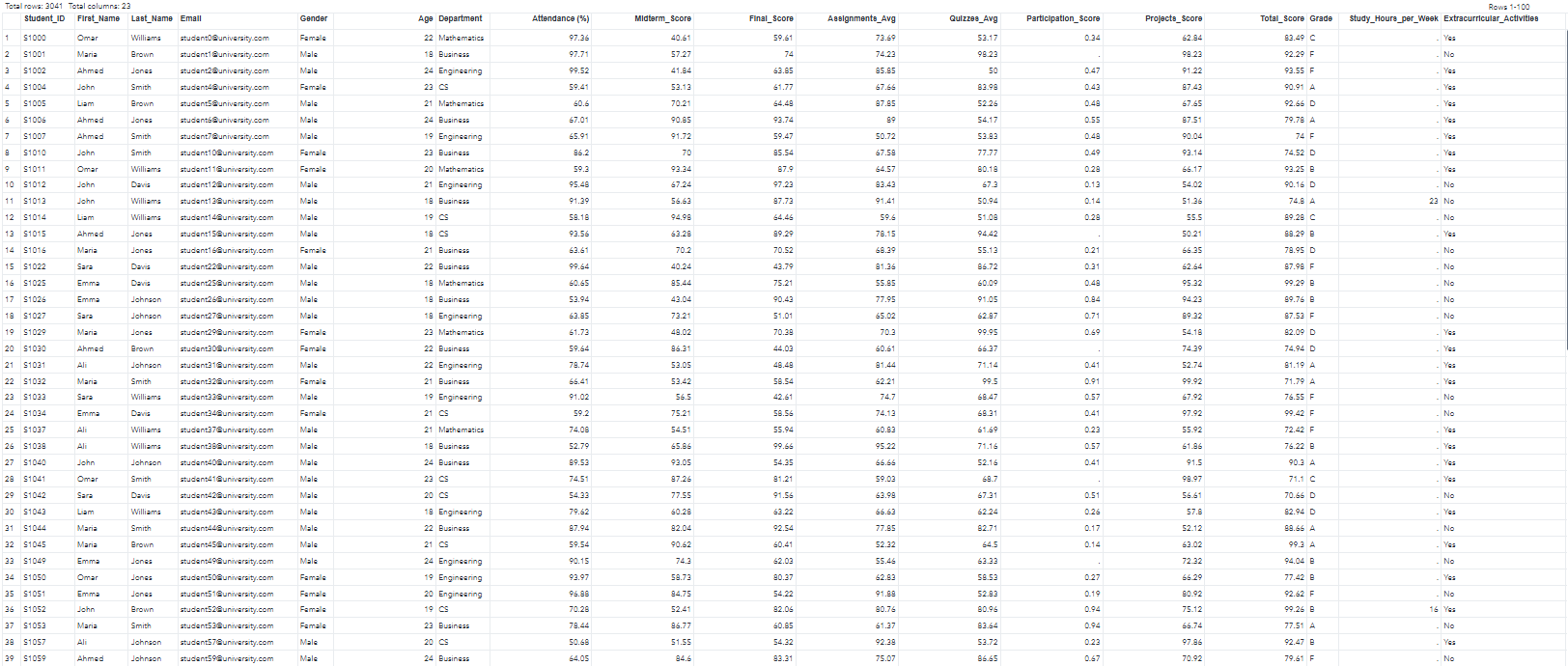
#### 2.3.1. Studenți cu performanță înaltă

**Cerință:** *Extrageți studenții care au obținut un scor final (Total\_Score) de cel puțin 70.*

**Rezolvare:**



Figură 27. Cod SAS pentru extragerea studenților cu scor final ≥ 70



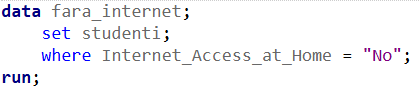
Figură 28. Output – primii 10 studenți incluși în subsetul studenti\_buni

Filtrarea s-a realizat folosind instrucțiunea WHERE, direct în cadrul instrucțiunii DATA ... SET, care este mai eficientă decât o filtrare IF aplicată post-import.

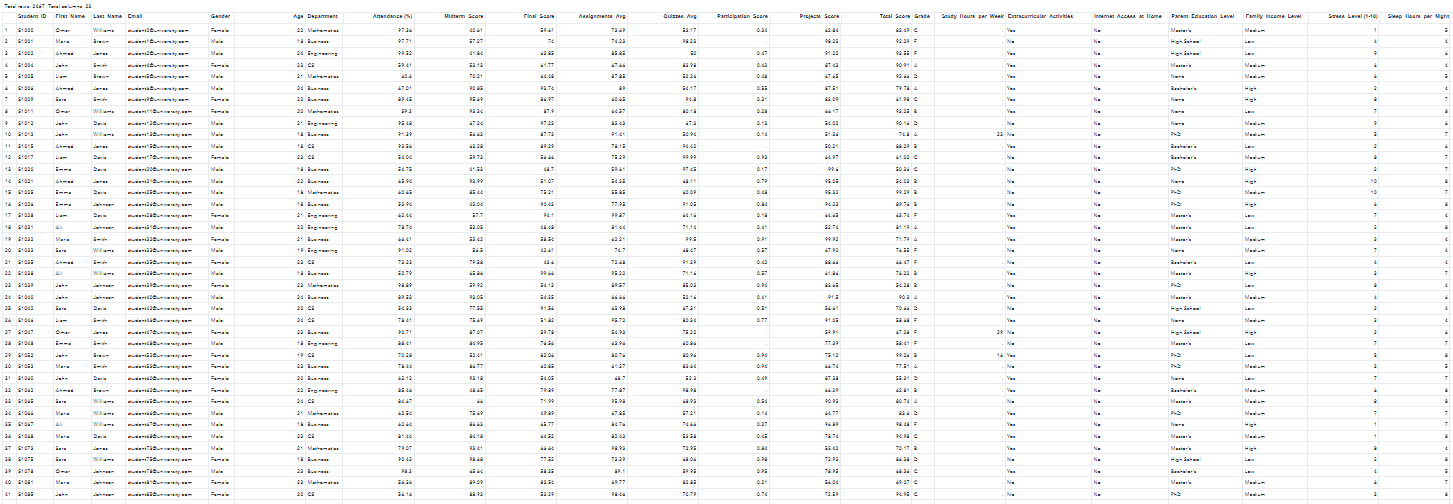
#### 2.3.2: Studenți fără acces la internet acasă

**Cerință:** *Extrageți studenții care nu au acces la internet acasă.*

**Rezolvare:**



Figură 29. Cod SAS pentru extragerea studenților fără acces la internet



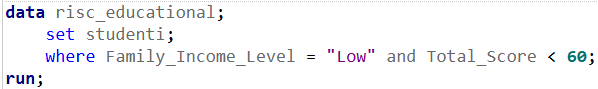
Figură 30. Output – primii 10 studenți fără internet acasă

Am extras un subset format din studenții care au declarat că nu au acces la internet acasă (Internet\_Access\_at\_Home = "No"). Acest grup poate fi analizat separat în comparație cu cei care beneficiază de conectivitate constantă.

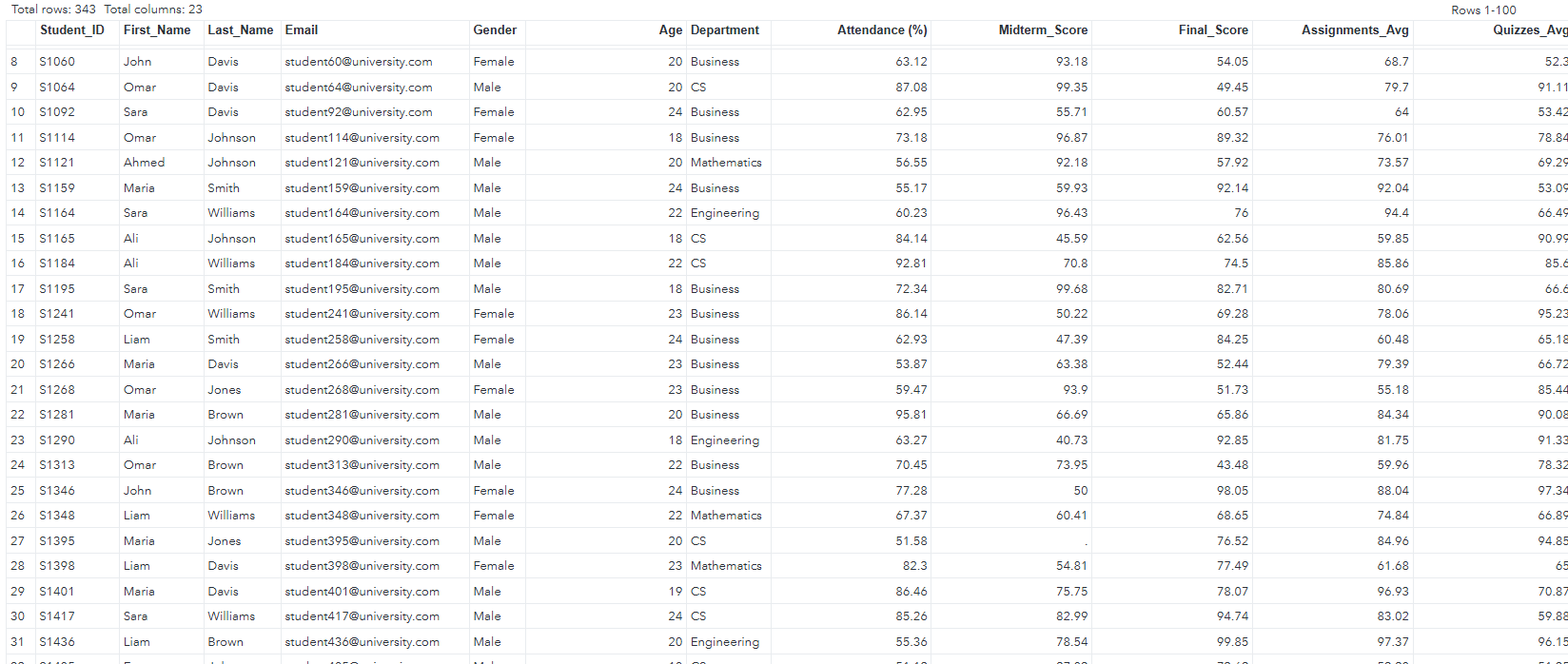
#### 2.3.3: Studenți din familii cu venit scăzut și scor sub 60

**Cerință:** *Întocmiți o analiză asupra studenților aflați în risc educațional.*

**Rezolvare:**



Figură 31. Cod SAS pentru selecția cazurilor cu risc educațional



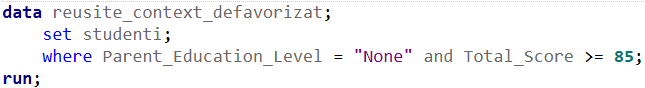
Figură 32. Output – studenți având risc educațional scăzut

Pentru a identifica posibile cazuri de risc educațional, s-a creat un subset alcătuit din studenții provenind din familii cu venituri reduse care au obținut un scor sub 60. Această combinație evidențiază o categorie vulnerabilă atât din punct de vedere socio-economic, cât și educațional.

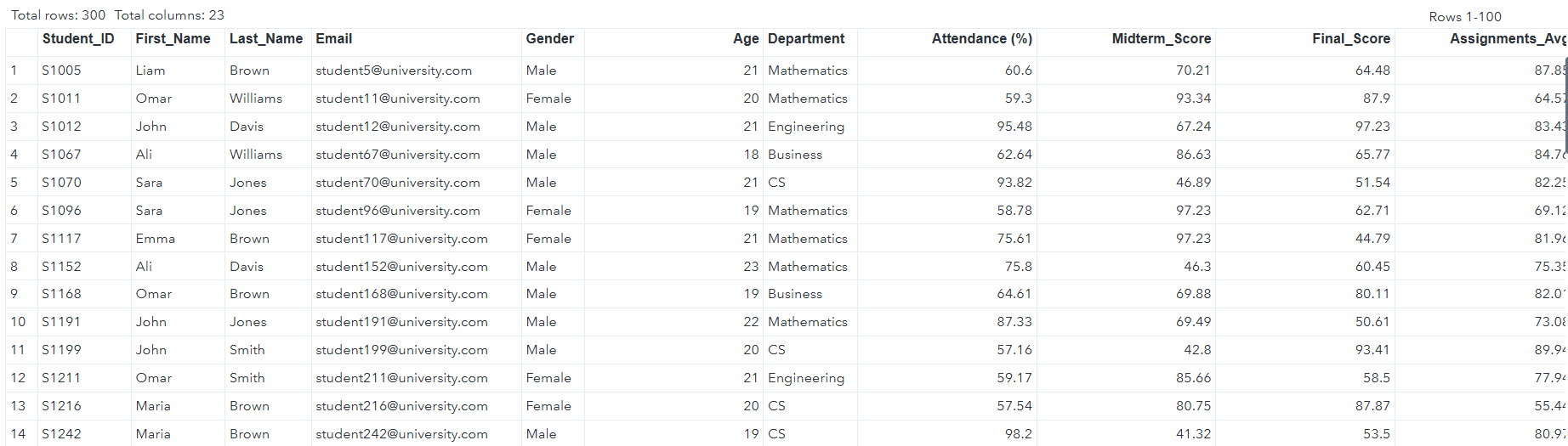
#### 2.3.4: Studenți cu părinți fără studii și scor ridicat (≥ 85)

**Cerință:** *Identificați exemplele de reușită academică în contexte defavorizate.*

**Rezolvare:**



Figură 33. Cod SAS pentru identificarea reușitelor în contexte defavorizate



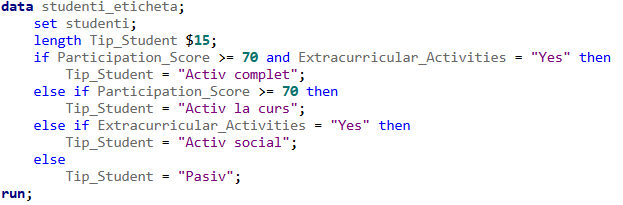
Figură 34. Output – Studeți cu reușite academice având contexte defavorizate

Pentru a evidenția cazurile pozitive, s-a extras un subset de studenți cu scor final ridicat (≥ 85), dar care provin din medii familiale cu părinți fără studii.

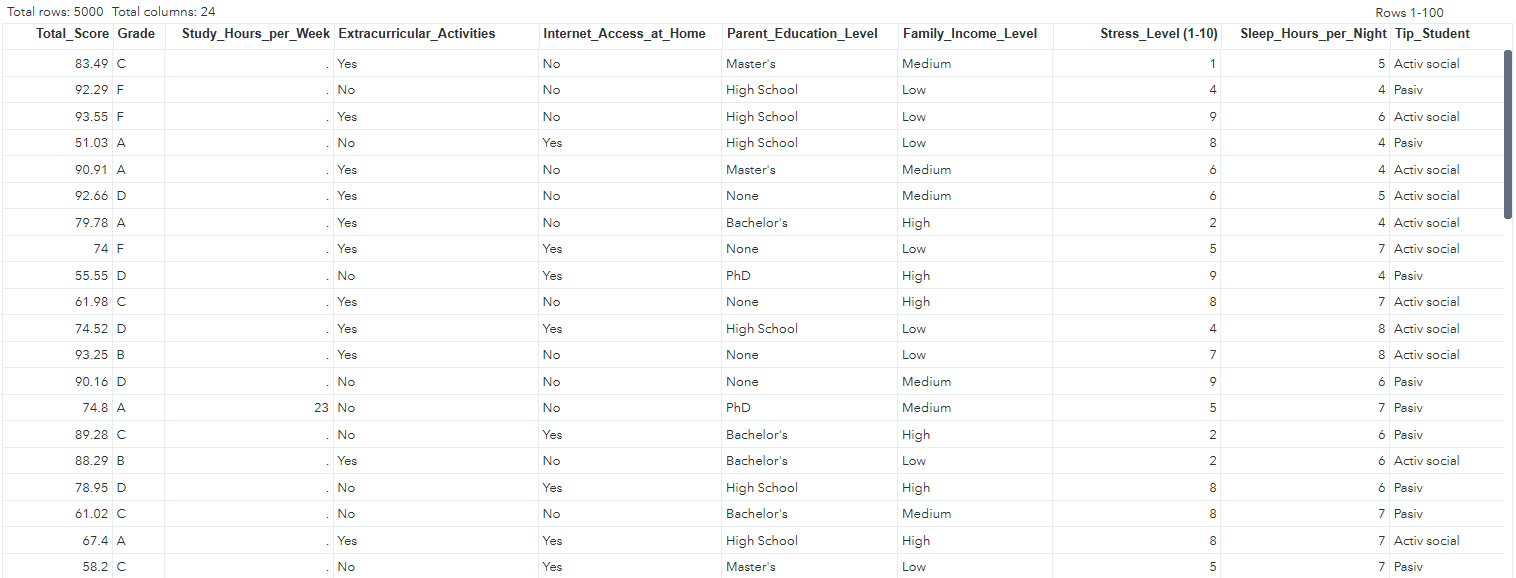
### 2.4. Procesare condițională și crearea unei noi variabile categorice

#### 2.4.1. Etichetarea studenților în funcție de implicare

**Cerință:** *Clasificați studenții în funcție de nivelul de implicare educațională și socială, prin combinarea scorului de participare la cursuri și a implicării în activități extracurriculare.*



Figură 35. Cod SAS pentru etichetarea studenților în funcție de implicare

****

Figură 36. Output – eticheta Tip\_Student aplicată

**Rezolvare:**

Pentru a obține o imagine clară asupra tipurilor de implicare ale studenților, am utilizat o structură condiționată *IF-THEN-ELSE* care etichetează fiecare student astfel:

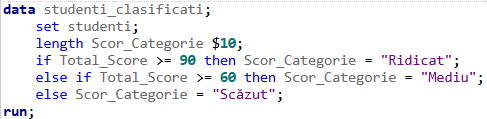
* „Activ complet” – participare ≥ 70 și activ extracurricular
* „Activ la curs” – doar participare ≥ 70
* „Activ social” – doar activ extracurricular
* „Pasiv” – niciuna

Această etichetare ne ajută la realizarea unor analize comparative între tipologii de studenți și pentru a ne da seama ce pattern-uri comportamentale avem noi, ca studenți

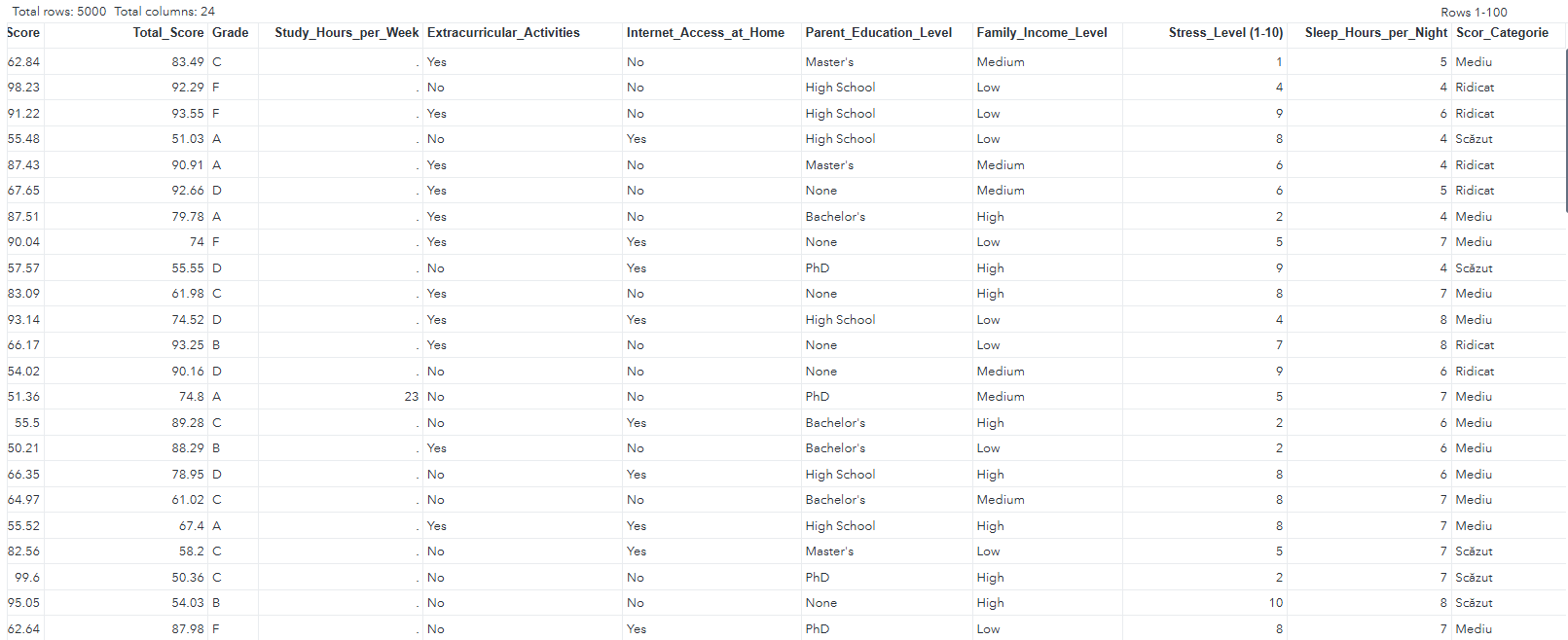
#### 2.4.2. Marcarea studenților eligibili pentru burse

**Cerință:** *Identificați automat studenții eligibili pentru burse, pe baza unor criterii de performanță și venit.*

**Rezolvare:**

****

Figură 37. Cod SAS pentru marcarea eligibilității la bursă

****

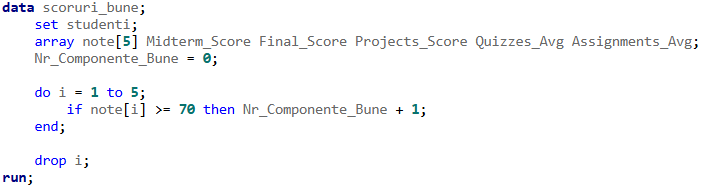
Figură 38. Output – studenții marcați cu pentru Eligibil\_Bursa

Pentru a selecta studenții eligibili pentru burse, pentru care s-a aplicat condiția Total\_Score ≥ 85. Cu ajutorul instrucțiunii *IF-THEN-ELSE*, a fost creată variabila Eligibil\_Bursa pentru cei care îndeplinesc criteriile de a primi bursă

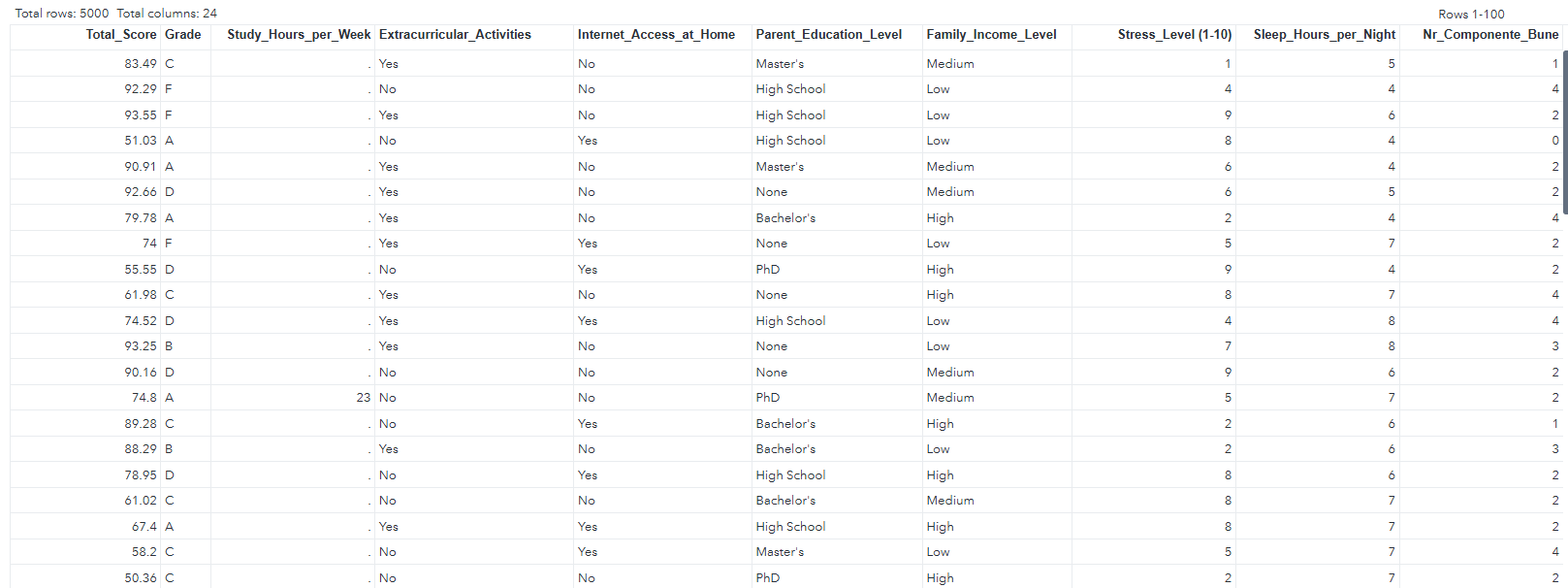
#### 2.4.3. Numărarea componentelor unde studentul a avut peste 70 de puncte

**Cerință:** *Calculați numărului de componente de evaluare în care un student a avut o performanță bună, definită ca scor ≥ 70.*

**Rezolvare:**

****

Figură 39. Cod SAS pentru numărarea componentelor cu scor ≥ 70

****

Figură 40. Output – numărul de componente cu performanță ridicată

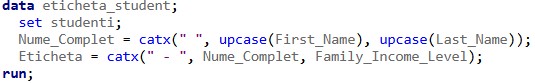
Am folosit o buclă DO pentru a parcurge cele cinci componente de evaluare (Midterm, Final, Projects, Quizzes, Assignments). Cu fiecare iterație, dacă scorul a fost ≥ 70, s-a incrementat contorul Nr\_Componente\_Bune. Această variabilă indică consistența performanței unui student și poate fi utilizată în analiza individuală sau comparativă.

### 2.5. Utilizarea funcțiilor SAS

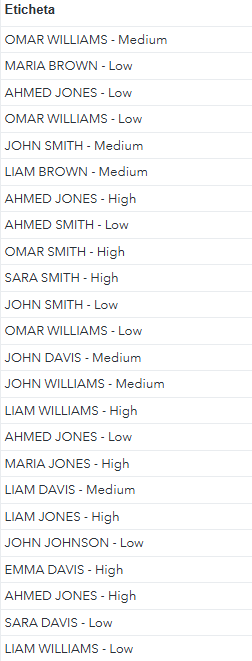
#### 2.5.1. Etichetă personalizată cu numele complet și tipul venitului

**Cerință:** *Obțineți o nouă coloană denumită Etichetă care reprezintă numele complet al studentului și venitul familiei sale.*

**Rezolvare:**



Figură 41. Cod SAS pentru eticheta studentului



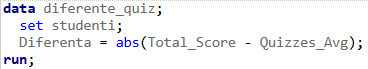
Figură 42. Output – etichete compuse

* upcase() → transformă textul în majuscule (pentru standardizare)
* catx(" ", ...) → concatenează cu separator spațiu între prenume și nume
* catx(" - ", ...) → combină numele complet cu nivelul venitului, folosind separator vizual

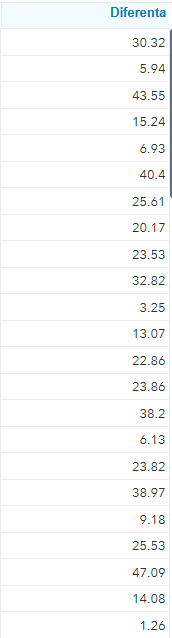
#### 2.5.2. Calcularea diferenței dintre scorul final și cel mediu al quizurilor (ABS)

**Cerință:** *Determinați diferența absolută dintre scorul total (Total\_Score) și media quizurilor (Quizzes\_Avg) pentru fiecare student.*

**Rezolvare:**



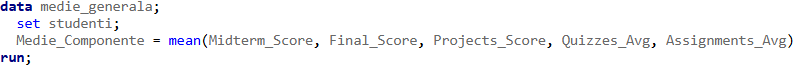
Figură 43. Cod SAS pentru calculul diferenței față de quizuri



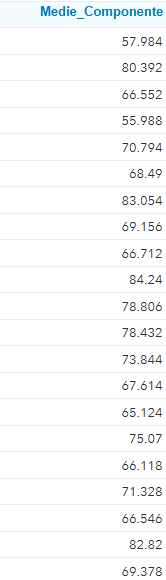
Figură 44. Output – diferențe absolute

#### 2.5.3. Calcularea diferenței dintre scorul final și cel mediu al quizurilor (ABS)

**Cerință:** *Calculați media aritmetică simplă a celor 5 componente de evaluare: Midterm, Final, Projects, Quizzes și Assignments.*



Figură 45. Cod SAS pentru calculul mediei generale pe componente



Figură 46. Output – media componentelor

Funcția mean calculează media doar din valorile nenule, astfel evităm erorile dacă lipsesc unele scoruri, și este utilă pentru clasificarea și analizarea generală a studenților.

### 2.6. Combinarea seturilor de date prin proceduri specifice SAS și SQL

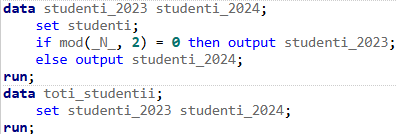
În analiza datelor, este adesea necesar să combinăm informații provenite din surse diferite. Cele mai comune modalități sunt: concatenarea (adaugă rânduri din mai multe tabele care au aceeași structuă), îmbinarea (combină 2 tabele pe baza unui identificator comun) și join (alăturări flexibile între tabele, ca în SQL).

#### 2.6.1 – Concatenarea a două fișiere de studenți (SET)

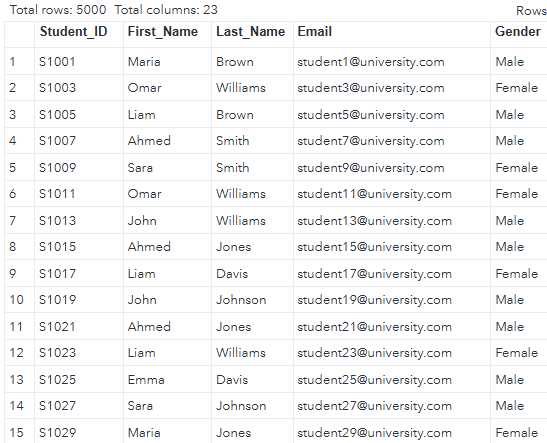
**Cerință:** *Se dorește simularea adăugării a două loturi de studenți (ex. anii 2023 și 2024) într-un tabel unic. Această operație este utilă atunci când se colectează date în serii temporale sau pe grupe.*

**Rezolvare:**

Pentru această demonstrație, s-au împărțit observațiile din tabelul studenti în două subseturi: studenti\_2023 și studenti\_2024. Apoi, s-a realizat concatenarea cu instrucțiunea SET. Acesta este un exerciție demonstrativ, deoarece nu dispunem de generații diferite de studenți, însă pentru un set real de date, această concatenare ne poate ajuta la analiza comparativă între generații.



Figură 47. Cod SAS pentru concatenarea fișierelor de studenți



Figură 48. Output – toți studenții combinați într-un singur tabel

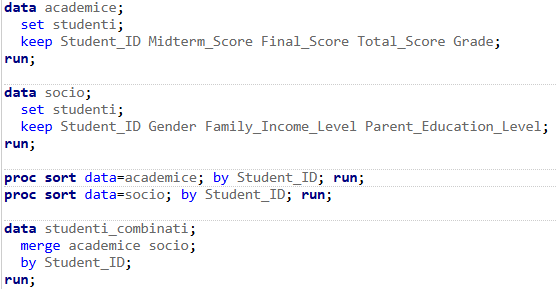
#### 2.6.2 – Îmbinarea a două tabele pe baza ID-ului (MERGE)

**Cerință:** *Separați datele academice de cele socio-demografice și uniți-le pe baza coloanei Student\_ID.*

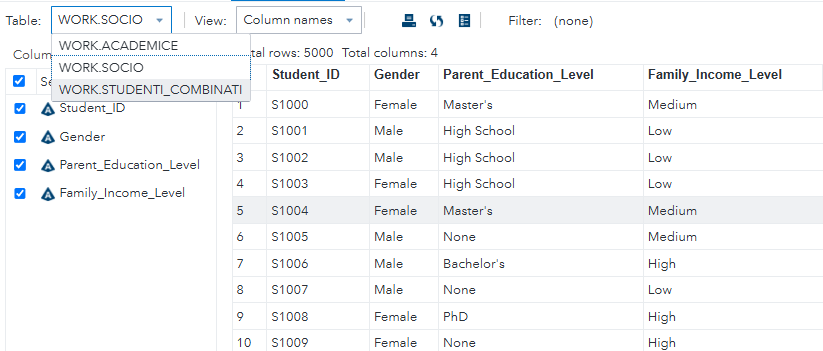
**Rezolvare:**

Acest exercițiu își propune să reunească într-un singur tabel informațiile academice și cele socio-demografice ale fiecărui student. Pentru asta, s-au separat mai întâi două seturi de date: unul cu scorurile și notele, celălalt cu informații precum genul, venitul familiei sau educația părinților. Apoi, prin funcția MERGE, cele două fișiere au fost combinate pe baza unui identificator comun, Student\_ID.

Această procedură este foarte utilă atunci când lucrăm cu fișiere diferite pentru aceeași populație. Ne ajută să avem într-un singur loc toate datele de care avem nevoie pentru analize mai complexe, cum ar fi influența factorilor sociali asupra performanței școlare.

**

Figură 49. Cod SAS pentru îmbinarea datelor academice și socio-economice

****

Figură 50. Output – tabelul combinat cu date socio-demografice.

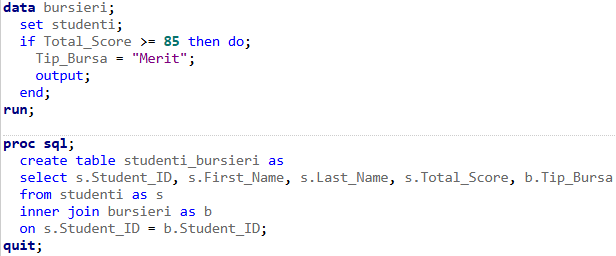
#### 2.6.3 – JOIN cu PROC SQL între studenți și bursieri

**Cerință:** *Se dorește extragerea doar a studenților care îndeplinesc condițiile de bursă și completarea datelor lor generale cu tipul bursei acordate.*

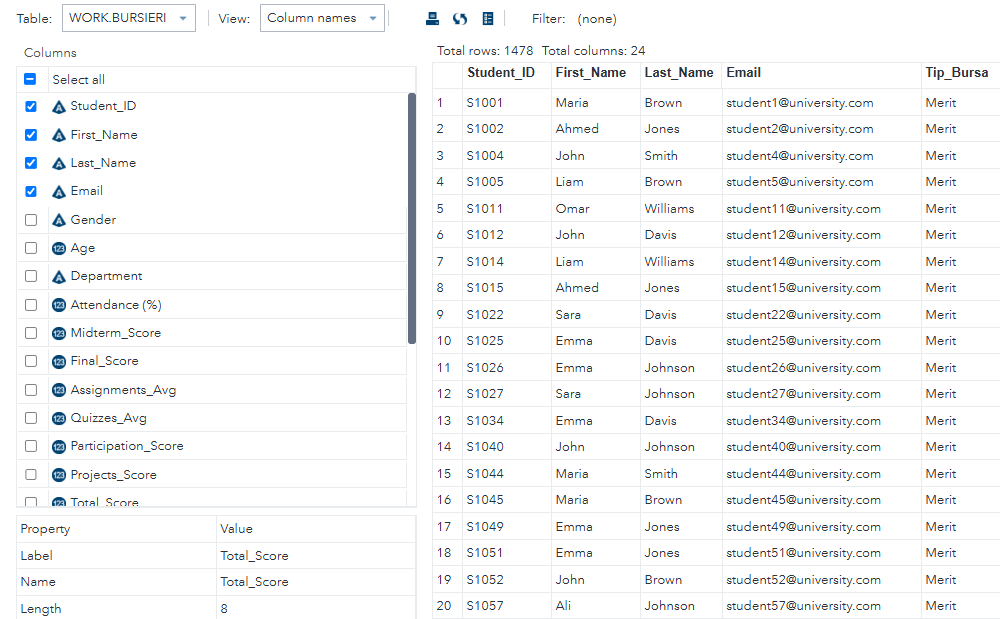
**Rezolvare:**

În acest exercițiu, am folosit PROC SQL pentru a face o selecție a studenților care se califică pentru bursă, pe baza scorului total. Am creat mai întâi un tabel cu doar acei studenți care au avut peste 85 de puncte și le-am atribuit o etichetă de tipul bursei (de exemplu, „Merit”). Apoi, folosind INNER JOIN, am combinat această listă cu tabelul complet al studenților, pentru a obține o variantă finală care conține doar bursierii, împreună cu datele lor personale.

Această metodă este foarte practică atunci când vrem să extragem rapid un grup țintă și să-i combinăm cu alte date – de exemplu, pentru a genera o listă cu bursierii care urmează să fie notificați sau sprijiniți financiar.



Figură 51. Cod SAS pentru alăturare cu PROC SQL JOIN



Figură 52. Output – lista bursierilor

### 2.7. Generarea graficelor

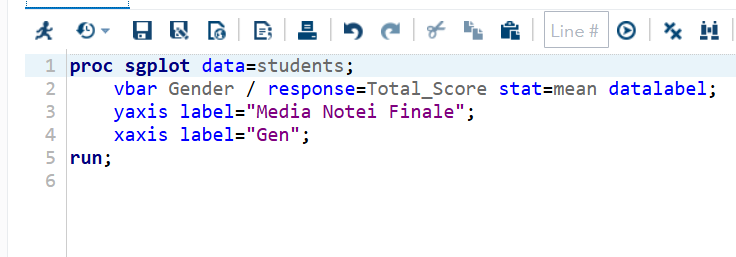
Graficele sunt o metodă esențială pentru a înțelege mai bine structura și tendințele datelor. În această secțiune vom prezentata câteva exemple de grafice utile în analiza datelor educaționale, care evidențiază distribuții, corelații și comparații între variabile.

2.7.1. Bar chart – Media scorurilor totale în funcție de gen

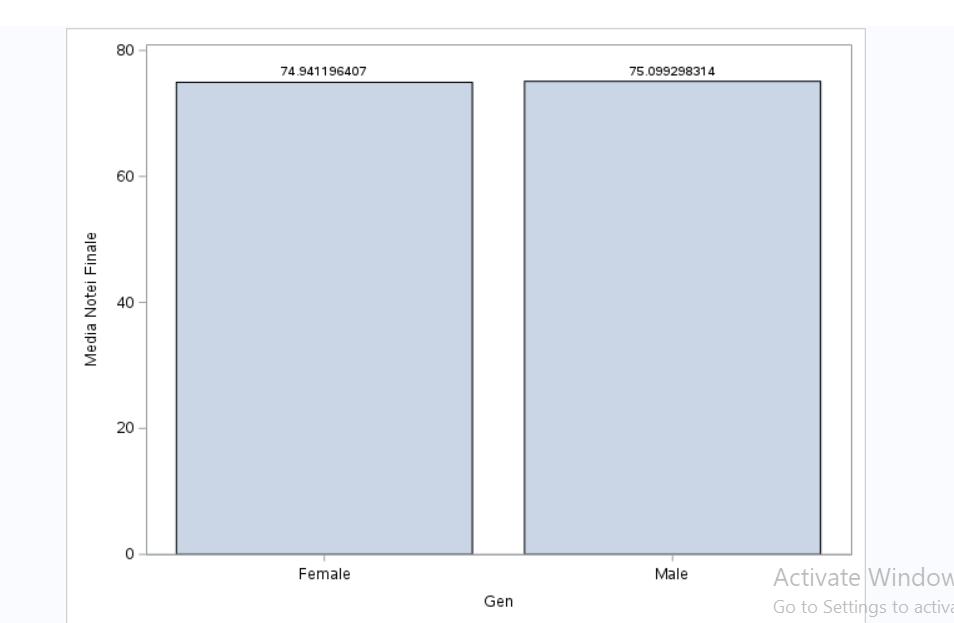
**Cerință:** *Se dorește compararea performanțelor academice între genuri prin afișarea mediei scorului total pentru fiecare categorie de gen (masculin și feminin).*

**Rezolvare:**

Pentru a evidenția diferențele de performanță între genuri, se creează un grafic cu bare verticale care ilustrează media notelor finale pentru fiecare categorie de gen (masculin/feminin). Acest tip de vizualizare este util pentru a observa tendințe generale în performanță în funcție de gen și poate ghida eventuale analize suplimentare sau intervenții educaționale.



Figură 53. Cod pentru generarea graficului cu media scorurilor totale pe gen



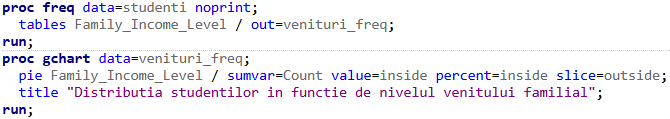
Figură 54. Media scorurilor finale în funcție de gen

Această reprezentare grafică sugerează că genul nu influențează semnificativ performanța finală, cel puțin în cadrul acestui eșantion de date. Ambele medii sunt aproape egale, ceea ce sprijină ideea unui mediu educațional echitabil, în care rezultatele sunt similare indiferent de gen. Această concluzie poate fi utilă în analiza echității academice sau în fundamentarea unor decizii legate de politici educaționale.

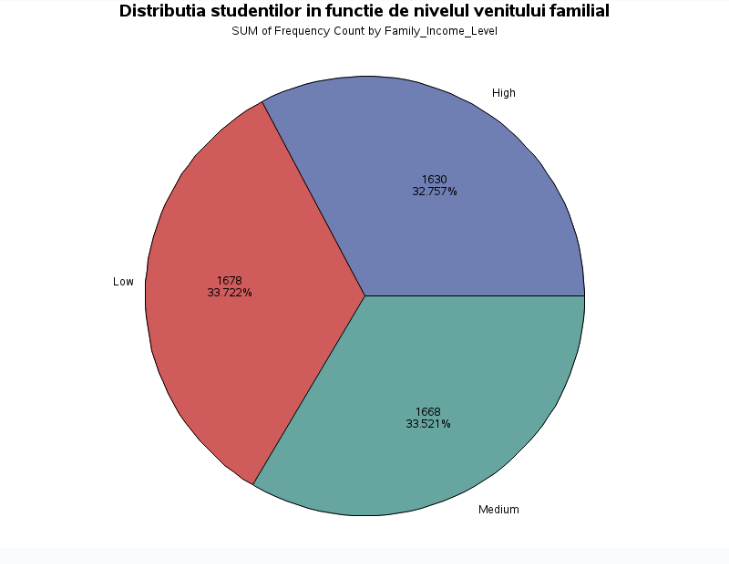
#### 2.7.2. Pie chart – Distribuția studenților pe niveluri de venit

**Cerință:** *Se dorește reprezentarea procentuală a studenților în funcție de nivelul venitului familial (scăzut, mediu, ridicat).*

**Rezolvare:**



Figură 55. Cod pentru generarea distribuției între studenți pe nivelurile de venit

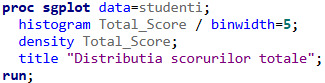


Figură 56. Pie chart – distribuția studenților pe niveluri de venit

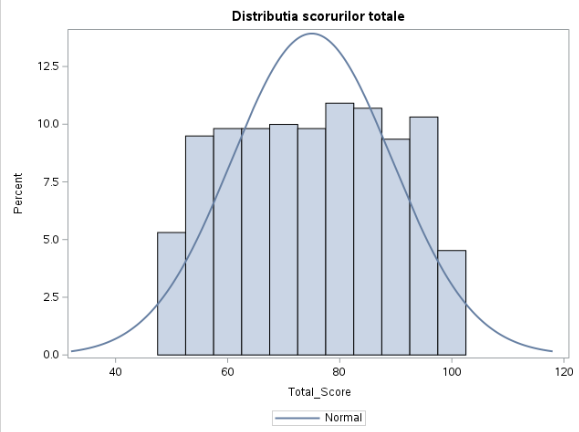
Acest grafic circular reprezintă vizual proporțiile studenților pe categorii de venit. Fiecare felie indică atât procentul, cât și numărul efectiv al studenților care se încadrează într-o anumită clasă socio-economică. Graficul este util pentru a înțelege cât de echilibrată sau dezechilibrată este distribuția financiară a eșantionului. Această distribuție aproape uniformă sugerează că setul de date nu este dezechilibrat din punct de vedere socio-economic

#### 2.7.3. Histogramă – Distribuția scorurilor totale

**Cerință:** *Se dorește analizarea distribuției scorurilor finale pentru a observa dacă valorile sunt concentrate în jurul unei medii sau dispersate.*



Figură 57. Cod - histogramă pentru distribuția scorurilor totale



Figură 58. Histogramă – distribuția scorurilor totale

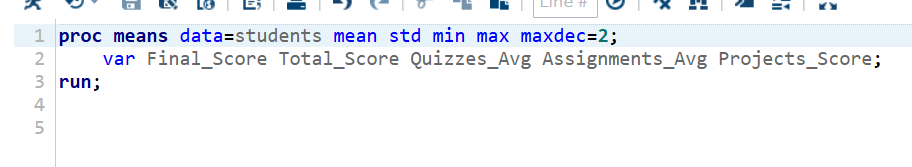
Histograma evidențiază frecvența scorurilor în intervale de câte 5 puncte. Linia de densitate oferă o imagine generală asupra formei distribuției (simetrică, concentrată, sau întinsă). Scorurile sunt distribuite relativ simetric, ceea ce sugerează o populație omogenă din punct de vedere al performanței. Există o ușoară concentrație în zona scorurilor medii (75–85), iar extremele (sub 50 și peste 100) sunt rare. Linia de densitate (curba albastră) confirmă tendința unei distribuții apropiate de normală.

### 2.8. Statistici descriptive

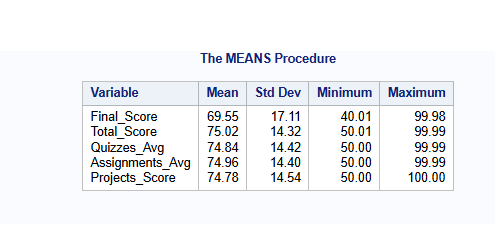
**Cerință:** *Utilizarea procedurilor statistice de bază.*

**Rezolvare:**

Pentru a analiza distribuția și variabilitatea performanței academice, se utilizează procedura PROC MEANS din SAS. Aceasta calculează statistici descriptive esențiale pentru un set de variabile numerice relevante, precum: nota finală (Nota\_Finala), scorul parțial (Scor\_Partial), media temelor (Media\_Teme), media testelor (Media\_Teste), scorul proiectelor (Scor\_Proiect) și scorul total (Total\_Scor).



Figură 59. Statistici descriptive pentru variabilele de performanță academică



Figură 60. Statistici descriptive ale variabilelor de performanță academică

Media scorurilor este relativ ridicată pentru toate categoriile, oscilând între *69.55* (pentru Final\_Score) și *75.02* (pentru Total\_Score), ceea ce sugerează o performanță general bună a studenților. Se observă că Final\_Score are cea mai mică medie și totodată cea mai mare abatere standard (*17.11*), indicând o variație mai largă a notelor la evaluarea finală comparativ cu celelalte componente.

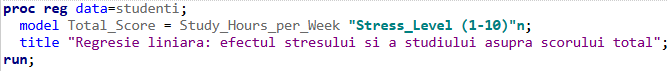
Valorile minime și maxime arată că toți indicatorii evaluați variază între aproximativ *50* și *100*, cu excepția Final\_Score, care are o valoare minimă de *40.01*, semnalând prezența unor performanțe semnificativ mai slabe în cadrul acestui scor.

### 2.9. Corelații între indicatorii de performanță academică

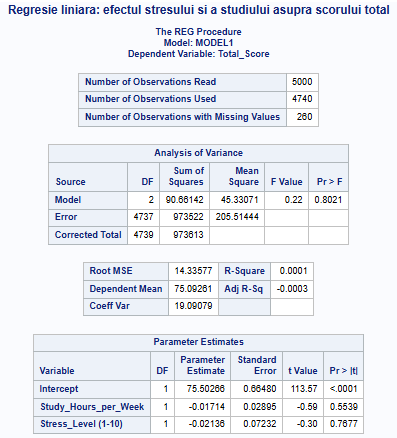
#### 2.9.1. Regresie liniară – Estimarea scorului total

**Cerință:** Să se realizeze analiza impactului orelor de studiu și al stresului asupra scorului total obținut de studenți.

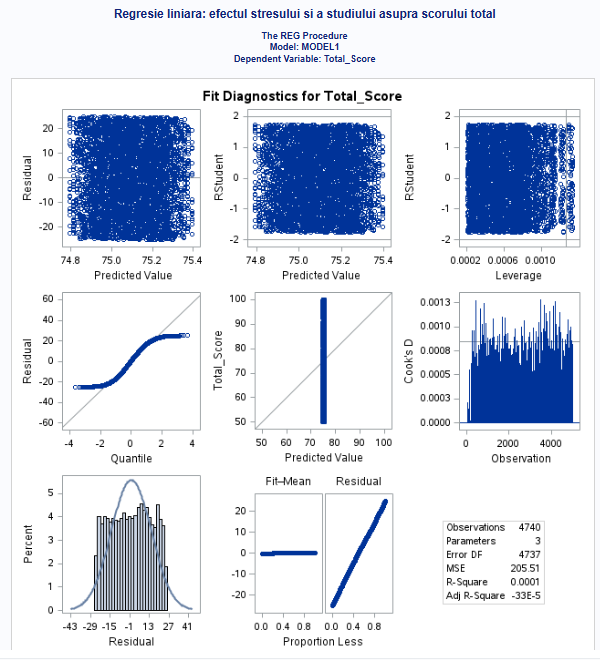
**Rezolvare:**

****

Figură 61. Cod SAS - regresia liniară

****

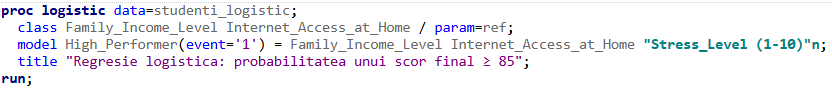
Figură 62. Output 1 - regresia liniară

****

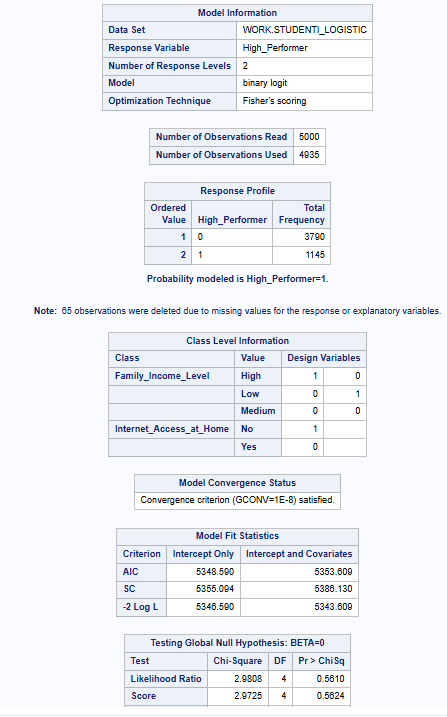
Figură 63. Output 2 - regresia liniară

Modelul de regresie liniară aplicat pentru a analiza influența stresului și a orelor de studiu asupra scorului total nu este semnificativ din punct de vedere statistic (p = 0.8021, R² = 0.0001). Coeficienții ambelor variabile sunt negativi și nesemnificativi, ceea ce sugerează că, în cadrul acestui eșantion, nici nivelul de stres și nici timpul dedicat studiului nu explică variația scorurilor academice. Diagramele de diagnostic confirmă lipsa unei relații clare, indicând o dispersie aleatoare a reziduurilor și un model slab ajustat.

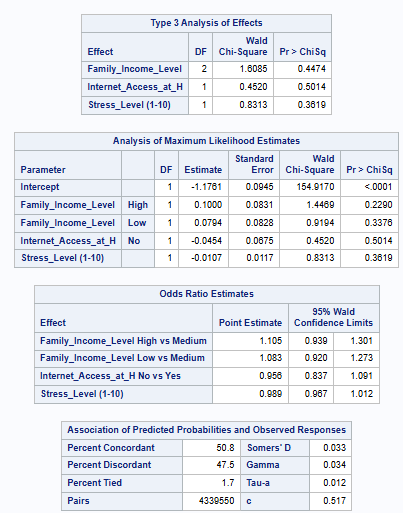
#### 2.9.2. Regresie logistică – Predicția scorului final ridicat



Figură 64. Cod SAS - regresia logistică



Figură 65. Output 1 - regresia logistică



Figură 66. Output 2 - regresia logistică

Modelul de regresie logistică a fost utilizat pentru a prezice probabilitatea ca un student să obțină un scor final ridicat (≥ 85), în funcție de venitul familial, accesul la internet și nivelul de stres. Rezultatele arată că niciunul dintre predictorii incluși nu este semnificativ statistic (p > 0.05), indicând că acești factori nu influențează în mod semnificativ probabilitatea de a obține un scor mare în acest eșantion. Odds ratio-urile calculate sugerează doar ușoare variații în șansele de performanță ridicată în funcție de venit sau accesul la internet, dar intervalele largi de încredere subliniază lipsa robusteții. Prin urmare, modelul nu are putere predictivă relevantă în această formulare.

# CONCLUZIE

Lucrarea de față a avut ca obiectiv principal analizarea performanței academice a studenților prin intermediul unor metode statistice moderne și al prelucrării datelor, utilizând instrumente precum Python (cu ajutorul bibliotecii Streamlit) și SAS. Procesul analitic a fost structurat riguros, parcurgând etapele fundamentale ale unui demers științific: importul și curățarea datelor, transformarea variabilelor, tratarea valorilor lipsă, analiza descriptivă și exploratorie, modelarea relațiilor dintre variabile, precum și aplicarea unor tehnici predictive relevante.

Rezultatele obținute evidențiază importanța unor factori precum scorurile la teme, teste sau proiecte, dar și a unor variabile contextuale (ex. gen, nivel de stres, acces la internet) în explicarea și anticiparea performanței academice. Modelele de regresie liniară și logistică au permis nu doar identificarea relațiilor semnificative dintre variabile, ci și formularea unor previziuni robuste, în special în cazul clasificării binare a performanței (sub și peste medie), cu un nivel de acuratețe ridicat.

De asemenea, utilizarea testului ANOVA și a matricei corelațiilor a oferit o înțelegere aprofundată a modului în care anumite categorii influențează rezultatele numerice, susținând luarea deciziilor bazate pe date concrete. Vizualizările generate au contribuit la claritatea interpretărilor și la o comunicare eficientă a rezultatelor.

În concluzie, proiectul demonstrează potențialul analizei statistice și al științei datelor în sprijinirea procesului educațional, oferind o bază solidă pentru intervenții direcționate, politici educaționale și strategii de îmbunătățire a performanței studenților.

## BIBLIOGRAFIE

1. Materiale puse la dispoziție în cadrul seminarului

## ANEXĂ

### FIGURI

[Figură 1. Interfața aplicației Streamlit pentru analiza vizuală a performanței studenților. 3](#_Toc198075035)

[Figură 2. Procentajul valorilor lipsă pentru fiecare variabilă din setul de date 5](#_Toc198075036)

[Figură 3. Vizualizarea grafică a valorilor lipsă pentru toate variabilele din setul de date utilizând heatmap-ul 6](#_Toc198075037)

[Figură 4. Figura 4. Selectarea metodei de transformare a variabilelor categorice 7](#_Toc198075038)

[Figură 5. Selectarea metodei de scalare pentru variabilele numerice 8](#_Toc198075039)

[Figură 6. Selectarea metodei de detectare a valorilor extreme 9](#_Toc198075040)

[Figură 7. Distribuția frecvenței pe gen în cadrul setului de date 10](#_Toc198075041)

[Figură 8. Distribuția frecvenței studenților pe departamente 11](#_Toc198075042)

[Figură 9. Distribuția scorurilor 12](#_Toc198075043)

[Figură 10. Matricea corelațiilor între variabilele numerice 13](#_Toc198075044)

[Figură 11. Compararea mediilor scorului total în funcție de nivelul de stres (1–10) 14](#_Toc198075045)

[Figură 12. Distribuția numărului de ore de somn pe noapte în funcție de gen și testul ANOVA 16](#_Toc198075046)

[Figură 13. Rezultatele regresiei liniare pentru predicția variabilei Total\_Score 18](#_Toc198075047)

[Figură 14. Predictori utilizați în regresia logistică pentru clasificarea performanței 19](#_Toc198075048)

[Figură 15. Matricea de confuzie 20](#_Toc198075049)

[Figură 16. Raport de clasificare – Regresie logistică (fără SelectKBest) 20](#_Toc198075050)

[Figură 17. Curba ROC – Regresie logistică 21](#_Toc198075051)

[Figură 18. Selectarea celor mai relevanți predictori folosind SelectKBest 22](#_Toc198075052)

[Figură 19. Matricea de confuzie după aplicarea SelectKBest 22](#_Toc198075053)

[Figură 20. Performanța modelului de regresie logistică folosind predictorii selectați 23](#_Toc198075054)

[Figură 21. Importul fișierului CSV într-un set de date SAS 23](#_Toc198075055)

[Figură 22. izualizarea structurii tabelului WORK.STUDENTS după importul fișierului CSV 24](#_Toc198075056)

[Figură 23. Definirea formatelor personalizate pentru variabilele Gender, Extracurricular\_Activities, Internet\_Access\_at\_Home, Parent\_Education\_Level și Family\_Income\_Level 25](#_Toc198075057)

[Figură 24. Aplicarea formatelor în cadrul PROC PRINT 25](#_Toc198075058)

[Figură 25. Aplicare PROC FREQ 26](#_Toc198075059)

[Figură 26. Output PROC FREQ 26](#_Toc198075060)

[Figură 27. Cod SAS pentru extragerea studenților cu scor final ≥ 70 27](#_Toc198075061)

[Figură 28. Output – primii 10 studenți incluși în subsetul studenti\_buni 27](#_Toc198075062)

[Figură 29. Cod SAS pentru extragerea studenților fără acces la internet 28](#_Toc198075063)

[Figură 30. Output – primii 10 studenți fără internet acasă 28](#_Toc198075064)

[Figură 31. Cod SAS pentru selecția cazurilor cu risc educațional 28](#_Toc198075065)

[Figură 32. Output – studenți având risc educațional scăzut 29](#_Toc198075066)

[Figură 33. Cod SAS pentru identificarea reușitelor în contexte defavorizate 29](#_Toc198075067)

[Figură 34. Output – Studeți cu reușite academice având contexte defavorizate 29](#_Toc198075068)

[Figură 35. Cod SAS pentru etichetarea studenților în funcție de implicare 30](#_Toc198075069)

[Figură 36. Output – eticheta Tip\_Student aplicată 30](#_Toc198075070)

[Figură 37. Cod SAS pentru marcarea eligibilității la bursă 31](#_Toc198075071)

[Figură 38. Output – studenții marcați cu pentru Eligibil\_Bursa 31](#_Toc198075072)

[Figură 39. Cod SAS pentru numărarea componentelor cu scor ≥ 70 32](#_Toc198075073)

[Figură 40. Output – numărul de componente cu performanță ridicată 32](#_Toc198075074)

[Figură 41. Cod SAS pentru eticheta studentului 32](#_Toc198075075)

[Figură 42. Output – etichete compuse 33](#_Toc198075076)

[Figură 43. Cod SAS pentru calculul diferenței față de quizuri 34](#_Toc198075077)

[Figură 44. Output – diferențe absolute 34](#_Toc198075078)

[Figură 45. Cod SAS pentru calculul mediei generale pe componente 34](#_Toc198075079)

[Figură 46. Output – media componentelor 35](#_Toc198075080)

[Figură 47. Cod SAS pentru concatenarea fișierelor de studenți 36](#_Toc198075081)

[Figură 48. Output – toți studenții combinați într-un singur tabel 36](#_Toc198075082)

[Figură 49. Cod SAS pentru îmbinarea datelor academice și socio-economice 37](#_Toc198075083)

[Figură 50. Output – tabelul combinat cu date socio-demografice. 37](#_Toc198075084)

[Figură 51. Cod SAS pentru alăturare cu PROC SQL JOIN 38](#_Toc198075085)

[Figură 52. Output – lista bursierilor 38](#_Toc198075086)

[Figură 53. Cod pentru generarea graficului cu media scorurilor totale pe gen 39](#_Toc198075087)

[Figură 54. Media scorurilor finale în funcție de gen 39](#_Toc198075088)

[Figură 55. Cod pentru generarea distribuției între studenți pe nivelurile de venit 40](#_Toc198075089)

[Figură 56. Pie chart – distribuția studenților pe niveluri de venit 40](#_Toc198075090)

[Figură 57. Cod - histogramă pentru distribuția scorurilor totale 41](#_Toc198075091)

[Figură 58. Histogramă – distribuția scorurilor totale 41](#_Toc198075092)

[Figură 59. Statistici descriptive pentru variabilele de performanță academică 42](#_Toc198075093)

[Figură 60. Statistici descriptive ale variabilelor de performanță academică 42](#_Toc198075094)

[Figură 61. Cod SAS - regresia liniară 42](#_Toc198075095)

[Figură 62. Output 1 - regresia liniară 43](#_Toc198075096)

[Figură 63. Output 2 - regresia liniară 43](#_Toc198075097)

[Figură 64. Cod SAS - regresia logistică 44](#_Toc198075098)

[Figură 65. Output 1 - regresia logistică 44](#_Toc198075099)

[Figură 66. Output 2 - regresia logistică 45](#_Toc198075100)