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

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). 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

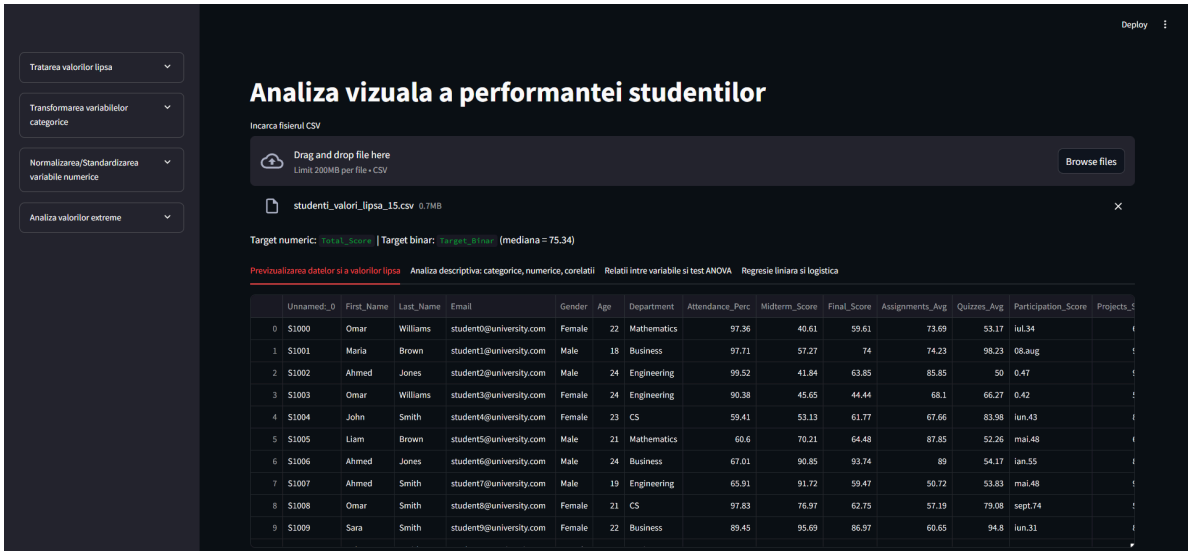
- **Student_ID**: Identificator unic al studentului.
- **First_Name**, **Last_Name**: Numele complet al studentului.
- **Email**: Adresa de email a studentului.
- **Gender**: Sexul biologic al studentului (*male* sau *female*).
- **Age**: Vârsta studentului în ani.
- **Department**: Facultatea sau departamentul 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.

1. 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.



	Unnamed:_0	First_Name	Last_Name	Email	Gender	Age	Department	Attendance_Perc	Midterm_Score	Final_Score	Assignments_Avg	Quizzes_Avg	Participation_Score	Projects_S
0	S1000	Omar	Williams	student0@university.com	Female	22	Mathematics	97.36	40.61	59.61	73.69	53.17	iul.34	
1	S1001	Maria	Brown	student1@university.com	Male	18	Business	97.71	57.27	74	74.23	98.23	08.aug	
2	S1002	Ahmed	Jones	student2@university.com	Male	24	Engineering	99.52	41.84	63.85	85.85	50	0.47	
3	S1003	Omar	Williams	student3@university.com	Female	24	Engineering	90.38	45.65	44.44	68.1	66.27	0.42	
4	S1004	John	Smith	student4@university.com	Female	23	CS	59.41	53.13	61.77	67.66	83.98	iun.43	
5	S1005	Liam	Brown	student5@university.com	Male	21	Mathematics	60.6	70.21	64.48	87.85	52.26	mai.48	
6	S1006	Ahmed	Jones	student6@university.com	Male	24	Business	67.01	90.85	93.74	89	54.17	ian.55	
7	S1007	Ahmed	Smith	student7@university.com	Male	19	Engineering	65.91	91.72	59.47	50.72	53.83	mai.48	
8	S1008	Omar	Smith	student8@university.com	Female	21	CS	97.83	76.97	62.75	57.19	79.08	sept.74	
9	S1009	Sara	Smith	student9@university.com	Female	22	Business	89.45	95.69	86.97	60.65	94.8	iun.31	

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:

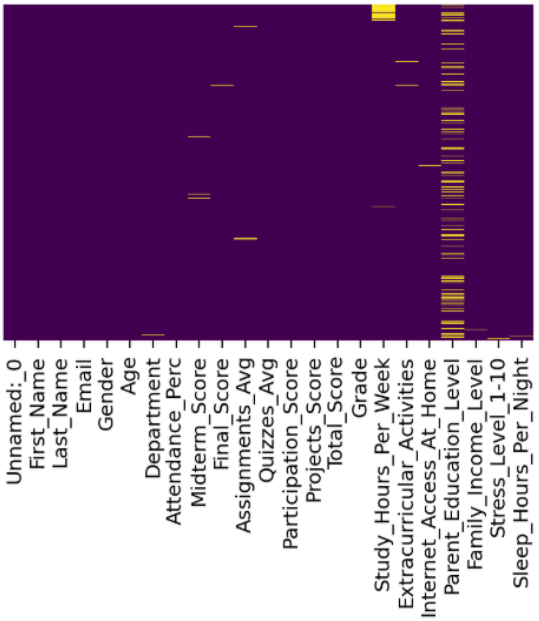
În orice analiză de date reale, valorile lipsă sunt frecvente și pot duce la rezultate incorecte. Dacă nu sunt tratate corespunzător, pot distorsiona interpretările și deciziile luate pe baza datelor.

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.

Valori lipsa

	0
Age	0.0400
Department	0.0600
Attendance_Perc	0.4000
Midterm_Score	0.7400
Final_Score	0.4800
Assignments_Avg	0.4400
Quizzes_Avg	0.3600
Participation_Score	0.3600
Projects_Score	0.2200
Total_Score	0.1000
Grade	0.1600
Study_Hours_Per_Week	4.7600
Extracurricular_Activities	0.3800
Internet_Access_At_Home	0.5000
Parent_Education_Level	21.0800
Family_Income_Level	0.4800
Stress_Level_1-10	0.3400
Sleep_Hours_Per_Night	0.2800

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

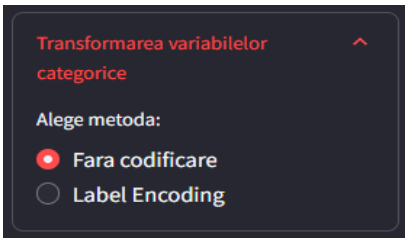
Cerință: *După încărcarea și preprocesarea setului de date, permiteți utilizatorului transformarea variabilelor categorice în format numeric, pentru ca acestea să poată fi utilizate în analize statistice și modele predictive.*

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]
```

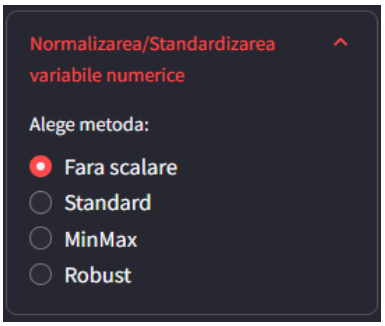
1.4. Normalizarea/Standardizarea variabilelor numerice

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])
```

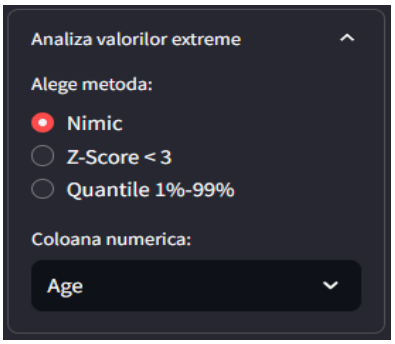
1.5. Eliminarea valorilor extreme (outliers)

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.*

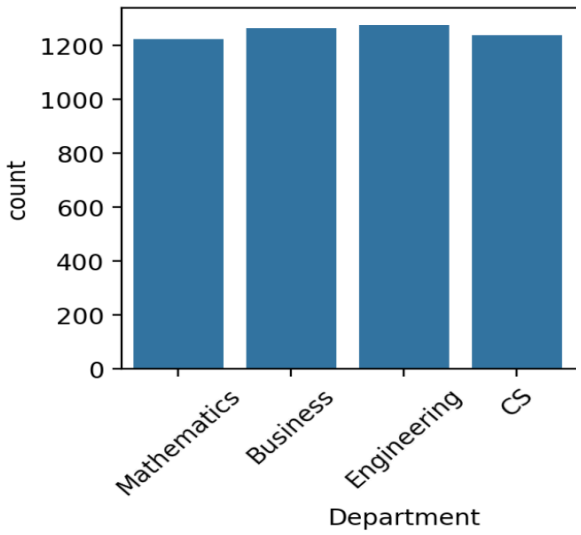
Rezolvare:

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":

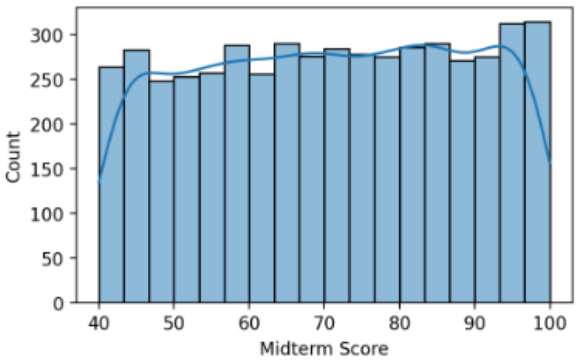
```
categorical_cols = [
    "Gender", "Department", "Grade", "Extracurricular_Activities",
    "Internet_Access_At_Home", "Family_Income_Level", "Stress_Level_1-10" ]
categorical_cols = [col for col in categorical_cols if col in df.columns]
selected_cat = st.selectbox("Coloană:", categorical_cols)
fig, 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.*

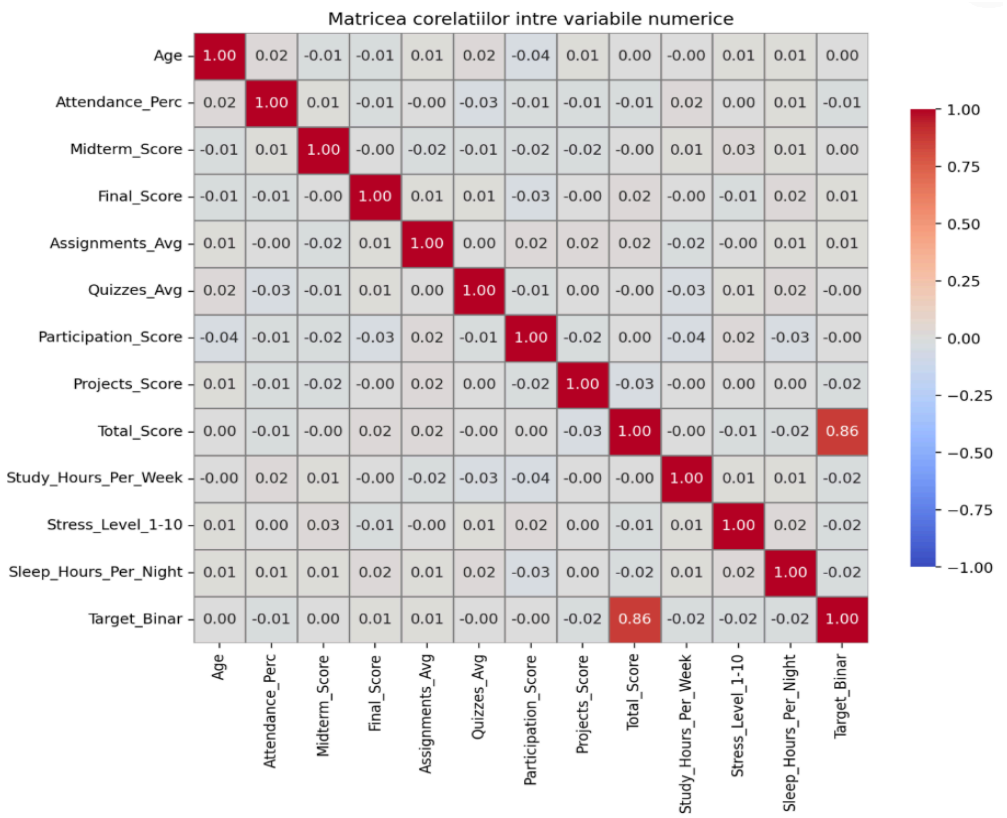
Rezolvare:

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)
```

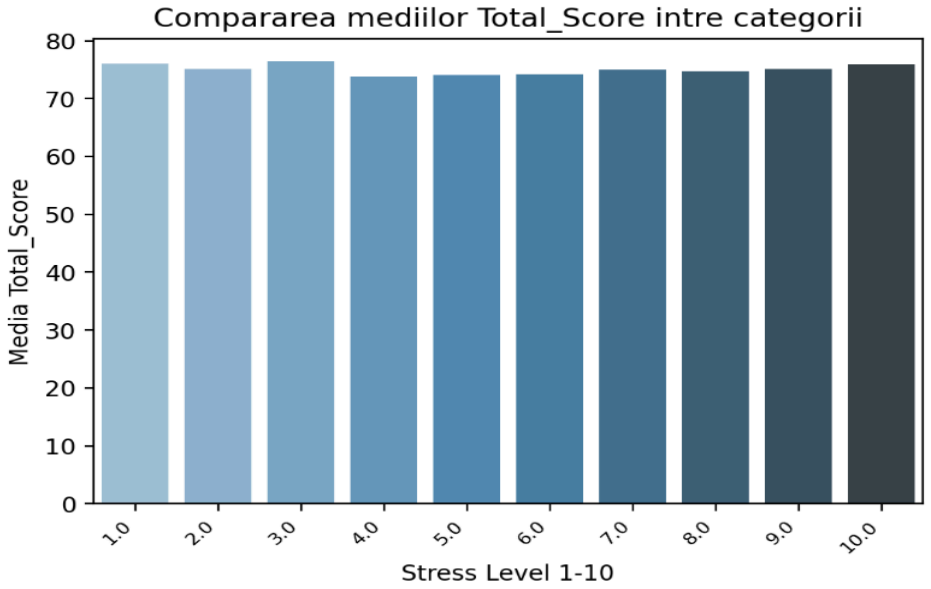
1.6.4. Compararea mediilor unei variabile numerice pe categorii

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.

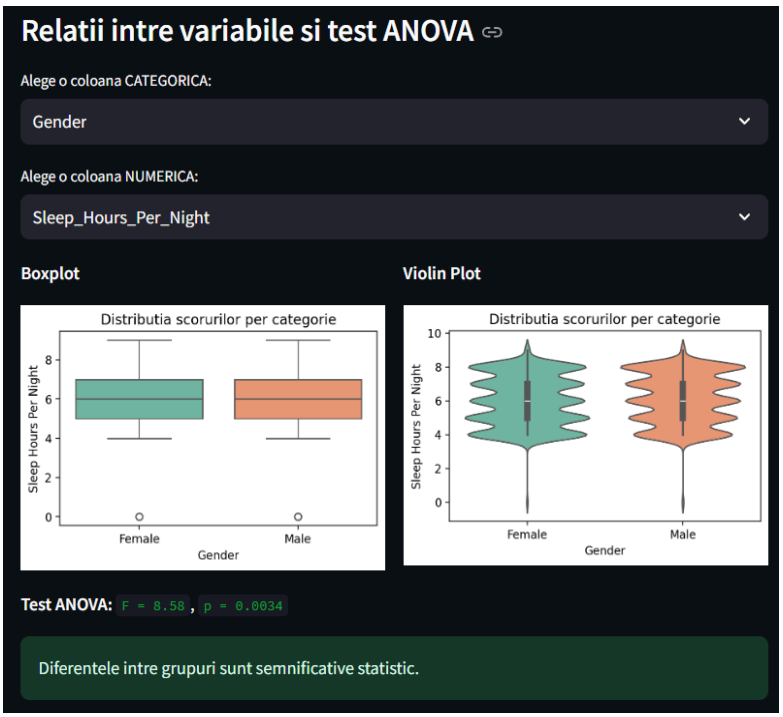
Rezolvare:

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.

	coef	std err	t	P>|t|	[0.025	0.975]
const	77.3849	2.951	26.219	0	71.599	83.171
Age	-0.0059	0.101	-0.059	0.953	-0.204	0.192
Attendance_Perc	-0.0036	0.014	-0.268	0.789	-0.03	0.023
Midterm_Score	0.0029	0.011	0.263	0.793	-0.019	0.025
Final_Score	0.0197	0.012	1.706	0.088	-0.003	0.042
Projects_Score	-0.0276	0.014	-2.011	0.044	-0.055	-0.001
Study_Hours_Per_Week	0.008	0.026	0.31	0.757	-0.043	0.059
Stress_Level_1-10	-0.0309	0.071	-0.436	0.663	-0.17	0.108
Sleep_Hours_Per_Night	-0.2502	0.14	-1.786	0.074	-0.525	0.025

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

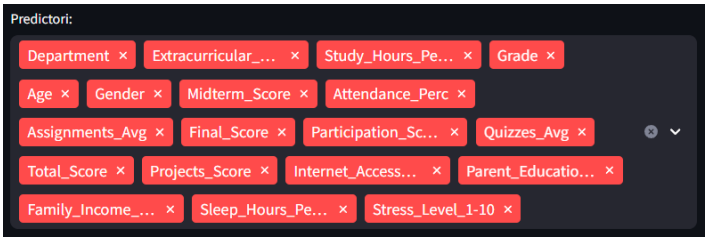
Cerință: Aplicația trebuie să permită utilizatorului antrenarea unui model de regresie logistică pentru a prezice probabilitatea ca un student să obțină un scor total peste medie.

Rezolvare:

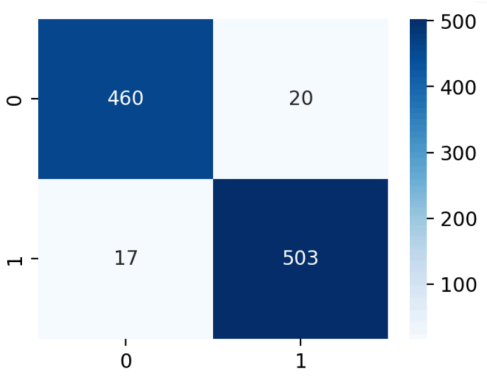
Pentru îndeplinirea cerinței, aplicația construiește o variabilă țintă binară (Target_Binar) pe baza valorii mediane a scorului total (Total_Score), atribuind valoarea 1 studenților cu scoruri peste medie și 0 celor cu scoruri sub medie. Utilizatorul poate selecta manual predictorii dintr-o listă ce 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.

Raport de clasificare

	precision	recall	f1-score	support
0	0.9644	0.9583	0.9613	480
1	0.9618	0.9673	0.9645	520
accuracy	0.963	0.963	0.963	0.963
macro avg	0.9631	0.9628	0.9629	1,000
weighted avg	0.963	0.963	0.963	1,000

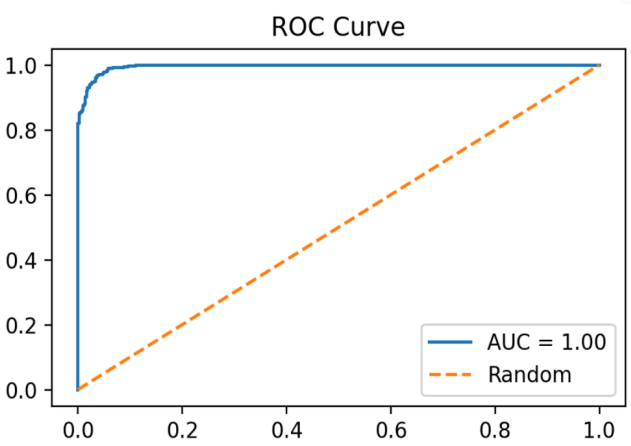
Acuratete: 0.96

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

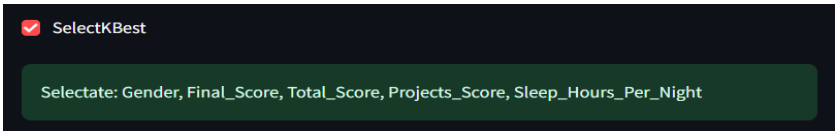
Cerință: *Pentru a obține un model predictiv mai precis și mai eficient, faceți ca aplicația permită selecția automată a celor mai relevante variabile (predictori) utilizate în clasificare.*

Rezolvare:

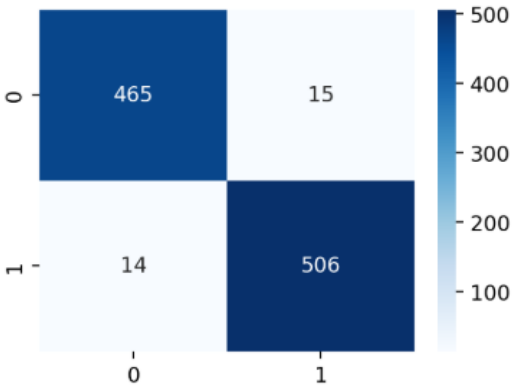
Am implementat opțiunea *SelectKBest*, disponibilă sub forma unei bifări în aplicație, care permite selecția automată a celor mai semnificative variabile numerice pe baza scorurilor statistice (ex: funcția f_regression). La activarea acestei opțiuni, aplicația reține doar predictorii cu cea mai mare 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.



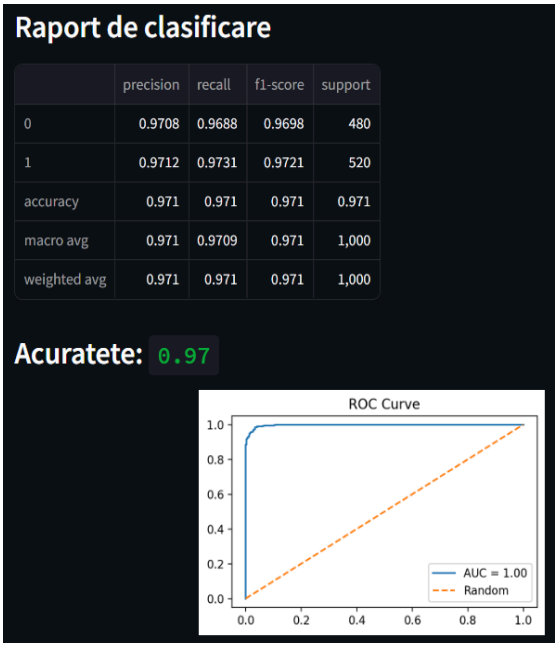
Figură 18. Selectarea celor mai relevanți predictori folosind SelectKBest



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

2. Pachetul SAS

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

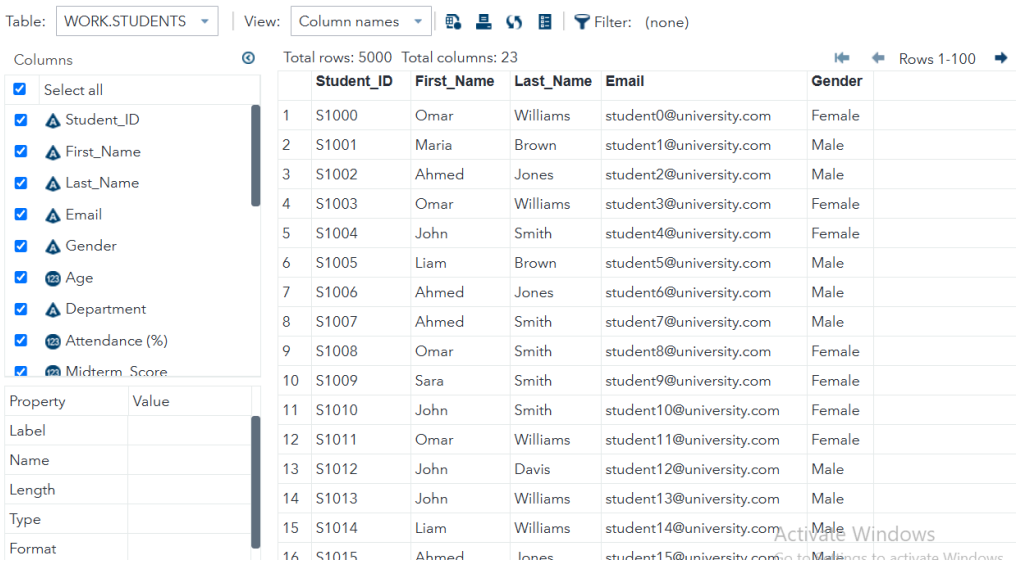
Cerință: *Crearea unui set de date SAS dintr-un fișier extern (.csv).*

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.

```
proc import datafile="/home/u64202433/Pachete_Software_Proiect/Students_Grading_Dataset.csv"
    out=work.students
    dbms=csv
    replace;
    delimiter=';';
    getnames=yes;
run;
```

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.

```
proc format;
    value $gender_fmt
        "Male" = "Barbat"
        "Female" = "Femeie";
    value $yn_fmt
        "Yes" = "Da"
        "No"  = "Nu";
    value $edu_fmt
        "High School" = "Liceu"
        "Bachelor's" = "Licenta"
        "Master's" = "Master"
        "PhD" = "Doctorat"
        "No formal education" = "Fara studii";
    value $income_fmt
        "Low" = "Venit scazut"
        "Medium" = "Venit mediu"
        "High" = "Venit ridicat";
run;
```

Figură 23. Definirea formatelor personalizate pentru variabilele Gender, Extracurricular_Activities, Internet_Access_at_Home, Parent_Education_Level și Family_Income_Level

Date formatate - primele 10 randuri

Obs	Gender	Extracurricular_Activities	Internet_Access_at_Home	Parent_Education_Level	Family_Income_Level
1	Femeie	Da	Nu	Master	Venit mediu
2	Barbat	Nu	Nu	Liceu	Venit scazut
3	Barbat	Da	Nu	Liceu	Venit scazut
4	Femeie	Nu	Da	Liceu	Venit scazut
5	Femeie	Da	Nu	Master	Venit mediu
6	Barbat	Da	Nu	None	Venit mediu
7	Barbat	Da	Nu	Licenta	Venit ridicat
8	Barbat	Da	Da	None	Venit scazut
9	Femeie	Nu	Da	Doctorat	Venit ridicat
10	Femeie	Da	Nu	None	Venit ridicat

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.

```
proc freq data=studenti;
    tables Gender Extracurricular_Activities Internet_Access_at_Home
           Parent_Education_Level Family_Income_Level;
    format Gender $gender_fmt.
           Extracurricular_Activities $yn_fmt.
           Internet_Access_at_Home $yn_fmt.
           Parent_Education_Level $edu_fmt.
           Family_Income_Level $income_fmt.;
    title "Frecvente formatate";
run;
```

Figură 25. Aplicare PROC FREQ

The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Femeie	2449	48.98	2449	48.98
Barbat	2551	51.02	5000	100.00

Extracurricular_Activities	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Nu	2481	49.81	2481	49.81
Da	2500	50.19	4981	100.00
Frequency Missing = 19				

Internet_Access_at_Home	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Nu	2467	49.59	2467	49.59
Da	2508	50.41	4975	100.00
Frequency Missing = 25				

Parent_Education_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Licenta	1011	20.35	1011	20.35
Liceu	938	18.88	1949	39.24
Master	989	19.91	2938	59.15
None	1021	20.56	3959	79.71
Doctorat	1008	20.29	4967	100.00
Frequency Missing = 33				

Family_Income_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Venit ridicat	1630	32.76	1630	32.76
Venit scazut	1678	33.72	3308	66.48
Venit mediu	1668	33.52	4976	100.00
Frequency Missing = 24				

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:

```
data studenti_buni;
    set studenti;
    where Total_Score >= 70;
run;
```

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

Total rows: 3041 Total columns: 23 Rows 1-100

	student_ID	First_Name	Last_Name	Email	Gender	Age	Department	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg	Quizzes_Avg	Participation_Score	Projects_Score	Total_Score	Grade	Study_Hours_per_Week	Extracurricular_Activities
1	S1000	Omar	Williams	student0@university.com	Female	22	Mathematics	97.36	40.61	59.61	73.49	53.17	0.34	62.84	83.49	C	.	Yes
2	S1001	Maria	Brown	student1@university.com	Male	18	Business	97.71	57.27	74	74.23	98.23	.	98.23	92.29	F	.	No
3	S1002	Ahmed	Jones	student2@university.com	Male	24	Engineering	99.52	41.84	63.85	85.85	50	0.47	91.22	93.55	F	.	Yes
4	S1004	John	Smith	student4@university.com	Female	23	CS	59.41	53.13	61.77	67.66	83.98	0.43	87.43	90.91	A	.	Yes
5	S1005	Liam	Brown	student5@university.com	Male	21	Mathematics	40.4	70.21	64.48	87.85	52.24	0.48	67.45	92.66	D	.	Yes
6	S1006	Ahmed	Jones	student6@university.com	Male	24	Business	67.01	90.85	93.74	89	54.17	0.55	87.51	79.78	A	.	Yes
7	S1007	Ahmed	Smith	student7@university.com	Male	19	Engineering	65.91	91.72	59.47	50.72	53.83	0.48	90.04	74	F	.	Yes
8	S1010	John	Smith	student10@university.com	Female	23	Business	86.2	70	85.54	67.58	77.77	0.49	93.14	74.52	D	.	Yes
9	S1011	Omar	Williams	student11@university.com	Female	20	Mathematics	59.3	93.24	87.9	64.57	80.19	0.28	64.17	93.25	B	.	Yes
10	S1012	John	Davis	student12@university.com	Male	21	Engineering	95.48	67.24	97.23	83.43	67.3	0.13	54.02	90.16	D	.	No
11	S1013	John	Williams	student13@university.com	Male	18	Business	91.39	56.63	87.73	91.41	50.94	0.14	51.36	74.8	A	23	No
12	S1014	Liam	Williams	student14@university.com	Male	19	CS	58.16	94.98	64.46	59.6	51.06	0.28	55.5	89.28	C	.	No
13	S1015	Ahmed	Jones	student15@university.com	Male	18	CS	93.54	63.28	89.29	78.15	94.42	.	50.21	88.29	B	.	Yes
14	S1016	Maria	Jones	student16@university.com	Female	21	Business	63.61	70.2	70.52	68.39	55.13	0.21	66.35	78.95	D	.	No
15	S1022	Sara	Davis	student22@university.com	Male	22	Business	99.44	40.24	43.79	91.36	84.72	0.31	62.44	67.98	F	.	No
16	S1025	Emma	Davis	student25@university.com	Male	18	Mathematics	60.65	85.44	75.21	55.85	60.09	0.48	95.32	99.29	B	.	No
17	S1026	Emma	Johnson	student26@university.com	Male	19	Business	53.94	43.04	90.43	77.95	91.05	0.84	94.23	89.74	B	.	No
18	S1027	Sara	Johnson	student27@university.com	Male	18	Engineering	63.85	73.21	51.01	65.02	62.87	0.71	89.32	87.53	F	.	No
19	S1029	Maria	Jones	student29@university.com	Female	23	Mathematics	61.73	48.02	70.38	70.3	99.95	0.49	54.18	82.09	D	.	Yes
20	S1030	Ahmed	Brown	student30@university.com	Female	22	Business	59.64	86.31	44.03	60.61	66.37	.	74.39	74.94	D	.	Yes
21	S1031	Ali	Johnson	student31@university.com	Male	22	Engineering	78.74	53.05	48.48	81.44	71.14	0.41	52.74	81.19	A	.	Yes
22	S1032	Maria	Smith	student32@university.com	Female	21	Business	66.41	53.42	58.54	62.21	99.5	0.91	99.92	71.79	A	.	Yes
23	S1033	Sara	Williams	student33@university.com	Male	19	Engineering	91.02	54.5	42.61	74.7	68.47	0.57	67.92	74.55	F	.	No
24	S1034	Emma	Davis	student34@university.com	Female	21	CS	59.2	75.21	59.56	74.13	68.31	0.41	97.92	99.42	F	.	No
25	S1037	Ali	Williams	student37@university.com	Male	21	Mathematics	74.09	54.51	55.94	60.83	61.49	0.23	55.92	72.42	F	.	Yes
26	S1038	Ali	Williams	student38@university.com	Male	18	Business	52.79	65.86	99.66	95.22	71.16	0.57	61.86	76.22	B	.	Yes
27	S1040	John	Johnson	student40@university.com	Male	24	Business	89.53	93.05	54.35	64.66	52.16	0.41	91.5	90.3	A	.	Yes
28	S1041	Omar	Smith	student41@university.com	Male	23	CS	74.51	87.26	81.21	59.03	68.7	.	98.97	71.1	C	.	Yes
29	S1042	Sara	Davis	student42@university.com	Male	20	CS	54.33	77.55	91.56	62.98	67.31	0.51	54.41	70.46	D	.	No
30	S1043	Liam	Williams	student43@university.com	Male	18	Engineering	79.62	60.28	63.22	66.63	62.24	0.26	57.8	82.94	D	.	Yes
31	S1044	Maria	Smith	student44@university.com	Male	22	Business	87.94	82.04	92.54	77.85	82.71	0.17	52.12	88.66	A	.	No
32	S1045	Maria	Brown	student45@university.com	Male	21	CS	59.54	90.62	60.41	52.32	84.5	0.14	63.02	99.3	A	.	Yes
33	S1049	Emma	Jones	student49@university.com	Male	24	Engineering	90.15	74.3	62.03	55.46	63.23	.	72.22	94.04	B	.	No
34	S1050	Omar	Jones	student50@university.com	Female	19	Engineering	93.97	59.73	80.37	62.83	58.53	0.27	66.29	77.42	B	.	Yes
35	S1051	Emma	Jones	student51@university.com	Female	20	Engineering	96.88	84.75	54.22	91.88	52.83	0.19	80.92	92.62	F	.	No
36	S1052	John	Brown	student52@university.com	Female	19	CS	70.28	52.41	82.06	80.76	80.96	0.94	75.12	99.26	B	16	Yes
37	S1053	Maria	Smith	student53@university.com	Female	23	Business	78.44	84.77	60.85	61.37	83.44	0.94	64.74	77.51	A	.	No
38	S1057	Ali	Johnson	student57@university.com	Male	20	CS	50.68	51.55	54.32	92.38	53.72	0.23	97.86	92.47	B	.	Yes
39	S1059	Ahmed	Johnson	student59@university.com	Male	24	Business	64.05	84.4	83.31	75.07	84.45	0.47	70.92	79.41	F	.	No

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.

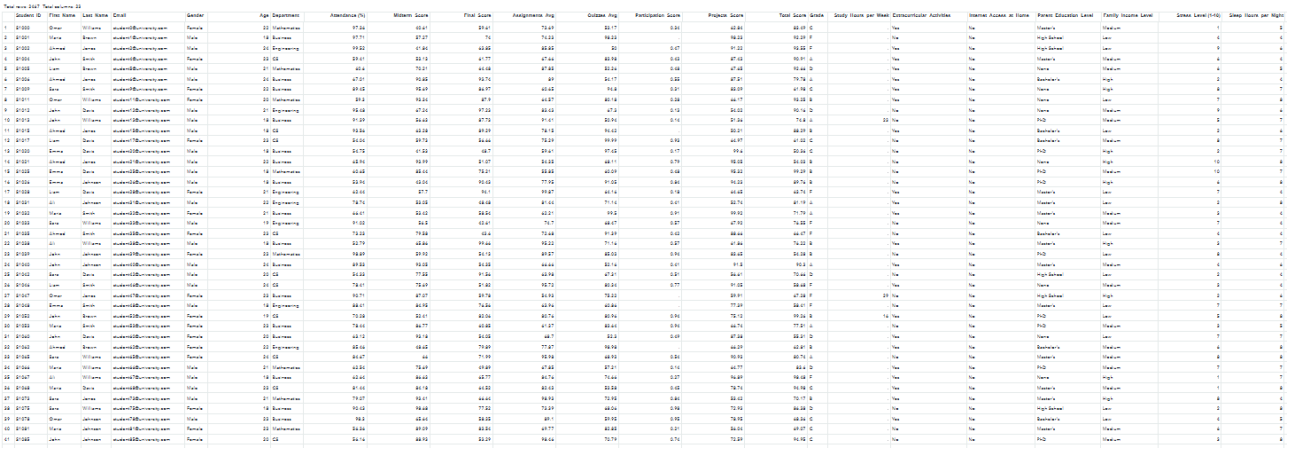
2.3.2: Studenți fără acces la internet acasă

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

Rezolvare:

```
data fara_internet;
    set studenti;
    where Internet_Access_at_Home = "No";
run;
```

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:

```
data risc_educational;
    set studenti;
    where Family_Income_Level = "Low" and Total_Score < 60;
run;
```

Figură 31. Cod SAS pentru selecția cazurilor cu risc educațional

Total rows: 343 Total columns: 23 Rows 1-100

	Student_ID	First_Name	Last_Name	Email	Gender	Age	Department	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg	Quizzes_Avg
8	S1060	John	Davis	student60@university.com	Female	20	Business	63.12	93.18	54.05	68.7	52.3
9	S1064	Omar	Davis	student64@university.com	Male	20	CS	87.08	99.35	49.45	79.7	91.11
10	S1092	Sara	Davis	student92@university.com	Female	24	Business	62.95	55.71	60.57	64	53.42
11	S1114	Omar	Johnson	student114@university.com	Female	18	Business	73.18	96.87	89.32	76.01	78.84
12	S1121	Ahmed	Johnson	student121@university.com	Male	20	Mathematics	56.55	92.18	57.92	73.57	69.29
13	S1159	Maria	Smith	student159@university.com	Male	24	Business	55.17	59.93	92.14	92.04	53.09
14	S1164	Sara	Williams	student164@university.com	Male	22	Engineering	60.23	96.43	76	94.4	66.49
15	S1165	Ali	Johnson	student165@university.com	Male	18	CS	84.14	45.59	62.56	59.85	90.99
16	S1184	Ali	Williams	student184@university.com	Male	22	CS	92.81	70.8	74.5	85.86	85.6
17	S1195	Sara	Smith	student195@university.com	Male	18	Business	72.34	99.68	82.71	80.69	66.6
18	S1241	Omar	Williams	student241@university.com	Female	23	Business	86.14	50.22	69.28	78.06	95.23
19	S1258	Liam	Smith	student258@university.com	Female	24	Business	62.93	47.39	84.25	60.48	65.18
20	S1266	Maria	Davis	student266@university.com	Male	23	Business	53.87	63.38	52.44	79.39	66.72
21	S1268	Omar	Jones	student268@university.com	Female	23	Business	59.47	93.9	51.73	55.18	85.44
22	S1281	Maria	Brown	student281@university.com	Male	20	Business	95.81	66.69	65.86	84.34	90.08
23	S1290	Ali	Johnson	student290@university.com	Male	18	Engineering	63.27	40.73	92.85	81.75	91.33
24	S1313	Omar	Brown	student313@university.com	Male	22	Business	70.45	73.95	43.48	59.96	78.32
25	S1346	John	Brown	student346@university.com	Female	24	Business	77.28	50	98.05	88.04	97.34
26	S1348	Liam	Williams	student348@university.com	Female	22	Mathematics	67.37	60.41	68.65	74.84	66.89
27	S1395	Maria	Jones	student395@university.com	Male	20	CS	51.58	.	76.52	84.96	94.85
28	S1398	Liam	Davis	student398@university.com	Female	23	Mathematics	82.3	54.81	77.49	61.68	65
29	S1401	Maria	Davis	student401@university.com	Male	19	CS	86.46	75.75	78.07	96.93	70.87
30	S1417	Sara	Williams	student417@university.com	Male	24	CS	85.26	82.99	94.74	83.02	59.88
31	S1436	Liam	Brown	student436@university.com	Male	20	Engineering	55.36	78.54	99.85	97.37	96.15

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:

```
data reusite_context_defavorizat;
    set studenti;
    where Parent_Education_Level = "None" and Total_Score >= 85;
run;
```

Figură 33. Cod SAS pentru identificarea reușitelor în contexte defavorizate

Total rows: 300 Total columns: 23 Rows 1-100

	Student_ID	First_Name	Last_Name	Email	Gender	Age	Department	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg
1	S1005	Liam	Brown	student5@university.com	Male	21	Mathematics	60.6	70.21	64.48	87.8
2	S1011	Omar	Williams	student11@university.com	Female	20	Mathematics	59.3	93.34	87.9	64.5
3	S1012	John	Davis	student12@university.com	Male	21	Engineering	95.48	67.24	97.23	83.4
4	S1067	Ali	Williams	student67@university.com	Male	18	Business	62.64	86.63	65.77	84.7
5	S1070	Sara	Jones	student70@university.com	Male	21	CS	93.82	46.89	51.54	82.2
6	S1096	Sara	Jones	student96@university.com	Female	19	Mathematics	58.78	97.23	62.71	69.1
7	S1117	Emma	Brown	student117@university.com	Female	21	Mathematics	75.61	97.23	44.79	81.9
8	S1152	Ali	Davis	student152@university.com	Male	23	Mathematics	75.8	46.3	60.45	75.3
9	S1168	Omar	Brown	student168@university.com	Male	19	Business	64.61	69.88	80.11	82.0
10	S1191	John	Jones	student191@university.com	Male	22	Mathematics	87.33	69.49	50.61	73.0
11	S1199	John	Smith	student199@university.com	Male	20	CS	57.16	42.8	93.41	89.9
12	S1211	Omar	Smith	student211@university.com	Female	21	Engineering	59.17	85.66	58.5	77.9
13	S1216	Maria	Brown	student216@university.com	Female	20	CS	57.54	80.75	87.87	55.4
14	S1242	Maria	Brown	student242@university.com	Male	19	CS	98.2	41.32	53.5	80.9

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.*

```
data studenti_eticheta;
    set studenti;
    length Tip_Student $15;
    if Participation_Score >= 70 and Extracurricular_Activities = "Yes" then
        Tip_Student = "Activ complet";
    else if Participation_Score >= 70 then
        Tip_Student = "Activ la curs";
    else if Extracurricular_Activities = "Yes" then
        Tip_Student = "Activ social";
    else
        Tip_Student = "Pasiv";
run;
```

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

Total rows: 5000 Total columns: 24 Rows 1-100

Total_Score	Grade	Study_Hours_per_Week	Extracurricular_Activities	Internet_Access_at_Home	Parent_Education_Level	Family_Income_Level	Stress_Level (1-10)	Sleep_Hours_per_Night	Tip_Student
83.49	C	.	Yes	No	Master's	Medium	1	5	Activ social
92.29	F	.	No	No	High School	Low	4	4	Pasiv
93.55	F	.	Yes	No	High School	Low	9	6	Activ social
51.03	A	.	No	Yes	High School	Low	8	4	Pasiv
90.91	A	.	Yes	No	Master's	Medium	6	4	Activ social
92.66	D	.	Yes	No	None	Medium	6	5	Activ social
79.78	A	.	Yes	No	Bachelor's	High	2	4	Activ social
74	F	.	Yes	Yes	None	Low	5	7	Activ social
55.55	D	.	No	Yes	PhD	High	9	4	Pasiv
61.98	C	.	Yes	No	None	High	8	7	Activ social
74.52	D	.	Yes	Yes	High School	Low	4	8	Activ social
93.25	B	.	Yes	No	None	Low	7	8	Activ social
90.16	D	.	No	No	None	Medium	9	6	Pasiv
74.8	A	23	No	No	PhD	Medium	5	7	Pasiv
89.28	C	.	No	Yes	Bachelor's	High	2	6	Pasiv
88.29	B	.	Yes	No	Bachelor's	Low	2	6	Activ social
78.95	D	.	No	Yes	High School	High	8	6	Pasiv
61.02	C	.	No	No	Bachelor's	Medium	8	7	Pasiv
67.4	A	.	Yes	Yes	High School	High	8	7	Activ social
58.2	C	.	No	Yes	Master's	Low	5	7	Pasiv

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:

```
data studenti_clasificati;
    set studenti;
    length Scor_Categorie $10;
    if Total_Score >= 90 then Scor_Categorie = "Ridicat";
    else if Total_Score >= 60 then Scor_Categorie = "Mediu";
    else Scor_Categorie = "Scăzut";
run;
```

Figură 37. Cod SAS pentru marcarea eligibilității la bursă

Total rows: 5000 Total columns: 24 — Rows 1-100

Score	Total_Score	Grade	Study_Hours_per_Week	Extracurricular_Activities	Internet_Access_at_Home	Parent_Education_Level	Family_Income_Level	Stress_Level (1-10)	Sleep_Hours_per_Night	Scor_Categorie
62.84	83.49	C	.	Yes	No	Master's	Medium	1	5	Mediu
98.23	92.29	F	.	No	No	High School	Low	4	4	Ridicat
91.22	93.55	F	.	Yes	No	High School	Low	9	6	Ridicat
55.48	51.03	A	.	No	Yes	High School	Low	8	4	Scăzut
87.43	90.91	A	.	Yes	No	Master's	Medium	6	4	Ridicat
67.65	92.66	D	.	Yes	No	None	Medium	6	5	Ridicat
87.51	79.78	A	.	Yes	No	Bachelor's	High	2	4	Mediu
90.04	74	F	.	Yes	Yes	None	Low	5	7	Mediu
57.57	55.55	D	.	No	Yes	PhD	High	9	4	Scăzut
83.09	61.98	C	.	Yes	No	None	High	8	7	Mediu
93.14	74.52	D	.	Yes	Yes	High School	Low	4	8	Mediu
66.17	93.25	B	.	Yes	No	None	Low	7	8	Ridicat
54.02	90.16	D	.	No	No	None	Medium	9	6	Ridicat
51.36	74.8	A	23	No	No	PhD	Medium	5	7	Mediu
55.5	89.28	C	.	No	Yes	Bachelor's	High	2	6	Mediu
50.21	88.29	B	.	Yes	No	Bachelor's	Low	2	6	Mediu
66.35	78.95	D	.	No	Yes	High School	High	8	6	Mediu
64.97	61.02	C	.	No	No	Bachelor's	Medium	8	7	Mediu
55.52	67.4	A	.	Yes	Yes	High School	High	8	7	Mediu
82.56	58.2	C	.	No	Yes	Master's	Low	5	7	Scăzut
99.6	50.36	C	.	No	No	PhD	High	2	7	Scăzut
95.05	54.03	B	.	No	No	None	High	10	8	Scăzut
62.64	87.98	F	.	No	Yes	PhD	Low	8	7	Mediu

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:

```
data scoruri_bune;
    set studenti;
    array note[5] Midterm_Score Final_Score Projects_Score Quizzes_Avg Assignments_Avg;
    Nr_Componente_Bune = 0;

    do i = 1 to 5;
        if note[i] >= 70 then Nr_Componente_Bune + 1;
    end;

    drop i;
run;
```

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

Total rows: 5000 Total columns: 24 — Rows 1-100

Total_Score	Grade	Study_Hours_per_Week	Extracurricular_Activities	Internet_Access_at_Home	Parent_Education_Level	Family_Income_Level	Stress_Level (1-10)	Sleep_Hours_per_Night	Nr_Componente_Bune
83.49	C	.	Yes	No	Master's	Medium	1	5	1
92.29	F	.	No	No	High School	Low	4	4	4
93.55	F	.	Yes	No	High School	Low	9	6	2
51.03	A	.	No	Yes	High School	Low	8	4	0
90.91	A	.	Yes	No	Master's	Medium	6	4	2
92.66	D	.	Yes	No	None	Medium	6	5	2
79.78	A	.	Yes	No	Bachelor's	High	2	4	4
74	F	.	Yes	Yes	None	Low	5	7	2
55.55	D	.	No	Yes	PhD	High	9	4	2
61.98	C	.	Yes	No	None	High	8	7	4
74.52	D	.	Yes	Yes	High School	Low	4	8	4
93.25	B	.	Yes	No	None	Low	7	8	3
90.16	D	.	No	No	None	Medium	9	6	2
74.8	A	23	No	No	PhD	Medium	5	7	2
89.28	C	.	No	Yes	Bachelor's	High	2	6	1
88.29	B	.	Yes	No	Bachelor's	Low	2	6	3
78.95	D	.	No	Yes	High School	High	8	6	2
61.02	C	.	No	No	Bachelor's	Medium	8	7	2
67.4	A	.	Yes	Yes	High School	High	8	7	2
58.2	C	.	No	Yes	Master's	Low	5	7	4
50.36	C	.	No	No	PhD	High	2	7	2

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:

```sas
data eticheta_student;
  set studenti;
  Nume_Complet = catx(" ", upcase(First_Name), upcase(Last_Name));
  Eticheta = catx(" - ", Nume_Complet, Family_Income_Level);
run;
```

Figură 41. Cod SAS pentru eticheta studentului

Eticheta
OMAR WILLIAMS - Medium
MARIA BROWN - Low
AHMED JONES - Low
OMAR WILLIAMS - Low
JOHN SMITH - Medium
LIAM BROWN - Medium
AHMED JONES - High
AHMED SMITH - Low
OMAR SMITH - High
SARA SMITH - High
JOHN SMITH - Low
OMAR WILLIAMS - Low
JOHN DAVIS - Medium
JOHN WILLIAMS - Medium
LIAM WILLIAMS - High
AHMED JONES - Low
MARIA JONES - High
LIAM DAVIS - Medium
LIAM JONES - High
JOHN JOHNSON - Low
EMMA DAVIS - High
AHMED JONES - High
SARA DAVIS - Low
LIAM WILLIAMS - Low

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:

```
data diferente_quiz;
  set studenti;
  Diferenta = abs(Total_Score - Quizzes_Avg);
run;
```

Figură 43. Cod SAS pentru calculul diferenței față de quizuri

Diferenta
30.32
5.94
43.55
15.24
6.93
40.4
25.61
20.17
23.53
32.82
3.25
13.07
22.86
23.86
38.2
6.13
23.82
38.97
9.18
25.53
47.09
14.08
1.26

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.*

```
data medie_generala;
  set studenti;
  Medie_Componente = mean(Midterm_Score, Final_Score, Projects_Score, Quizzes_Avg, Assignments_Avg)
run;
```

Figură 45. Cod SAS pentru calculul mediei generale pe componente

Medie_Componente
57.984
80.392
66.552
55.988
70.794
68.49
83.054
69.156
66.712
84.24
78.806
78.432
73.844
67.614
65.124
75.07
66.118
71.328
66.546
82.82
69.378

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 exercițiie 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.

```
data studenti_2023 studenti_2024;
    set studenti;
    if mod(_N_, 2) = 0 then output studenti_2023;
    else output studenti_2024;
run;
data toti_studentii;
    set studenti_2023 studenti_2024;
run;
```

Figură 47. Cod SAS pentru concatenarea fișierelor de studenți

Total rows: 5000 Total columns: 23 Rows

	Student_ID	First_Name	Last_Name	Email	Gender
1	S1001	Maria	Brown	student1@university.com	Male
2	S1003	Omar	Williams	student3@university.com	Female
3	S1005	Liam	Brown	student5@university.com	Male
4	S1007	Ahmed	Smith	student7@university.com	Male
5	S1009	Sara	Smith	student9@university.com	Female
6	S1011	Omar	Williams	student11@university.com	Female
7	S1013	John	Williams	student13@university.com	Male
8	S1015	Ahmed	Jones	student15@university.com	Male
9	S1017	Liam	Davis	student17@university.com	Female
10	S1019	John	Johnson	student19@university.com	Male
11	S1021	Ahmed	Jones	student21@university.com	Male
12	S1023	Liam	Williams	student23@university.com	Female
13	S1025	Emma	Davis	student25@university.com	Male
14	S1027	Sara	Johnson	student27@university.com	Male
15	S1029	Maria	Jones	student29@university.com	Female

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.

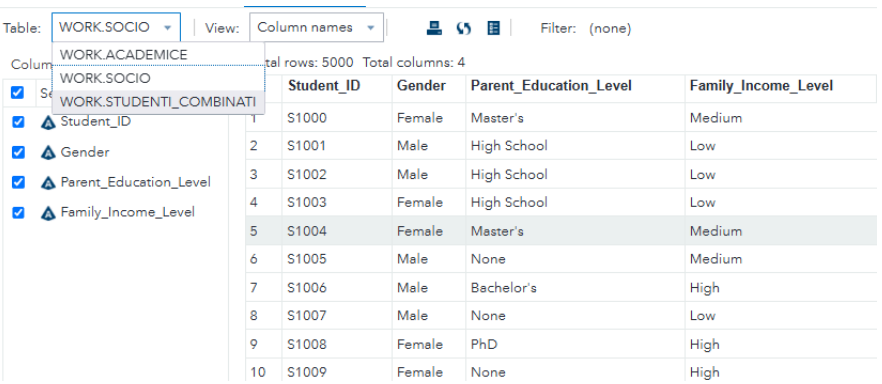
```
data academice;
  set studenti;
  keep Student_ID Midterm_Score Final_Score Total_Score Grade;
run;

data socio;
  set studenti;
  keep Student_ID Gender Family_Income_Level Parent_Education_Level;
run;

proc sort data=academice; by Student_ID; run;
proc sort data=socio; by Student_ID; run;

data studenti_combinati;
  merge academice socio;
  by Student_ID;
run;
```

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



	Student_ID	Gender	Parent_Education_Level	Family_Income_Level
1	S1000	Female	Master's	Medium
2	S1001	Male	High School	Low
3	S1002	Male	High School	Low
4	S1003	Female	High School	Low
5	S1004	Female	Master's	Medium
6	S1005	Male	None	Medium
7	S1006	Male	Bachelor's	High
8	S1007	Male	None	Low
9	S1008	Female	PhD	High
10	S1009	Female	None	High

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.

```
data bursieri;
  set studenti;
  if Total_Score >= 85 then do;
    Tip_Bursa = "Merit";
    output;
  end;
run;

proc sql;
  create table studenti_bursieri as
  select s.Student_ID, s.First_Name, s.Last_Name, s.Total_Score, b.Tip_Bursa
  from studenti as s
  inner join bursieri as b
  on s.Student_ID = b.Student_ID;
quit;
```

Figură 51. Cod SAS pentru alăturare cu PROC SQL JOIN

Table: WORK.BURSIERI | View: Column names | Filter: (none)

Columns

- Select all
- Student_ID
- First_Name
- Last_Name
- Email
- Gender
- Age
- Department
- Attendance (%)
- Midterm_Score
- Final_Score
- Assignments_Avg
- Quizzes_Avg
- Participation_Score
- Projects_Score
- Total_Score

Property	Value
Label	Total_Score
Name	Total_Score
Length	8

Total rows: 1478 Total columns: 24

	Student_ID	First_Name	Last_Name	Email	Tip_Bursa
1	S1001	Maria	Brown	student1@university.com	Merit
2	S1002	Ahmed	Jones	student2@university.com	Merit
3	S1004	John	Smith	student4@university.com	Merit
4	S1005	Liam	Brown	student5@university.com	Merit
5	S1011	Omar	Williams	student11@university.com	Merit
6	S1012	John	Davis	student12@university.com	Merit
7	S1014	Liam	Williams	student14@university.com	Merit
8	S1015	Ahmed	Jones	student15@university.com	Merit
9	S1022	Sara	Davis	student22@university.com	Merit
10	S1025	Emma	Davis	student25@university.com	Merit
11	S1026	Emma	Johnson	student26@university.com	Merit
12	S1027	Sara	Johnson	student27@university.com	Merit
13	S1034	Emma	Davis	student34@university.com	Merit
14	S1040	John	Johnson	student40@university.com	Merit
15	S1044	Maria	Smith	student44@university.com	Merit
16	S1045	Maria	Brown	student45@university.com	Merit
17	S1049	Emma	Jones	student49@university.com	Merit
18	S1051	Emma	Jones	student51@university.com	Merit
19	S1052	John	Brown	student52@university.com	Merit
20	S1057	Ali	Johnson	student57@university.com	Merit

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

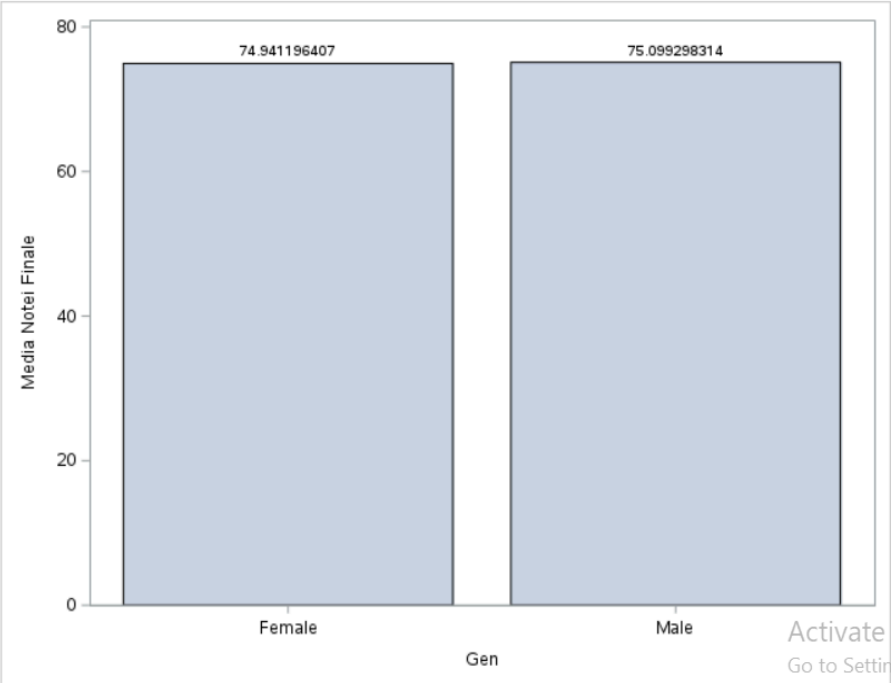
Cerință: *Se dorește compararea performanțelor academice între genuri prin afișarea mediei scorului total pentru fiecare categorie de gen (masculin și feminin).*

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.

```
proc sgplot data=students;
    vbar Gender / response=Total_Score stat=mean datalabel;
    yaxis label="Media Notei Finale";
    xaxis label="Gen";
run;
```

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.

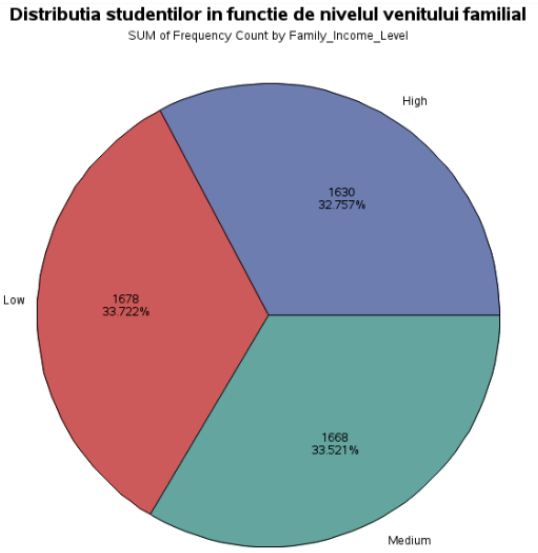
2.7.2. Pie chart – Distribuția studenților pe niveluri de venit

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

Rezolvare:

```
proc freq data=studenti noprint;
  tables Family_Income_Level / out=venituri_freq;
run;
proc gchart data=venituri_freq;
  pie Family_Income_Level / sumvar=Count value=inside percent=inside slice=outside;
  title "Distributia studentilor in functie de nivelul venitului familial";
run;
```

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



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

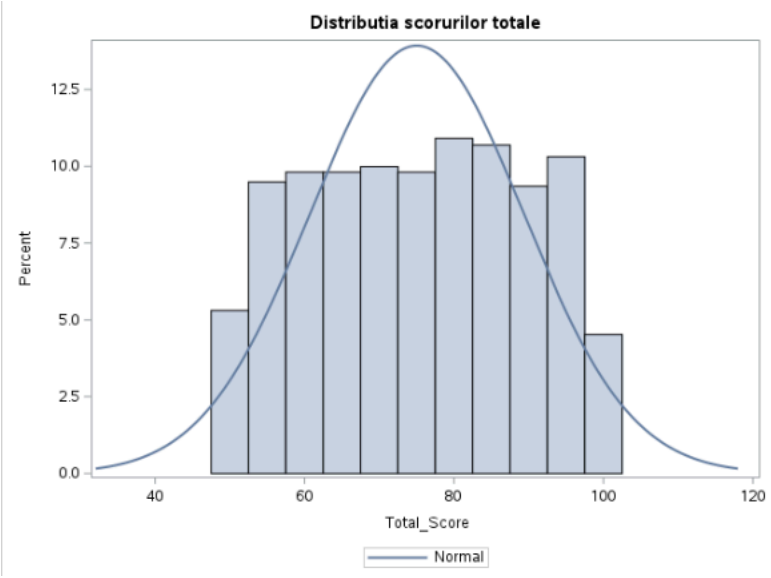
Acest grafic circular reprezintă vizual proporțiile studenților pe categorii de venit. Fiecare felie indică atât procentul, cât și numărul efectiv al studenților care se încadrează într-o anumită clasă socio-economică. Graficul este util pentru a înțelege cât de echilibrată sau dezechilibrată este distribuția 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.*

```
proc sgplot data=studenti;
  histogram Total_Score / binwidth=5;
  density Total_Score;
  title "Distributia scorurilor totale";
run;
```

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).

```
1 proc means data=students mean std min max maxdec=2;
2     var Final_Score Total_Score Quizzes_Avg Assignments_Avg Projects_Score;
3 run;
```

Figură 59. Statistici descriptive pentru variabilele de performanță academică

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum
Final_Score	69.55	17.11	40.01	99.98
Total_Score	75.02	14.32	50.01	99.99
Quizzes_Avg	74.84	14.42	50.00	99.99
Assignments_Avg	74.96	14.40	50.00	99.99
Projects_Score	74.78	14.54	50.00	100.00

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:

```
proc reg data=studenti;
  model Total_Score = Study_Hours_per_Week "Stress_Level (1-10)"n;
  title "Regresie liniara: efectul stresului si a studiului asupra scorului total";
run;
```

Figură 61. Cod SAS - regresia liniară

Regresie liniara: efectul stresului si a studiului asupra scorului total

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Score

Number of Observations Read	5000
Number of Observations Used	4740
Number of Observations with Missing Values	260

Analysis of Variance

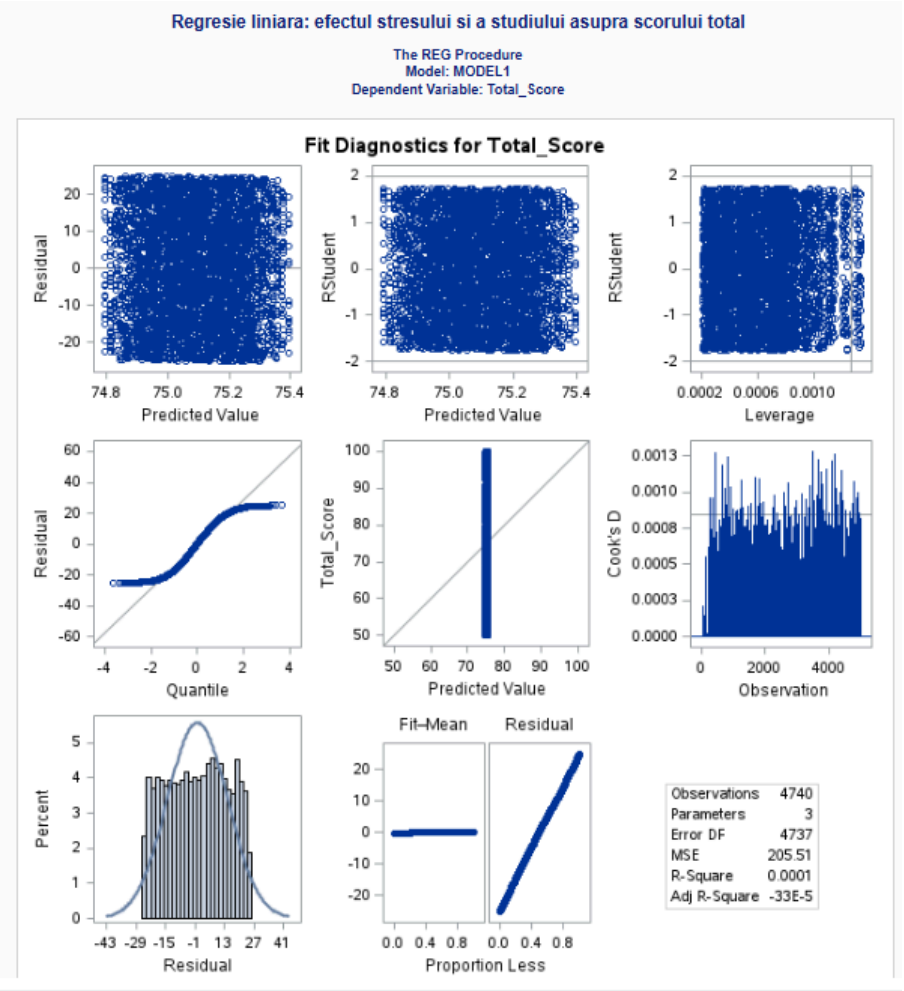
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	90.66142	45.33071	0.22	0.8021
Error	4737	973522	205.51444		
Corrected Total	4739	973613			

Root MSE	14.33577	R-Square	0.0001
Dependent Mean	75.09261	Adj R-Sq	-0.0003
Coeff Var	19.09079		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	75.50266	0.66480	113.57	<.0001
Study_Hours_per_Week	1	-0.01714	0.02895	-0.59	0.5539
Stress_Level (1-10)	1	-0.02136	0.07232	-0.30	0.7677

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 ($p = 0.8021$, $R^2 = 0.0001$). 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

```
proc logistic data=studenti_logistic;
  class Family_Income_Level Internet_Access_at_Home / param=ref;
  model High_Performer(event='1') = Family_Income_Level Internet_Access_at_Home "Stress_Level (1-10)"n;
  title "Regresie logistica: probabilitatea unui scor final ≥ 85";
run;
```

Figură 64. Cod SAS - regresia logistică

Model Information	
Data Set	WORK.STUDENTI_LOGISTIC
Response Variable	High_Performer
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	5000
Number of Observations Used	4935

Response Profile		
Ordered Value	High_Performer	Total Frequency
1	0	3790
2	1	1145

Probability modeled is High_Performer=1.

Note: 65 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Family_Income_Level	High	1	0
	Low	0	1
	Medium	0	0
Internet_Access_at_Home	No	1	
	Yes	0	

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	5348.590	5353.609
SC	5355.094	5386.130
-2 Log L	5346.590	5343.609

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.9808	4	0.5610
Score	2.9725	4	0.5624

Figură 65. Output 1 - regresia logistică

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Family_Income_Level	2	1.6085	0.4474
Internet_Access_at_H	1	0.4520	0.5014
Stress_Level (1-10)	1	0.8313	0.3619

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1761	0.0945	154.9170	<.0001
Family_Income_Level	High	1	0.1000	0.0831	1.4469	0.2290
Family_Income_Level	Low	1	0.0794	0.0828	0.9194	0.3376
Internet_Access_at_H	No	1	-0.0454	0.0675	0.4520	0.5014
Stress_Level (1-10)		1	-0.0107	0.0117	0.8313	0.3619

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Family_Income_Level High vs Medium	1.105	0.939	1.301
Family_Income_Level Low vs Medium	1.083	0.920	1.273
Internet_Access_at_H No vs Yes	0.956	0.837	1.091
Stress_Level (1-10)	0.989	0.967	1.012

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	50.8	Somers' D	0.033
Percent Discordant	47.5	Gamma	0.034
Percent Tied	1.7	Tau-a	0.012
Pairs	4339550	c	0.517

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