

ACADEMIA DE STUDII ECONOMICE BUCUREȘTI
FACULTATEA DE CIBERNETICĂ STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ



PROIECT PACHETE SOFTWARE

Analiza indicelui de depresie în rândul studenților

Profesor coordonator:
Oprea Simona Vasilica

Studenti:
Pichermayer Ruxandra-Theodora
Preda Maria-Andreea
Grupa 1097
Seria E

Cuprins

Introducere	3
Metodologia de lucru a proiectului	5
Python.....	5
Construirea modelului de predicție	5
Preprocesarea datelor	5
Prelucrarea datelor lipsă	6
Tratarea valorilor extreme	8
Standardizarea datelor	8
Matricea de corelație	9
Regresie logistică	11
Raportul de clasificare	12
Matricea de confuzie	13
Interpretare grafică	14
Interpretare tabel	16
Concluzie:	17
Programare SAS.....	17
1. Crearea unui set de date SAS din fișiere externe	17
a) Descrierea problemei	17
b) Informații necesare pentru rezolvare	17
c) Eliminarea coloanelor	18
2. Crearea și folosirea de formate definite de utilizator	19
a) Definirea problemei	19
b) Informații necesare pentru rezolvare	19
c) Metode de calcul, algoritmi, formule de calcul utilizate	19
d) Rezolvare:	20
3. Frecvența depresiei în funcție de gen	21
a) Definirea problemei	21
b) Informații necesare	21
c) Metodă de calcul	21
d) Rezolvare:	21
e) Interpretare:	22
5. Interogare SQL asupra setului de date	22
a) Definirea problemei	22
b) Informații necesare pentru rezolvare	22
c) Metodă de calcul	22
d) Rezolvare:	23
e) Interpretare:	23
5. Statistici descriptive pentru stres și performanță	23
a) Definirea problemei	23
b) Informații necesare	23
c) Metodă de calcul	24
d) Rezolvare	24
e) Interpretare	24

6. Gruparea CGPA în categorii	25
a) Definirea problemei	25
b) Informații necesare	25
c) Rezolvare:	25
d) Interpretare.....	26
7. Corelație între stres și depresie	26
a) Definirea problemei	26
b) Informații necesare	26
c) Rezolvare	27
d) Interpretare.....	27
8. Grafic: depresie în funcție de durata somnului	28
a) Definirea problemei	28
b) Informații necesare	28
c) Rezolvare	28
d) Interpretare.....	28
Concluzie	29

Introducere

Conform celor mai recente cercetări în domeniul psihologiei și al sănătății mintale, depresia este una dintre cele mai răspândite afecțiuni în rândul tinerilor, în special al studenților. Aceasta se manifestă printr-o stare persistentă de tristețe, lipsă de energie, scăderea interesului pentru activitățile zilnice, tulburări ale somnului și ale apetitului, precum și prin dificultăți de concentrare sau senzația de inutilitate. Contextul universitar, marcat de presiunea academică, instabilitate financiară, izolare socială și adaptarea la un mediu nou, contribuie semnificativ la dezvoltarea acestor simptome.

Deși depresia este o problemă frecvent întâlnită, deseori rămâne nediagnosticată sau ignorată, ceea ce poate duce la consecințe grave, atât în plan personal, cât și academic sau social. Identificarea timpurie a factorilor asociați cu riscul de depresie poate reprezenta un pas esențial în prevenirea și gestionarea acestei tulburări în rândul studenților.

În acest context, analiza datelor devine un instrument valoros, oferind posibilitatea de a explora relațiile dintre variabile precum stilul de viață, obiceiurile de somn, mediul de trai sau gradul de stres academic, și riscul de depresie.

Prin urmare, în cadrul acestui proiect ne propunem să realizăm o analiză detaliată a unui set de date reale privind sănătatea mintală a studenților. Vom utiliza limbajul Python și biblioteca Streamlit pentru a dezvolta o aplicație interactivă ce integrează tehnici de preprocesare, analiză exploratorie, vizualizare și modelare predictivă. De asemenea, pentru o analiză în profunzime a acestui subiect, am consolidat studiul lucrării noastre prin folosirea programului SAS. Scopul final este de a evidenția factorii cei mai relevanți care pot contribui la apariția depresiei și de a oferi o perspectivă obiectivă asupra acestei problematice din ce în ce mai actuale.

Descrierea setului de date

Setul de date utilizat în acest proiect, intitulat „Student Depression Dataset”, conține informații colectate de la un eșantion de studenți și are ca scop evidențierea factorilor care pot influența starea lor mintală, în special apariția depresiei. Acest set include un mix de variabile demografice, sociale, academice și comportamentale.

Printre cele mai relevante coloane se regăsesc:

Gender, Age, City, și Degree – caracteristici demografice de bază;

Sleep Duration, Sleep Quality, Stress, Anxiety, Workload, Social Media Use, Academic Performance – variabile ce reflectă obiceiuri de viață, nivelul de stres, și interacțiunea cu factorii de mediu;

Depression – coloana țintă, reprezentând prezența sau absența simptomelor depresive.

Datele sunt etichetate și în format prelucrabil, cu posibile valori lipsă și extreme, ceea ce oferă o oportunitate valoroasă de a aplica tehnici de curățare, codificare și normalizare. Setul este ideal pentru analiză exploratorie și pentru construirea de modele predictive, deoarece integrează o gamă variată de factori psihosociali și stiluri de viață, toate raportate la riscul de depresie.

Prin analizarea acestui set de date, proiectul își propune să aducă în atenție trăsături semnificative care ar putea semnala predispoziția unui student de a dezvolta simptome depresive, facilitând astfel intervenții mai rapide și mai eficiente în context educațional și psihologic.

Diversitatea variabilelor din setul de date ne permite realizarea unei analize exploratorii cuprinzătoare și dezvoltarea unui model statistic prin care am putea prezice diagnosticul. De asemenea, cu ajutorul acestei analize, pot fi evidențiați factori care ar putea fi luați în considerare la elaborarea unor strategii de prevenție sau de intervenție în situații grave.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Id	Gender	Age	City	Professor	Academic	Work Pres	CGPA	Study Sati	Job Satisf	Sleep Dur	Dietary Ha	Degree	Have you	Work/Stuc	Financial	Family His	Depression
2	2	Male	33	Visakhapatnam	Student	5	0	8.97	2	0	5-6 hours	Healthy	B.Pharm	Yes	3	1	No	1
3	8	Female	24	Bangalore	Student	2	0	5.9	5	0	5-6 hours	Moderate	BSc	No	3	2	Yes	0
4	26	Male	31	Srinagar	Student	3	0	7.03	5	0	Less than	Healthy	BA	No	9	1	Yes	0
5	30	Female	28	Varanasi	Student	3	0	5.59	2	0	7-8 hours	Moderate	BCA	Yes	4	5	Yes	1
6	32	Female	25	Jaipur	Student	4	0	8.13	3	0	5-6 hours	Moderate	M.Tech	Yes	1	1	No	0
7	33	Male	29	Pune	Student	2	0	5.7	3	0	Less than	Healthy	PhD	No	4	1	No	0
8	52	Male	30	Thane	Student	3	0	9.54	4	0	7-8 hours	Healthy	BSc	No	1	2	No	0
9	56	Female	30	Chennai	Student	2	0	8.04	4	0	Less than	Unhealthy	Class 12	No	0	1	Yes	0
10	59	Male	28	Nagpur	Student	3	0	9.79	1	0	7-8 hours	Moderate	B.Ed	Yes	12	3	No	1
11	62	Male	31	Nashik	Student	2	0	8.38	3	0	Less than	Moderate	LLB	Yes	2	5	No	1
12	83	Male	24	Nagpur	Student	3	0	6.1	3	0	5-6 hours	Moderate	Class 12	Yes	11	1	Yes	1
13	91	Male	33	Vadodara	Student	3	0	7.03	4	0	Less than	Healthy	BE	Yes	10	2	Yes	0
14	94	Male	27	Kalyan	Student	5	0	7.04	1	0	Less than	Moderate	M.Tech	No	10	1	Yes	1
15	100	Female	19	Rajkot	Student	2	0	8.52	4	0	Less than	Unhealthy	Class 12	No	6	2	Yes	0
16	103	Female	19	Kalyan	Student	5	0	5.64	5	0	Less than	Moderate	Class 12	Yes	4	5	Yes	1
17	106	Male	29	Srinagar	Student	3	0	8.58	3	0	More than	Moderate	M.Tech	Yes	10	2	Yes	1
18	120	Male	25	Nashik	Student	5	0	6.51	2	0	Less than	Unhealthy	M.Ed	Yes	2	5	Yes	1
19	132	Female	20	Ahmedabad	Student	5	0	7.25	3	0	5-6 hours	Healthy	Class 12	Yes	10	3	No	1
20	139	Male	19	Chennai	Student	2	0	7.83	2	0	7-8 hours	Unhealthy	Class 12	No	6	3	No	0
21	145	Male	25	Kalyan	Student	3	0	9.93	3	0	5-6 hours	Moderate	B.Ed	No	8	3	Yes	1
22	161	Male	29	Kolkata	Student	3	0	8.74	4	0	5-6 hours	Moderate	B.Ed	Yes	1	1	No	0
23	162	Male	29	Kolkata	Student	3	0	6.73	3	0	7-8 hours	Moderate	M.Tech	No	0	1	No	0
24	166	Female	25	Ahmedabad	Student	3	0	5.57	3	0	More than	Unhealthy	MSc	Yes	10	5	No	1
25	172	Male	23	Thane	Student	1	0	8.59	4	0	7-8 hours	Healthy	BHM	No	11	3	No	0
26	173	Male	18	Bangalore	Student	4	0	7.1	3	0	More than	Unhealthy	Class 12	Yes	11	5	Yes	1
27	176	Female	20	Mumbai	Student	5	0	8.58	5	0	7-8 hours	Moderate	Class 12	No	2	2	Yes	1
28	186	Male	31	Ahmedabad	Student	2	0	6.08	5	0	7-8 hours	Moderate	LLB	Yes	3	3	Yes	1
29	193	Male	25	Lucknow	Student	3	0	7.25	3	0	More than	Unhealthy	M.Ed	Yes	10	5	No	1
30	208	Male	33	Indore	Student	5	0	5.74	2	0	Less than	Moderate	M.Pharm	No	8	3	Yes	0

Metodologia de lucru a proiectului

Analiza setului de date cu privire la depresia în rândul studenților este structurată în două părți, fiecare dintre ele fiind divizată în cerințe specifice, corespunzătoare celor două pachete software pe care le-am utilizat: **Python** și **SAS**. Fiecare dintre acestea urmărește utilizarea unui set minim de funcționalități – conform cerințelor, ajutându-ne să atingem obiectivul general al proiectului: identificarea factorilor semnificativi asociați cu diagnosticul de depresie și dezvoltarea unor modele de predicție cu care putem estima probabilitatea de apariție a bolii.

Python

Construirea modelului de predicție

Preprocesarea datelor

În prima parte a proiectului, vom analiza setul de date cu ajutorul limbajului Python. Prelucrările pot fi vizualizate în cadrul unei interfețe interactive pe care am implementat-o cu biblioteca Streamlit.

În urma încărcării fișierului CSV, am constatat că trei dintre coloanele importate nu aduc valoare analitică relevantă pentru obiectivele studiului nostru. Prin urmare, am decis să le eliminăm: Profession, Work_Pressure și Job_Satisfaction.

```
df.drop(columns=['Profession','Work_Pressure','Job_Satisfaction'], inplace=True)
```

Vom începe analiza datelor cu un prim pas destul de important, și anume preprocesarea datelor. Cum se poate observa mai sus, setul de date conține coloane precum Gender, Sleep Quality, Stress, Anxiety ce conțin valori pe care calculatorul nu le poate interpreta cu ușurință.

```
24 # Data set preprocess
25 df['Gender'] = df['Gender'].map({'Male': 0, 'Female': 1})
26 df['Suicidal_Thoughts'] = df['Have you ever had suicidal thoughts?'].map({'No': 0, 'Yes': 1})
27 df.drop(columns=['Have you ever had suicidal thoughts?'], inplace=True)
28
29 # Ordinal Encoding
30 diet_map = {'Unhealthy': 0, 'Moderate': 1, 'Healthy': 2}
31 df['Dietary Habits'] = df['Dietary Habits'].map(diet_map)
32
33 sleep_map = {
34     'Less than 5 hours': 0,
35     '5-6 hours': 1,
36     '7-8 hours': 2,
37     'More than 8 hours': 3
38 }
39 df['Sleep Duration'] = df['Sleep Duration'].map(sleep_map)
40
41 non_numeric_cols = df.select_dtypes(include=['object']).columns.to_list()
42 if non_numeric_cols:
43     labelEncoder = LabelEncoder()
44     for col in non_numeric_cols:
45         df[col] = labelEncoder.fit_transform(df[col].astype(str))
```

Coloana *Gender* este transformată într-o variabilă binară (0 = masculin, 1 = feminin).

Coloana *Have you ever had suicidal thoughts?* este redenumită *Suicidal_Thoughts* și mapată la 0 pentru „No” și 1 pentru „Yes”.

După mapare, coloana originală este eliminată din setul de date pentru a evita redundanța.

Coloana *Dietary Habits* este codificată în mod ordonat, reflectând gradul de sănătate al alimentației: 0: Unhealthy, 1: Moderate, 2: Healthy.

Se identifică toate coloanele care au tipul de date object (de obicei stringuri).

Fiecare coloană este transformată numeric folosind LabelEncoder din scikit-learn. Această tehnică atribuie fiecărei etichete un număr întreg unic.

După acest preprocessing:

- Toate coloanele din setul de date au valori numerice.
- Setul este pregătit pentru etapele ulterioare: împărțirea în seturi de antrenare/testare, scalare, antrenarea modelelor etc.

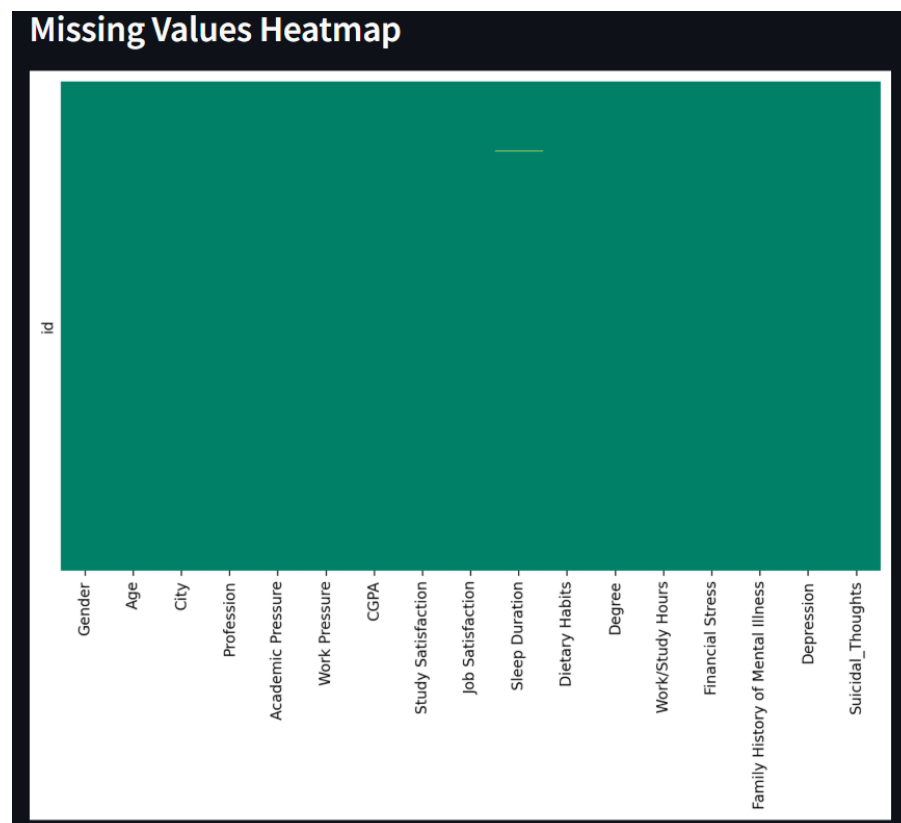
Setul de date după procesare:

id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfai
2	0	33	51	11	5	0	8.97	
8	1	24	3	11	2	0	5.9	
26	0	31	44	11	3	0	7.03	
30	1	28	49	11	3	0	5.59	
32	1	25	16	11	4	0	8.13	
33	0	29	39	11	2	0	5.7	
52	0	30	46	11	3	0	9.54	
56	1	30	6	11	2	0	8.04	
59	0	28	33	11	3	0	9.79	
62	0	31	37	11	2	0	8.38	

Prelucrarea datelor lipsă

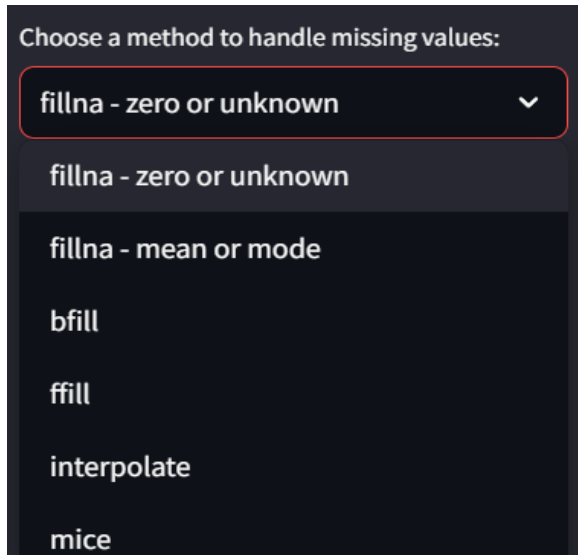
De asemenea, pentru o analiză atentă și corectă a datelor am calculat și procentul valorilor lipsă pe care l-am afișat cu ajutorul unui HeatMap.

```
80 # Calculate the percentage of missing values for each column
81 missing_percentage = df.isnull().mean() * 100
82 missing_percentage = missing_percentage[missing_percentage > 0]
83
84 if not missing_percentage.empty:
85     missing_df = pd.DataFrame(missing_percentage, columns=["Percentage of Missing Values"])
86     st.table(missing_df)
87 else:
88     st.write("No missing values in the dataset.")
89
90 st.subheader("Missing Values Heatmap")
91 plt.figure(figsize=(10, 6))
92 sns.heatmap(df.isnull(), cbar=False, cmap = 'summer', yticklabels=False)
93 st.pyplot(plt)
```



După cum se poate observa, datele puse la dispoziție sunt bine alese, iar valorile lipsa nu reprezintă un impediment major.

Această secțiune din aplicație aplică o metodă de tratare a valorilor lipsă (NaN) în funcție de opțiunea selectată de utilizator (`missing_value_method`). Diferite metode sunt disponibile pentru a gestiona lipsa valorilor în mod adaptat la contextul analizei.



Choose a method to handle missing values:

fillna - zero or unknown

fillna - zero or unknown

fillna - mean or mode

bfill

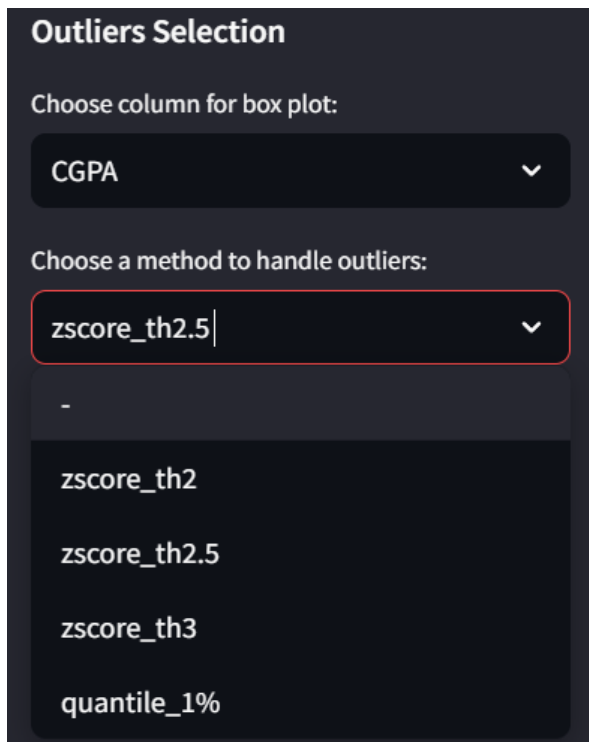
ffill

interpolate

mice

Tratarea valorilor extreme

Dupa pregatirea coloanelor cu valori lipsa, vom trata valorile extreme. Precum mai sus, utilizatorul poate alege ce metoda sa utilizeze, precum IQR sau prin Z-score.



Outliers Selection

Choose column for box plot:

CGPA

Choose a method to handle outliers:

zscore_th2.5

-

zscore_th2

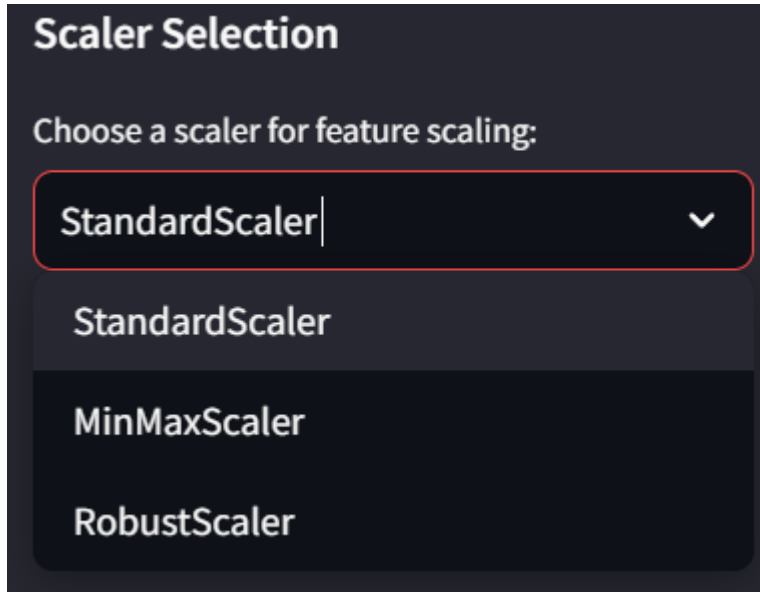
zscore_th2.5

zscore_th3

quantile_1%

Standardizarea datelor

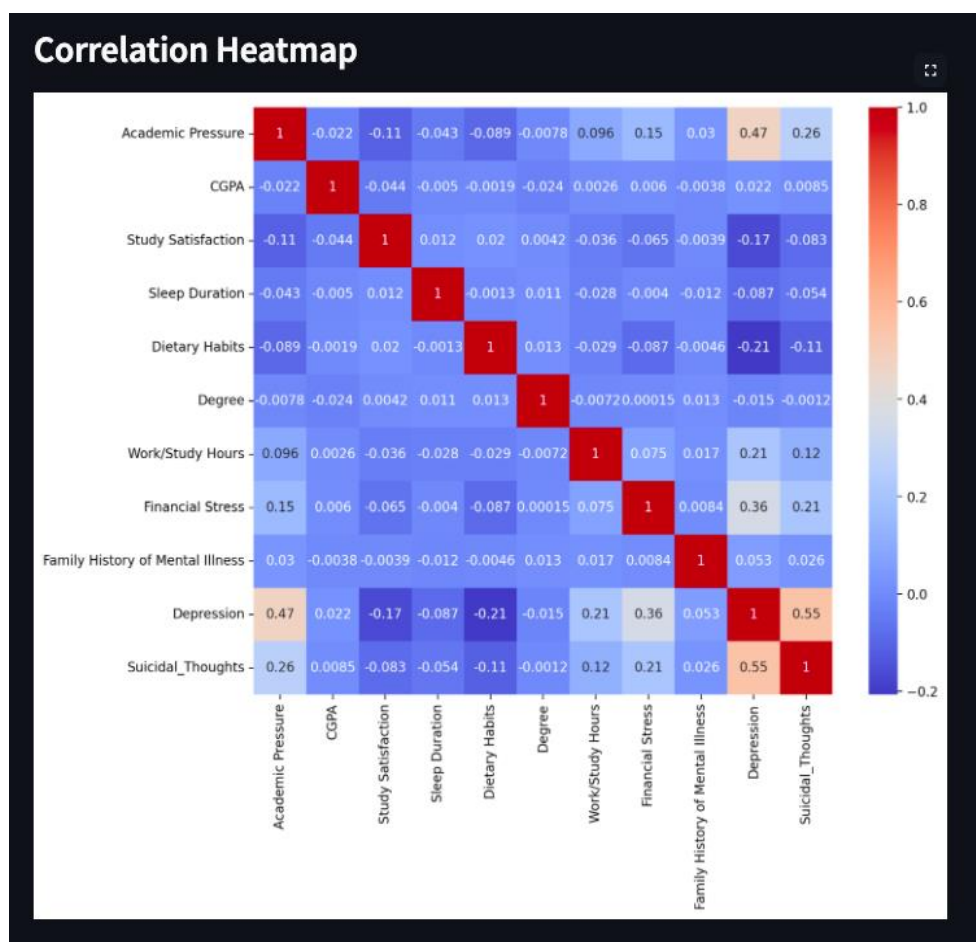
Ultimul pas de preprocesare va fi standardizarea datelor. Pentru acest pas vom folosi un scaler ales de utilizator prin interfata creata pe coloanele de interes, anume cele numerice.



```
71 # sidebar for choosing a scaler
72 st.sidebar.subheader("Scaler Selection")
73 scaler_option = st.sidebar.selectbox(
74     "Choose a scaler for feature scaling:",
75     ("StandardScaler", "MinMaxScaler", "RobustScaler")
76 )
77
```

Matricea de corelatie

Urmatorul pas este indentificarea corelatiilor dintre variabile prin **matricea de corelatie**. Aceasta este un instrument esențial pentru a înțelege relațiile dintre variabilele numerice dintr-un set de date.



Observand aceasta matrice, putem trage concluziile:

1. Ce factori sunt cele mai puternic corelați pozitiv cu depresia?

- **Gândurile suicidale** (**Suicidal Thoughts** → **Depression: 0.55**)
Cel mai puternic asociază faptul că au existat gânduri suicidale cu scorul de depresie, ceea ce era de așteptat din perspectivă clinică.
- **Presiunea academică** (**Academic Pressure** → **Depression: 0.47**)
Studentii care raportează un nivel mai mare de stres academic au, în medie, scoruri de depresie semnificativ mai mari.
- **Stresul financiar** (**Financial Stress** → **Depression: 0.36**)
Grijile legate de bani se asociază de asemenea cu un risc mai mare de simptome depressive.
- **Orele petrecute la curs/studii** (**Work/Study Hours** → **Depression: 0.21**)
Un volum mai mare de lucru/învățare se corelează cu o ușoară creștere a depresiei.

2. Ce factori sunt moderant corelați negativ (posibil protectori)?

- **Satisfacția față de studiu** (Study Satisfaction → Depression: **-0.17**)
Studentii mulțumiți de experiența lor academică tind să aibă scoruri mai mici de depresie.
- **Obiceiurile alimentare sănătoase** (Dietary Habits → Depression: **-0.11**)
O dietă mai sănătoasă se asociază cu niveluri ușor reduse de depresie.
- **Durata somnului** (Sleep Duration → Depression: **-0.09**)
Mai multe ore de somn par să scadă în mică măsură riscul de depresie.

3. Factorii neutri sau aproape de zero

- **CGPA, Family History of Mental Illness** au corelații cu Depression foarte aproape de 0, ceea ce sugerează că în acest set de date nu sunt indicatori direcți ai simptomelor depresive.

4. Alte relații interesante între variabile

- **Academic Pressure și Suicidal Thoughts** (+0.26)
Studentii foarte stresați academic raportează mai des gânduri suicidale.

Regresie logistică

După ce am explorat relațiile dintre variabile cu ajutorul matricii de corelație – identificând factori precum presiunea academică, stresul financiar și gândurile suicidale drept cei mai puternic asociați cu scorul de depresie – am trecut la un model de **regresie logistică**. Scopul acestei analize de regresie a fost de a cuantifica impactul fiecărui predictor asupra probabilității de a dezvolta simptome depresive și de a verifica semnificația statistică a coeficienților obținuți. Astfel, am transformat simpla asociere măsurată prin corelații într-un instrument predictiv, capabil să estimeze riscul individual de depresie și să ne ghideze spre factorii cei mai relevanți pentru intervenții viitoare.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
y_prob = model.predict_proba(X_test)[:, 1]

st.write("### Confusion Matrix")
cm = confusion_matrix(y_test, y_pred)
st.write(pd.DataFrame(cm, columns=["Predicted 0", "Predicted 1"], index=["Actual 0", "Actual 1"]))

st.write("### Classification Report")
report_dict = classification_report(y_test, y_pred, output_dict=True)
report_df = pd.DataFrame(report_dict).transpose()
st.dataframe(report_df.style
              .background_gradient(cmap='Greens')
              .format("{:.2f}")
              .set_properties(**{'text-align': 'center'})
              .set_table_styles(
                  [{ 'selector': 'th', 'props': [('text-align', 'center')]}]
              ))
```

Atât matricea de confuzie, cât și raportul de clasificare sunt instrumente esențiale pentru a evalua performanța unui model de clasificare binară (depresie vs. non-depresie).

Raportul de clasificare

Raportul de clasificare (sklearn's `classification_report`) sintetizează pentru fiecare clasă (0 și 1) următoarele:

Metrică	Definiție
Precision (Precizie)	$TP / (TP + FP)$ – proporția predicțiilor pozitive care au fost corecte
Recall (Sensitivitate)	$TP / (TP + FN)$ – proporția cazurilor pozitive (depresivi) pe care modelul le-a detectat
F1-score	Media armonică între precision și recall: $2 \cdot (\text{precision} \cdot \text{recall}) / (\text{precision} + \text{recall})$
Support	Numărul de exemple reale din fiecare clasă în setul de test

Matricea de confuzie

Confusion Matrix				
	Predicted 0	Predicted 1		
Actual 0	1,834	509		
Actual 1	402	2,836		

Classification Report				
	precision	recall	f1-score	support
0	0.82	0.78	0.80	2343.00
1	0.85	0.88	0.86	3238.00
accuracy	0.84	0.84	0.84	0.84
macro avg	0.83	0.83	0.83	5581.00
weighted avg	0.84	0.84	0.84	5581.00

Interpretare matrice de confuzie:

- **1834** studenți fără depresie au fost identificați corect.
- **2836** studenți cu depresie au fost identificați corect.

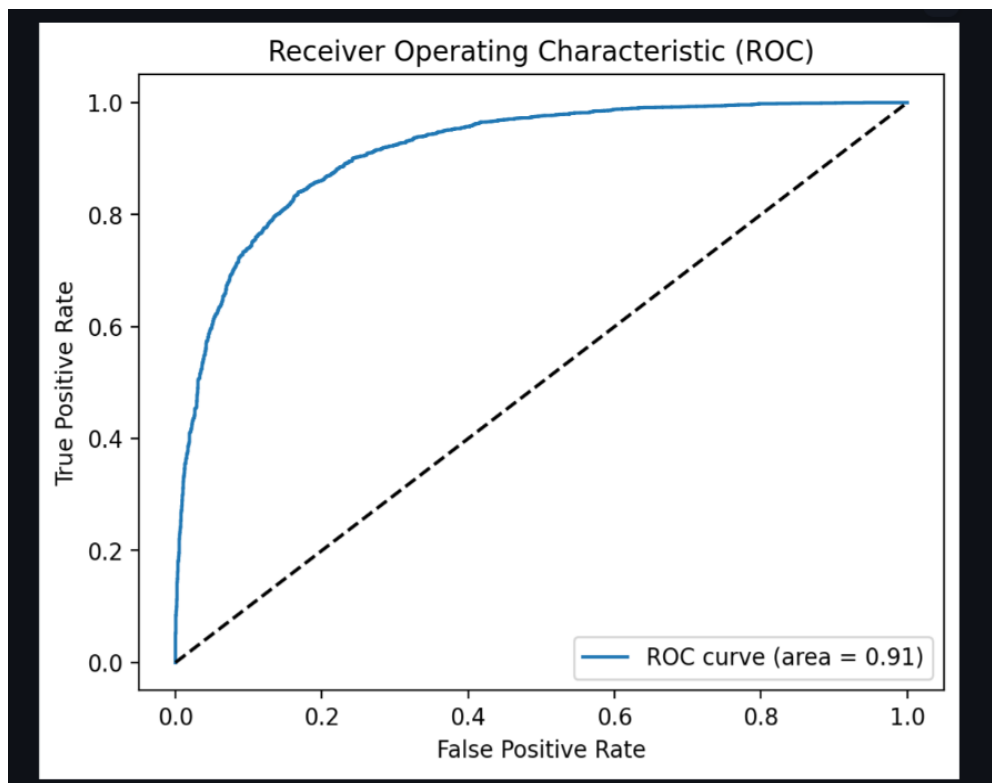
- **509** non-depresivi au fost etichetați greșit ca depresivi → *False Alarm*.
- **402** depresivi au fost ignorați → *Cazuri ratate*, lucru important în context medical.
- **Precision 0.82 (clasa 0)**: 82% din cei prezis ca *non-depresivi* au fost corecți.
- **Recall 0.78 (clasa 0)**: doar 78% din toți non-depresivii reali au fost recunoscuți → modelul mai confundă unii non-depresivi cu depresivi

Interpretare raport de clasificare:

- **Precision 0.85 (clasa 1)**: 85% din cei marcați ca *depresivi* chiar au depresie.
- **Recall 0.88 (clasa 1)**: 88% dintre toți depresivii reali au fost corect detectați.
- **F1-score 0.86 (clasa 1)**: echilibru bun între precizie și sensibilitate.
- **Accuracy totală: 84%** – proporția totală a predicțiilor corecte.
- **Macro avg** (media simplă): 0.83 → echilibrată între clase.
- **Weighted avg** (media ponderată în funcție de mărimea claselor): 0.84 → confirmă balanța bună.

Concluzie după analiza matricii de confuzie

Modelul nostru de clasificare reușește să prezică corect cazurile de depresie în 84% din situații, cu un **recall de 88% pentru clasa „depresie”**, ceea ce este esențial în contextul prevenirii și intervenției timpurii. Matricea de confuzie ne arată că doar o mică parte din cazurile depresive sunt omise (FN = 402), iar raportul de clasificare evidențiază un echilibru sănătos între precizie și sensibilitate, în special pentru clasa de interes. Aceste rezultate ne indică faptul că modelul este potrivit pentru o primă evaluare automată a riscului de depresie în rândul studenților.



Curba ROC (Receiver Operating Characteristic) – un instrument excelent pentru a evalua performanța unui model de clasificare binară.

Interpretare grafică

- Linia **punctată diagonală** (linia de la (0,0) la (1,1)) este modelul aleator (fără nicio putere de clasificare).
- Curba albastră este modelul tău – cu cât este mai aproape de colțul **(0,1)**, cu atât modelul este mai performant.
- **AUC (Area Under the Curve) = 0.91**

Curba ROC confirmă robustețea modelului nostru, cu un **AUC de 0.91**, indicând o capacitate foarte bună de a distinge între cazurile de depresie și non-depresie. Acest scor evidențiază faptul că modelul face față bine compromisului între sensibilitate și specificitate, oferind predicții fiabile chiar și în prezența dezechilibrelor de clasă.

Feature Importance (Coefficients)

	Feature	Coefficient
15	Suicidal_Thoughts	1.2298
4	Academic Pressure	1.1703
13	Financial Stress	0.8078
1	Age	-0.5325
12	Work/Study Hours	0.4338
10	Dietary Habits	-0.4305
7	Study Satisfaction	-0.3359
9	Sleep Duration	-0.2149
14	Family History of Mental Illness	0.1428
6	CGPA	0.0928

Interpretare tabel

Caracteristică	Coeficient	Interpretare
Suicidal_Thoughts	+1.2298	Cel mai puternic predictor al depresiei. Cu cât o persoană are mai multe gânduri suicidare, cu atât probabilitatea de a fi depresiv crește considerabil.
Academic Pressure	+1.1703	Presiunea academică ridicată este strâns legată de depresie.
Financial Stress	+0.8078	Problemele financiare cresc riscul de depresie.
Age	-0.5325	Cu cât vârsta este mai mare, probabilitatea depresiei scade.
Work/Study Hours	+0.4338	Orele îndelungate de muncă/studiu contribuie la apariția depresiei.
Dietary Habits	-0.4305	Alimentația sănătoasă reduce probabilitatea depresiei.
Study Satisfaction	-0.3359	Studentii mulțumiți de procesul educațional sunt mai puțin predispuși la depresie.
Sleep Duration	-0.2149	Somnul adecvat are un rol protector împotriva depresiei.
Family History	+0.1428	Istoricul familial este un factor de risc, dar cu influență moderată.
CGPA	+0.0928	Coeficient foarte mic, deci performanța academică are un impact slab.

Coeficienții regresiei logistice ne-au oferit o perspectivă clară asupra factorilor care influențează probabilitatea de apariție a depresiei. Am observat că cele mai importante caracteristici predictive sunt gândurile suicidare, presiunea academică și stresul financiar. În schimb, factori precum durata somnului, satisfacția în studii și obiceiurile alimentare sănătoase acționează ca factori protectivi. Astfel, modelul nu doar că face predicții precise, ci oferă și insighturi valoroase asupra fenomenului psihologic investigat.

Concluzie:

Modelul construit nu doar că face predicții precise, ci servește și ca un instrument de suport pentru universități, consilieri școlari și factori decizionali, oferind o înțelegere profundă asupra factorilor psihologici și comportamentali care influențează sănătatea mintală a studenților. În final, acest demers contribuie la prevenirea și identificarea timpurie a depresiei, într-un context academic tot mai solicitant.

Analiza corelațiilor a oferit o primă privire de ansamblu asupra relațiilor dintre variabile, indicând asocieri semnificative între depresie și factori precum presiunea academică, stresul financiar și gândurile suicidare. Aceste corelații ne-au ajutat să selectăm caracteristici relevante pentru antrenarea modelului.

Regresia logistică s-a dovedit a fi un model eficient, cu o **acuratețe de 84%** și un **AUC de 0.91**, ceea ce arată o capacitate bună de a distinge între cazurile depresive și non-depresive. De asemenea, **matricea de confuzie și raportul de clasificare** au indicat un echilibru bun între precizie și recall pentru ambele clase, dovedind robustețea modelului.

Prin analiza coeficienților, am putut extrage insight-uri valoroase: **gândurile suicidare, presiunea academică și stresul financiar** sunt cei mai importanți predictorii ai depresiei, în timp ce **obiceiurile alimentare sănătoase, satisfacția în studii și un somn adecvat** reduc probabilitatea apariției acesteia.

Programare SAS

Mai departe, am utilizat limbajul de programare SAS pentru a continua analiza datelor colectate.

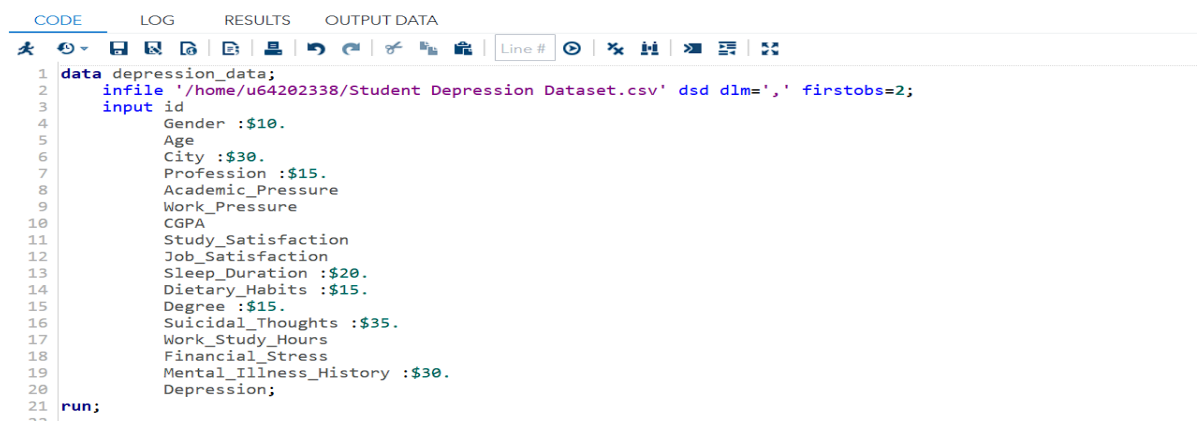
1. Crearea unui set de date SAS din fișiere externe

a) Descrierea problemei

Să se importe fișierele externe folosite în crearea proiectului și să se elimine coloanele nefolositoare pentru analiză

b) Informații necesare pentru rezolvare

- Înainte de a scrie codul, avem nevoie de importarea fișierelor în SAS Studio.
- După importare, avem nevoie de path-ul fiecărui fișier pentru a conduce direct la sursă.



```
1 data depression_data;
2   infile '/home/u64202338/Student Depression Dataset.csv' dsd dlm=',' firstobs=2;
3   input id
4         Gender :$10.
5         Age
6         City :$30.
7         Profession :$15.
8         Academic_Pressure
9         Work_Pressure
10        CGPA
11        Study_Satisfaction
12        Job_Satisfaction
13        Sleep_Duration :$20.
14        Dietary_Habits :$15.
15        Degree :$15.
16        Suicidal_Thoughts :$35.
17        Work_Study_Hours
18        Financial_Stress
19        Mental_Illness_History :$30.
20        Depression;
21 run;
```

În urma execuției codului, avem următorul output:

CODE	LOG	RESULTS	OUTPUT DATA
Table:	WORK.DEPRESSION_DATA	View:	Column names
Columns	Total rows: 27901 Total columns: 15		
<input checked="" type="checkbox"/> Select all			
<input checked="" type="checkbox"/> ID			
<input checked="" type="checkbox"/> Gender			
<input checked="" type="checkbox"/> Age			
<input checked="" type="checkbox"/> City			
<input checked="" type="checkbox"/> Academic_Pressure			
<input checked="" type="checkbox"/> CGPA			
<input checked="" type="checkbox"/> Study_Satisfaction			
<input checked="" type="checkbox"/> Sleep_Duration			
<input checked="" type="checkbox"/> Dietary_Habits			
<input checked="" type="checkbox"/> Degree			
<input checked="" type="checkbox"/> Suicidal_Thoughts			
<input checked="" type="checkbox"/> Work_Study_Hours			
<input checked="" type="checkbox"/> Financial_Stress			
<input checked="" type="checkbox"/> Mental_Illness_History			
<input checked="" type="checkbox"/> Depression			
Property	Value		
Label			
Name			
Length			
▼			

c) Eliminarea coloanelor

În urma încărcării fișierului CSV, am constatat că trei dintre coloanele importate nu aduc valoare analitică relevantă pentru obiectivele studiului nostru. Prin urmare, am decis să le eliminăm: Profession, Work Pressure și Job Satisfaction.

```
23 data depression_data;  
24     set depression_data;  
25     drop Profession Work_Pressure Job_Satisfaction;  
26 run;
```

În urma execuției codului, avem următorul output:

CODE
LOG
RESULTS
OUTPUT DATA

Table: WORK.DEPRESSION_DATA
View: Column names
Filter: (none)

Columns
Select all
id
Gender
Age
City
Academic_Pressure
CGPA
Study_Satisfaction
Sleep_Duration
Dietary_Habits
Degree
Suicidal_Thoughts
Work_Study_Hours
Financial_Stress
Mental_Illness_History
Depression

Total rows: 27901
Total columns: 15

		id	Gender	Age	City	Academic_Pressure
1		2	Male	33	Visakhapatnam	5
2		8	Female	24	Bangalore	2
3		26	Male	31	Srinagar	3
4		30	Female	28	Varanasi	3
5		32	Female	25	Jaipur	4
6		33	Male	29	Pune	2
7		52	Male	30	Thane	3
8		56	Female	30	Chennai	2
9		59	Male	28	Nagpur	3
10		62	Male	31	Nashik	2
11		83	Male	24	Nagpur	3
12		91	Male	33	Vadodara	3
13		94	Male	27	Kalyan	5
14		100	Female	19	Rajkot	2
15		103	Female	19	Kalyan	5
16		106	Male	29	Srinagar	3
17		120	Male	25	Nashik	5
18		132	Female	20	Ahmedabad	5
19		139	Male	19	Chennai	2
20		145	Male	25	Kalyan	3
21		161	Male	29	Kolkata	3

Property
Value

Label
Name
Length

Messages: 22
User: u64202338

2. Crearea și folosirea de formate definite de utilizator

a) Definirea problemei

Datele inițiale conțin mai multe variabile de tip text (character) care exprimă informații relevante pentru analiză (de exemplu: Gender, Suicidal_Thoughts, Dietary_Habits, Sleep_Duration, Degree, Mental_Illness_History).

Pentru a putea folosi aceste variabile în analize statistice și în modele predictive, este necesară **transformarea acestora în valori numerice codificate**.

b) Informații necesare pentru rezolvare

Este necesară o înțelegere clară a valorilor din fiecare variabilă textuală și a logicii de transformare. De exemplu:

- Gender: „Male” → 0, „Female” → 1
- Dietary_Habits: „Unhealthy” → 0, „Moderate” → 1, „Healthy” → 2
- Sleep_Duration: codificare ordinală în funcție de numărul de ore

c) Metode de calcul, algoritmi, formule de calcul utilizate

Pentru codificarea variabilelor calitative de tip text (character) în valori numerice am utilizat procedura `PROC FORMAT`, care permite definirea de formate personalizate de utilizator în SAS.

Prin `PROC FORMAT` am asociat fiecare valoare text cu un cod numeric reprezentat sub formă de text (ex: 'Male' = '0', 'Female' = '1'). Aceasta este o metodă eficientă pentru codificare ordonată, în special în cazul variabilelor categorice sau ordinale.

Ulterior, am aplicat aceste formate folosind funcția `PUT()` pentru a obține codul numeric sub formă de text și `INPUT()` pentru conversia rezultată într-o variabilă numerică utilizabilă în analize statistice sau modele predictive.

Formula generală utilizată:

Variabilă_numerica = INPUT(PUT(Variabilă_text, \$FORMAT.), 8.);

d) Rezolvare:

După rularea blocului de cod, variabilele care anterior conțineau text sunt acum transformate în valori numerice.

```
31 /* 2. Definirea formatelor */
32 proc format;
33   value $gender_fmt 'Male' = '0' 'Female' = '1';
34   value $thoughts_fmt 'No' = '0' 'Yes' = '1';
35   value $diet_fmt
36     'Unhealthy' = '0'
37     'Moderate' = '1'
38     'Healthy' = '2';
39   value $sleep_fmt
40     'Less than 5 hours' = '0'
41     '5-6 hours' = '1'
42     '7-8 hours' = '2'
43     'More than 8 hours' = '3';
44
45   value $mental_fmt 'Yes' = '1' 'No' = '0';
46 run;
47
48 /* Aplicarea formatelor și creare coloane noi */
49 data depression_data_formatted;
50   set depression_data;
51
52   Gender_num = input(put(Gender, $gender_fmt.), 8.);
53   Suicidal_Thoughts_num = input(put(Suicidal_Thoughts, $thoughts_fmt.), 8.);
54   Dietary_Habits_num = input(put(Dietary_Habits, $diet_fmt.), 8.);
55   Sleep_Duration_num = input(put(Sleep_Duration, $sleep_fmt.), 8.);
56   Mental_Illness_History_num = input(put(Mental_Illness_History, $mental_fmt.), 8.);
57 run;
```

Rezultat:

id	Gender	Age	City	Academic_Pressure	CGPA	Study_Satisfaction	Sleep_Duration	Dietary_Habits	Degree	Suicidal_Thoughts	Work_Study_Hours	Financial_Stress	Mental_Illness_History	Gender_num	Suicidal_Thoughts_num	Dietary_Habits_num	Sleep_Duration_num	Mental_Illness_History_num
1	0	1	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

3. Frecvența depresiei în funcție de gen

a) Definirea problemei

Dorim să aflăm câte persoane suferă de depresie în funcție de gen (bărbați sau femei).

Această analiză ne ajută să înțelegem dacă există o diferență semnificativă între frecvența depresiei la femei față de bărbați.

b) Informații necesare

Vom folosi variabilele Gender și Depression, deja transformate în valori numerice:

- Gender: 0 = masculin, 1 = feminin

- Depression: 0 = fără depresie, 1 = cu depresie

De asemenea, au fost definite formate personalizate (PROC FORMAT) pentru a afișa aceste coduri numeric sub formă de etichete text (ex. 0 → „Male”).

c) Metodă de calcul

Pentru analiza frecvențelor combinate ale celor două variabile (Gender_num și Depression), s-a utilizat procedura PROC FREQ din SAS. Aceasta afișează un tabel de contingență, în care este indicat:

- numărul de persoane depressive și non-depressive
- grupate în funcție de gen

d) Rezolvare:

```

52 title "Frecventa depresiei in functie de gen";
53 proc freq data=depression_data_formatted;
54     tables Gender_num * Depression;
55     format Gender_num gender_fmt.
56           Depression depression_fmt.;
57 run;
58

```

e) Interpretare:

Frecventa depresiei in functie de gen				
The FREQ Procedure				
Frequency Percent Row Pct Col Pct	Table of Gender by Depression			
	Gender	Depression		
		0	1	Total
0		6432	9115	15547
		23.05	32.67	55.72
		41.37	58.63	
		55.62	55.80	
1		5133	7221	12354
		18.40	25.88	44.28
		41.55	58.45	
		44.38	44.20	
Total		11565	16336	27901
		41.45	58.55	100.00

Analiza arată că rata depresiei este **aproape egală** între bărbați și femei (58.6% vs. 58.4%), ceea ce sugerează că genul nu este un factor determinant semnificativ în apariția depresiei în acest eșantion.

Cu toate acestea:

- Numărul total al bărbaților este mai mare (55.7% din total), deci și numărul absolut al bărbaților cu depresie este mai mare (9115 vs. 7221).

- b) Procentajele apropiate sugerează că intervențiile sau politicile de sprijin psihologic ar trebui să fie **uniform distribuite** între sexe.

5. Interogare SQL asupra setului de date

a) Definirea problemei

Dorim să analizăm câți studenți sunt în fiecare categorie de gen (masculin și feminin), folosind limbajul SQL în cadrul SAS.

b) Informații necesare pentru rezolvare

Este necesar să lucrăm cu variabila Gender, care a fost anterior codificată astfel:

- 0 = masculin
- 1 = feminin

c) Metodă de calcul

Se folosește procedura PROC SQL pentru a selecta și grupa datele în funcție de gen și a calcula numărul de observații.

d) Rezolvare:

```
60 /* 4. Interogare SQL asupra setului de date */
61 proc sql;
62     select Gender_num format=gender_fmt.,
63            count(*) as Numar_Studenti
64     from depression_data_formatted
65     group by Gender_num;
66 quit;
67
```

e) Interpretare:

Rezultatul afișează câți studenți de sex masculin și câți de sex feminin sunt în eșantion.

Această informație este utilă pentru a înțelege distribuția pe gen a datasetului, înainte de interpretarea altor factori (precum depresia, stresul etc.).

Gender	Numar_Studenti
0	15547
1	12354

5. Statistici descriptive pentru stres și performanță

a) Definirea problemei

Dorim să analizăm nivelul mediu al presiunii academice, stresului financiar și performanței academice (CGPA) în rândul studenților, pentru a înțelege distribuția acestor factori care pot contribui la starea lor psihică și la apariția depresiei.

b) Informații necesare

Pentru această analiză sunt folosite următoarele variabile numerice din setul de date `depression_data`:

- `Academic_Pressure`: un scor de la 1 la 5 care măsoară presiunea academică percepută de student.
- `Financial_Stress`: un scor de la 1 la 5 care indică nivelul stresului financiar.
- `CGPA`: media generală ponderată (Cumulative Grade Point Average), pe o scară de la 0 la 10.

c) Metodă de calcul

Am utilizat procedura `PROC MEANS` din SAS pentru a calcula statistici descriptive pentru cele 3 variabile menționate:

- `mean` – media aritmetică
- `std` – abaterea standard
- `min` și `max` – valorile extreme

d) Rezolvare

```
70 proc means data=depression_data mean std min max maxdec=2;  
71     var Academic_Pressure Financial_Stress CGPA;  
72 run;
```

e) Interpretare

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum
Academic_Pressure	3.14	1.38	0.00	5.00
Financial_Stress	3.14	1.44	1.00	5.00
CGPA	7.66	1.47	0.00	10.00

Tabelul de mai sus prezintă statistici descriptive pentru cele trei variabile analizate: presiunea academică (Academic_Pressure), stresul financiar (Financial_Stress) și media generală (CGPA).

- **Academic_Pressure** are o medie de **3.14** și o abatere standard de **1.38**, cu un minim de 0 și un maxim de 5. Acest lucru indică o presiune academică moderată, dar cu variație mare între studenți.
- **Financial_Stress** are aceeași medie de **3.14**, dar o abatere standard ușor mai mare (**1.44**), ceea ce sugerează o distribuție variabilă a nivelului de stres financiar.
- **CGPA** are o medie ridicată de **7.66**, cu o abatere standard de **1.47**, ceea ce înseamnă că, în medie, studenții au performanțe academice bune, dar există și cazuri cu performanță scăzută (valoare minimă = 0).

6. Gruparea CGPA în categorii

a) Definirea problemei

Variabila CGPA (Cumulative Grade Point Average) exprimă media generală a performanței academice a studenților pe o scară de la 0 la 10. Deoarece aceasta este o variabilă numerică continuă, interpretarea directă a valorilor poate fi dificilă atunci când dorim să comparăm grupuri sau să corelăm performanța academică cu alți factori psihosociali, precum depresia, stresul financiar sau presiunea academică.

Prin urmare, scopul acestei etape este de a transforma variabila CGPA într-o variabilă categorială, care să permită clasificarea studenților în funcție de nivelul performanței academice. Această abordare facilitează analiza statistică, vizualizarea datelor și interpretarea rezultatelor în contexte comparative (ex: frecvența depresiei în funcție de nivelul CGPA).

b) Informații necesare

Pentru realizarea clasificării, am avut nevoie de următoarele:

- Variabila numerică CGPA, prezentă în setul de date, exprimă media generală a fiecărui student.
- Praguri de clasificare stabilite pe baza unor repere educaționale și empirice:
 - $CGPA < 2 \rightarrow$ nivel de performanță scăzut (posibil risc de eșec academic)
 - $2 \leq CGPA < 3 \rightarrow$ nivel mediu (performanță modestă, dar acceptabilă)
 - $CGPA \geq 3 \rightarrow$ nivel ridicat (performanță bună sau foarte bună)

Pe baza acestor praguri, s-a creat o variabilă nouă denumită CGPA_Level, de tip text (character), cu valorile:

- "Scăzut"
- "Mediu"
- "Ridicat"

c) Rezolvare:

```
74 data depression_data;  
75   set depression_data;  
76   if CGPA < 2 then CGPA_Level = "Sczut";  
77   else if 2 <= CGPA < 3 then CGPA_Level = "Mediu";  
78   else if CGPA >= 3 then CGPA_Level = "Ridicat";  
79 run;  
80  
81 proc freq data=depression_data;  
82   tables CGPA_Level;  
83 run;
```

d) Interpretare

The FREQ Procedure

CGPA_Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ridicat	27892	99.97	27892	99.97
Sczut	9	0.03	27901	100.00

Distribuția rezultatelor arată că aproape toți studenții (99.97%) au un CGPA ridicat, în timp ce doar 9 studenți din întregul eșantion se încadrează în categoria cu performanță scăzută.

Această observație poate avea mai multe interpretări:

- Pozitiv: Sistemul educațional funcționează eficient, iar studenții reușesc să obțină note bune în ciuda presiunii academice și a stresului financiar.
- Critic: Este posibil ca datele să fie dezechilibrate sau să provină dintr-o sursă care favorizează performanța ridicată (ex: doar studenți olimpici, universitate de top).

7. Corelație între stres și depresie

a) Definirea problemei

Scopul acestei analize este de a investiga dacă există o corelație semnificativă între depresie și următorii factori:

- presiunea academică (Academic_Pressure)
- stresul financiar (Financial_Stress)
- numărul de ore lucrate sau petrecute pentru studiu (Work_Study_Hours)

Prin această analiză, putem evidenția dacă acești factori contribuie într-un mod semnificativ la apariția depresiei și dacă merită incluși în modele predictive sau politici educaționale de prevenție.

b) Informații necesare

Pentru a analiza relația dintre depresie și factorii menționați, folosim procedura PROC CORR din SAS, care calculează coeficienți de corelație Pearson între variabilele numerice.

Variabilele implicate în analiză sunt:

- **Depression:** variabilă dependentă (0 = fără depresie, 1 = cu depresie), de tip numeric binar.
- **Academic_Pressure:** scor numeric (de la 1 la 5), exprimă presiunea resimțită de student în legătură cu performanța școlară.
- **Financial_Stress:** scor numeric (de la 1 la 5), exprimă dificultățile financiare.

c) Rezolvare

```
86 proc corr data=depression_data;
87     var Academic_Pressure Financial_Stress Work_Study_Hours;
88     with Depression;
89 run;
```

d) Interpretare

The CORR Procedure						
1 With Variables:		Depression				
3 Variables:		Academic_Pressure Financial_Stress Work_Study_Hours				

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Depression	27901	0.58550	0.49264	16336	0	1.00000
Academic_Pressure	27901	3.14121	1.38146	87643	0	5.00000
Financial_Stress	27898	3.13987	1.43735	87596	1.00000	5.00000
Work_Study_Hours	27901	7.15698	3.70764	199687	0	12.00000

Pearson Correlation Coefficients			
Prob > r under H0: Rho=0			
Number of Observations			
	Academic_Pressure	Financial_Stress	Work_Study_Hours
Depression	0.47483 <.0001 27901	0.36359 <.0001 27898	0.20856 <.0001 27901

- Presiunea academică are cea mai mare corelație pozitivă cu depresia ($r = 0.47$). Asta sugerează că studenții care resimt mai intens presiunea academică tind să manifeste mai frecvent simptome de depresie.
- Stresul financiar are, de asemenea, o corelație pozitivă semnificativă ($r = 0.36$) cu depresia. Astfel, studenții cu dificultăți financiare prezintă o probabilitate crescută de a suferi de depresie, ceea ce susține importanța programelor de burse și sprijin financiar.
- Numărul de ore lucrate/studiate are o corelație pozitivă slabă ($r = 0.21$) cu depresia. Asta sugerează că un volum ridicat de muncă sau studiu poate contribui ușor la starea de stres sau epuizare, dar nu este un predictor principal.

Analiza realizată în SAS confirmă rezultatele obținute în Python. Valorile coeficienților de corelație dintre depresie și factorii analizați (presiune academică, stres financiar și ore de lucru/studiu) sunt foarte apropiate în ambele cazuri.

8. Grafic: depresie în funcție de durata somnului

a) Definirea problemei

Durata somnului este un factor important care poate influența sănătatea mintală, inclusiv riscul de depresie. Prin această analiză dorim să vizualizăm legătura dintre durata somnului raportată de studenți și prezența sau absența depresiei.

Mai exact, vrem să observăm dacă există un tipar: de exemplu, dacă persoanele care dorm mai puțin de 5 ore prezintă o incidență mai mare a depresiei comparativ cu cele care dorm între 7–8 ore sau mai mult.

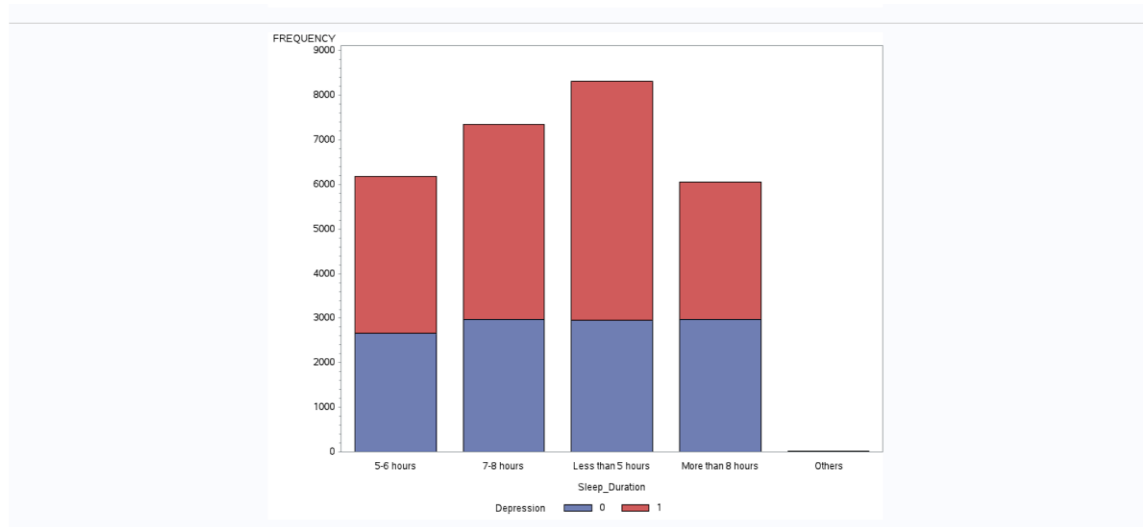
b) Informații necesare

Folosim variabilele Sleep_Duration (0–3) și Depression.

c) Rezolvare

```
91 proc gchart data=depression_data;
92     vbar Sleep_Duration / subgroup=Depression;
93 run;
```

d)Interpretare



Graficul cu bare arată distribuția studenților în funcție de durata somnului (Sleep_Duration), defalcată pe prezența depresiei (Depression – 1 = cu depresie, 0 = fără depresie).

- Cea mai mare frecvență a cazurilor de depresie (culoare roșie) apare la studenții care dorm „less than 5 hours”, urmați de cei care dorm „7-8 hours”.
- Categoria „More than 8 hours” are un număr similar de cazuri cu celelalte, dar cu o proporție ceva mai mică de depresie.
- Studenții care dorm „5-6 hours” par a avea cea mai mică incidență a depresiei dintre categoriile principale.
- Categoria „Others” are o frecvență neglijabilă și poate fi ignorată sau curățată din date.

Concluzie

Analiza realizată în SAS a oferit o imagine clară asupra principalilor factori asociați cu depresia în rândul studenților. Am utilizat o serie de proceduri statistice, transformări de date și vizualizări pentru a înțelege mai bine relațiile dintre variabile precum presiunea academică, stresul financiar, durata somnului, performanța academică și prezența depresiei.

Rezultatele au arătat că:

- Presiunea academică și stresul financiar sunt semnificativ corelate cu depresia, fiind posibili factori de risc.
- Durata somnului influențează incidența depresiei: studenții care dorm mai puțin de 5 ore prezintă un număr mai mare de cazuri.
- Majoritatea studenților au performanțe academice ridicate ($CGPA \geq 3$), însă acest lucru nu garantează absența depresiei.

În concluzie, analiza SAS confirmă faptul că depresia în rândul studenților este un fenomen complex, influențat de mai mulți factori psihosociali, iar instrumentele utilizate permit extragerea de informații valoroase din date pentru luarea unor decizii informate.