Projekt ZZN 2025/26

Data z výzkumu stresu studentů

Část druhá: řešení

Autoři: Jakub Vlk (xvlkja07), Michal Plšek (xpplsek04)

Datum odevzdání: 3.12.2025

Rekapitulace zadání

Segmentace studentů dle dostupných dat

Nalezení skupiny studentů s podobnými stresovými predispozicemi. Bylo by nejspíš vhodné provádět analýzu pro každý vzdělávací stupeň separátně, protože vliv některých atributů bude mezi nimi mít velmi pravděpodobně jinou úroveň (Academic pressure from your home, Nezdravé návyky apod.). Na základě jejich množství a charakteristiky je lze zobrazit, stanovit skupiny studenů zvládající podobný stres podobnými způsoby (pro vizualizaci bude nejprve potřeba provést PCA na převod do max. 3 atributů). Na základě těchto skupin získáme obecný pohled na zvládání stresu mezi studenty.

Analýza souvisejících atributů

Bude užitečné provést korelační analýzu a zjistit, které atributy spolu jak souvisejí. Na základě toho by se daly vyspecifikovat důležité a nedůležité parametry pro modelování.

Prediktivní model pro predikci chování při vystavení stresu

Jaký typ chování student pravděpodobně zvolí, když se pod tlak dostane. Konkrétně nás zajímá, jak spolehlivě dokážeme na základě vstupních dat (akademický stupeň, tlaky, prostředí) odhadnout, zda student bude inklinovat k 'emocionálnímu zhroucení' na rozdíl od konstruktivnějších strategií, jako je 'analýza situace' nebo 'vyhledání sociální podpory'. Cílem je vytvořit systém včasného varování. Pokud model u nově příchozího studenta identifikuje vysokou pravděpodobnost neefektivní reakce na stres, můžeme mu proaktivně nabídnout workshopy zaměřené na budování odolnosti a nácvik efektivních copingových strategií dříve,
než se u něj negativní vzorce chování zafixují.

Popis dat

Název sloupce (EN)	Název sloupce (CZ)	Popis	Typ sloupce	Možné hodnoty
Timestamp	Časová značka	Čas vložení záznamu	datetime	–
Your Academic Stage	Vzdělávací stupeň	high-school / undergraduate / postgraduate	kategorické	high-school, undergraduate, post-graduate
Peer pressure	Tlak sociálního okolí	Jak velký tlak pociťujete od vrstevníků	kategorické (1–5)	1 = nejlepší, 5 = nejhorší
Academic pressure from your home	Tlak na prospěch z domácnosti	Jak velký tlak je vyvíjen doma	kategorické (1–5)	1 = nejlepší, 5 = nejhorší
Study Environment	Studijní prostředí	Míra rušnosti ve studiu	kategorické	Peaceful / Noisy / Disrupted
What coping strategy you use as a student?	Strategie zvládání stresu	Jak reagujete na studijní stres	kategorické	Analyze intellectually / Social support / Emotional breakdown
Do you have any bad habits like smoking, drinking daily?	Nezdravé návyky	Kouření, alkohol, drogy…	kategorické	Yes / No / Prefer not to say
What would you rate the academic competition in your student life	Akademická rivalita	Jak silná je akademická soutěživost	kategorické (1–5)	1 = nejlepší, 5 = nejhorší
Rate your academic stress index	Index akademického stresu	Celkové hodnocení stresu	kategorické (1–5)	1 = nejlepší, 5 = nejhorší

U kategorických číselných hodnot je 1 = nejlepší, 5 = nejhorší (známkování jako ve škole)

Řešení

Rozbor a před zpracování dat

Před zahájením řešení jsem převedli všechny kategorických (textových) proměnných. Pro účely analýzy byly tyto atributy transformovány na číselné reprezentace pomocí následujících metod:

A) Ordinální kódování (Ordinal Encoding)
U atributů, které vykazují přirozené logické uspořádání (hierarchii), bylo zvoleno celočíselné mapování zachovávající toto pořadí.

- **Vzdělávací stupeň (Your Academic Stage):** Atribut byl převeden na vzestupnou škálu od 1 do 3, aby reflektoval postupující úroveň vzdělání.

  High-school - 1
  Undergraduate - 2
  Post-graduate - 3

- **Studijní prostředí (Study Environment):** Zde bylo zvoleno kódování, které koresponduje s předpokládanou mírou stresu (horší podmínky = vyšší hodnota). Tím byla zajištěna konzistence s ostatními stresovými faktory (kde 5 znamená nejvyšší stres).

Peaceful (Klidné) -> 1
Noisy (Hlučné) - 2
Disrupted (Narušované) - 3

B) Binární kódování
U atributů nabývajících pouze dvou stavů byla provedena binarizace.

- **Nezdravé návyky (Bad habits):** Odpovědi byly převedeny na indikátor přítomnosti zlozvyku.

  Yes - 1
  No - 0

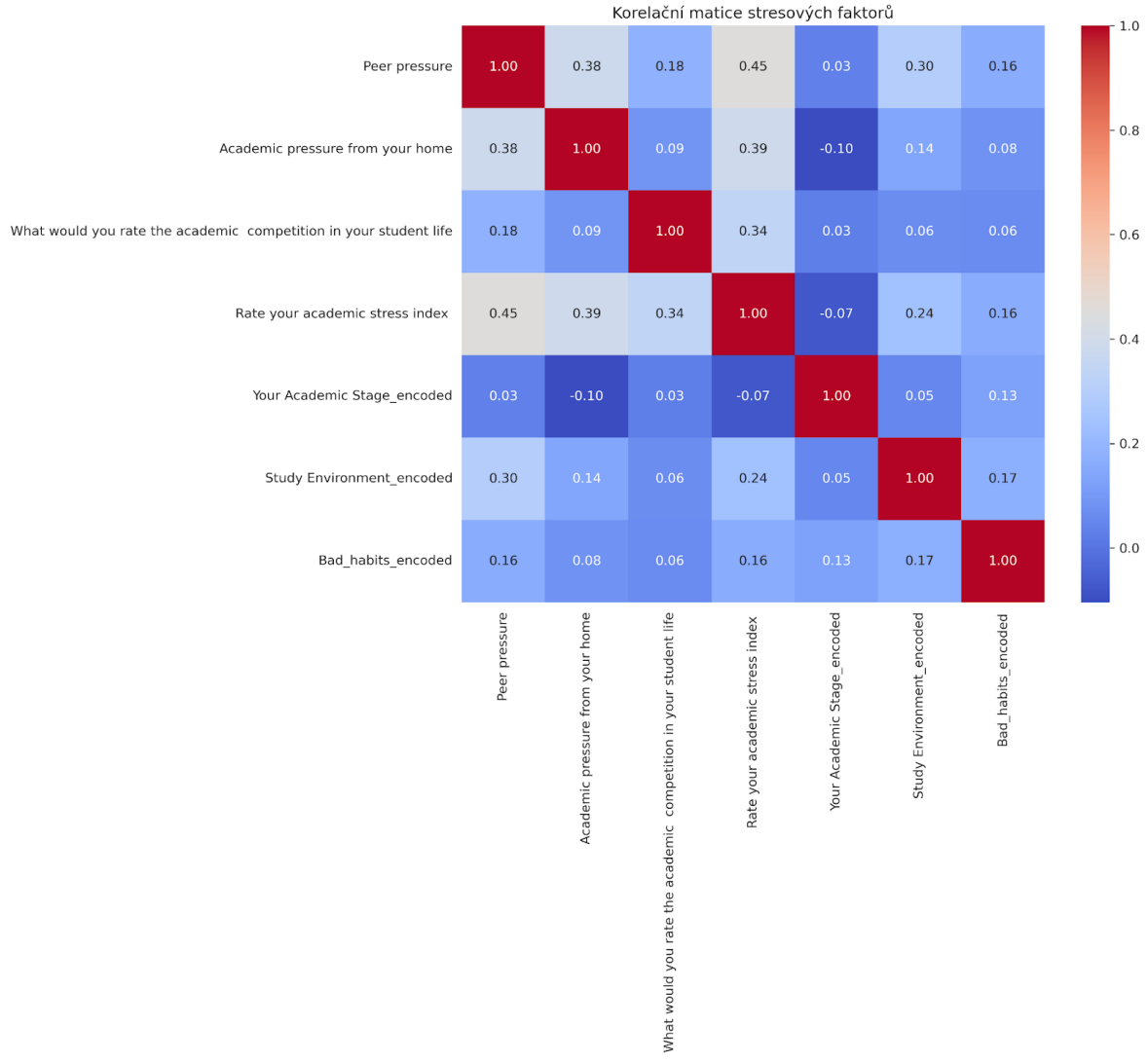
- *(Poznámka: Hodnoty "Prefer not to say" byly ošetřeny jako chybějící hodnoty).*

C) Ponechání číselných škál
Atributy, které již v dotazníku obsahovaly číselné hodnocení (např. *Peer pressure*, *Academic pressure*), byly ponechány v původním rozsahu 1–5.

Poznámka: Pro metody kdy to dávalo smysl tak jsem použili adhoc encoding.

Analýza souvisejících atributů

Provedená korelační analýza měla za cíl identifikovat faktory, které nejvýznamněji ovlivňují celkový index stresu (Rate your academic stress index). Z výsledné matice vyplývá, že nejsilnějším identifikovaným prediktorem stresu je tlak sociálního okolí (Peer pressure) s korelačním koeficientem 0,45, následovaný tlakem rodiny na studijní výsledky (Academic pressure from home) s hodnotou 0,39. Tyto hodnoty indikují středně silnou pozitivní závislost a naznačují, že externí očekávání a sociální faktory hrají při vzniku stresu klíčovou roli, často významnější než samotné studijní podmínky.



Dále byla pozorována slabší, avšak existující souvislost s vnímáním akademické rivality (0,34) a kvalitou studijního prostředí (0,24). Zásadním zjištěním je naopak faktická absence lineární závislosti mezi stupněm studia a stresem (korelace -0,07 u atributu Your Academic Stage). Tento výsledek vyvrací předpoklad, že by s vyšším stupněm vzdělání (např. u doktorandů) automaticky rostla míra stresu – data ukazují, že středoškoláci prožívají srovnatelnou zátěž. Vzhledem k tomu, že žádný z jednotlivých atributů nevykazuje extrémně silnou korelaci (např. nad 0,7), nelze na základě těchto výsledků hovořit o jednoznačném původci či nepůvodci stresu.

Segmentace studentů dle dostupných dat

Pro segmentaci uživatelů jsme použili všechny atributy kromě stresu. Vyřadili jsme taková data, která byla nekompletní (NaN a podobně).

Shlukování nebudeme provádět pro střední školy (High school) ani pro postgraduální studia (Post-graduate), protože ani jedna z těchto skupin neobsahuje dostatečný počet vzorků, u kterých by se dalo předpokládat průkazné nalezení nějakých množin.

Vzhledem k tomu jsme se rozhodli porovnat pomocí shlukovací analýzy všechny studenty s těmi na vysoké škole.

Clustering

V první fázi shlukové analýzy byla pro redukci dimenzionality aplikována standardní metoda **PCA (Principal Component Analysis)**. Jak je však patrné z přiloženého srovnání, tato lineární metoda nedokázala efektivně separovat jednotlivé skupiny studentů. Vzhledem k diskrétní povaze vstupních dat (převážně ordinální škály 1–5) vytvářela PCA lineární artefakty a překryvy, které znemožňovaly jasnou identifikaci shluků.

Z tohoto důvodu jsme přistoupili k testování pokročilejších nelineárních metod (t-SNE, UMAP, Kernel PCA) prostřednictvím analýzy citlivosti (Sensitivity Analysis). Na základě vizuálního porovnání separability shluků byla jako nejvhodnější metoda zvolena **UMAP (Uniform Manifold Approximation and Projection)**.

Pro finální model byl zvolen parametr **n_neighbors=10**. Tato konfigurace se ukázala jako optimální pro oba analyzované datasety z následujících důvodů:

1. **Respektování nelinearity:** Na rozdíl od PCA dokáže UMAP zachytit komplexní vztahy v datech, kde neexistuje silná lineární korelace, a efektivně "rozbalit" shluky do 2D prostoru.
2. **Důraz na lokální strukturu:** Volba nižší hodnoty sousedů (10) umožňuje algoritmu zaměřit se na lokální topologii dat. Zatímco vyšší hodnoty (30–50) vedly ke slévání dat do jedné masy (globální pohled), hodnota 10 umožnila izolovat menší, specifické podskupiny studentů (např. rizikové skupiny), což bylo primárním cílem analýzy.
3. **Stabilita výsledků:** Při tomto nastavení vykazovaly výsledné shluky nejvyšší stabilitu a logickou interpretovatelnost napříč oběma zkoumanými datasety.

Interpretace shluků - Vysokoškoláci

Cluster 0 představuje referenční skupinu s nejnižším naměřeným indexem stresu o hodnotě 3,18. Studenti v této skupině disponují velmi kvalitním studijním zázemím, nemají rizikové návyky a vnímají nejnižší míru akademické soutěživosti. Jedná se o stabilní segment s ideálními podmínkami pro studium.

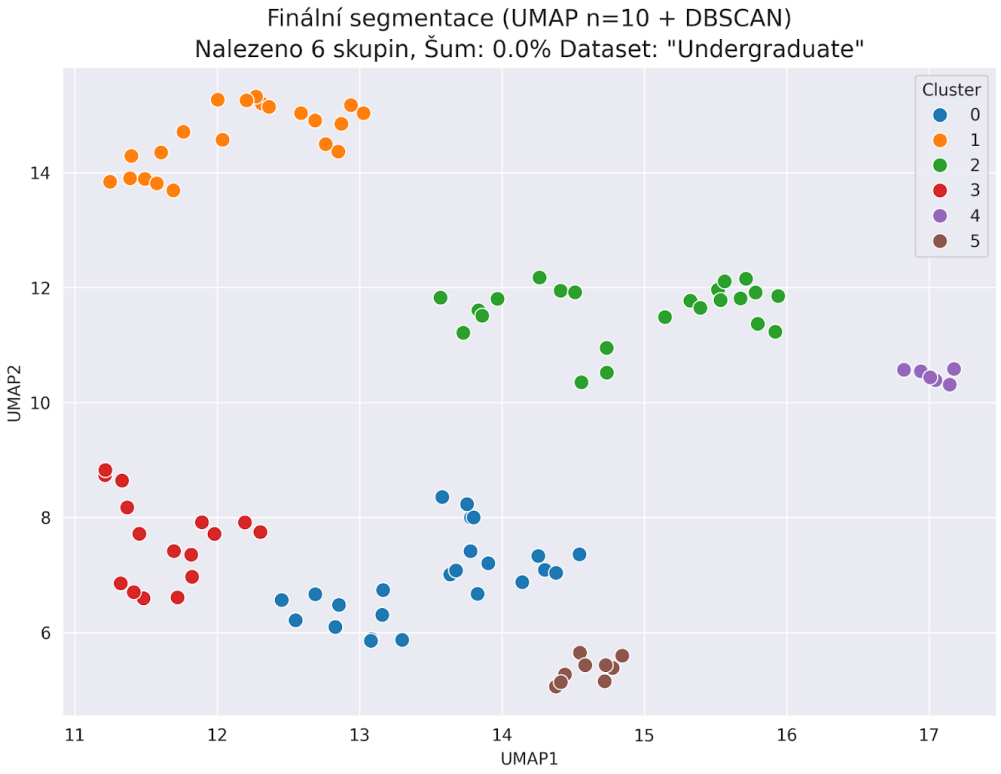
Cluster 1 zahrnuje studenty, kteří mají srovnatelně dobré materiální podmínky jako předchozí skupina, avšak vykazují vyšší hladinu stresu na úrovni 3,53. Klíčovým rozdílem je zde intenzivní vnímání akademické soutěživosti s hodnotou 4,20. Stres zde pramení primárně z osobních ambicí a rivality, nikoliv z externích překážek.

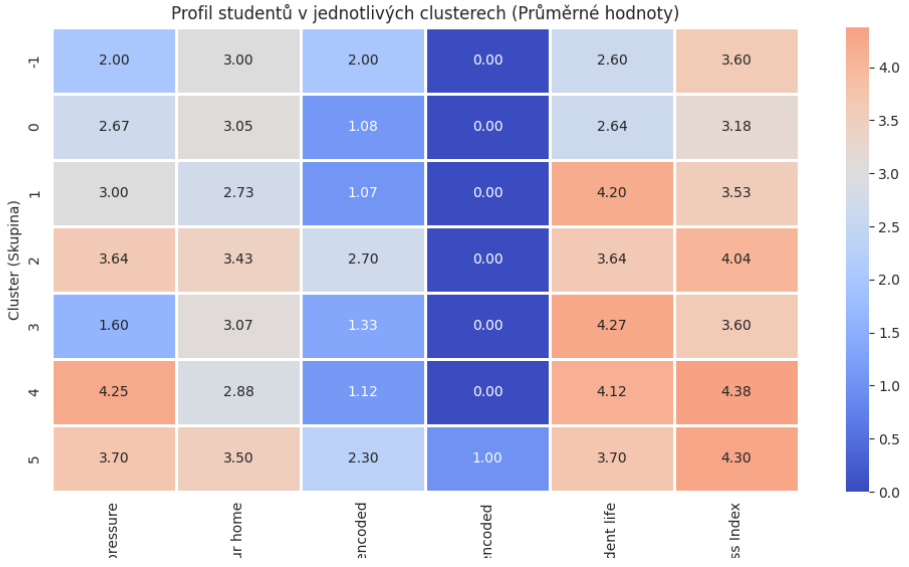
Cluster 2 je definován nejhoršími studijními podmínkami ze všech skupin, kde hodnota prostředí dosahuje 2,70. V kombinaci se zvýšeným tlakem ze strany rodiny to vede k vysoké hladině stresu 4,04. Jedná se o skupinu, jejíž psychická zátěž je přímým důsledkem nevhodného fyzického zázemí.

Cluster 3 tvoří specifický profil studentů, kteří jsou imunní vůči tlaku vrstevníků, jehož hodnota je zde absolutně nejnižší. Přesto pociťují silnou akademickou soutěživost a střední míru stresu 3,60. Pravděpodobně jde o introvertně laděné jedince, kteří ignorují sociální dynamiku, ale soustředí se na výkon.

Cluster 4 byl identifikován jako nejrizikovější skupina s nejvyšším indexem stresu 4,38. Paradoxně mají tito studenti výborné studijní prostředí, avšak čelí extrémnímu tlaku vrstevníků s hodnotou 4,25. Data ukazují, že sociální očekávání má na tuto skupinu devastující dopad.

Cluster 5 je unikátní skupina se stoprocentním výskytem špatných návyků. Tento faktor se kombinuje s horším prostředím a vysokým tlakem ze všech stran, což ústí ve velmi vysoký stres 4,30. Zde se prokazuje silná souvislost mezi nezdravým životním stylem a psychickou zátěží.





Interpretace shluků - Celý dataset

Cluster 1 představuje nejméně stresovanou skupinu v celém datasetu s indexem stresu 2,90. Tito studenti vykazují nízké hodnoty ve všech sledovaných stresorech, mají dobré zázemí a nepociťují výrazný tlak okolí.

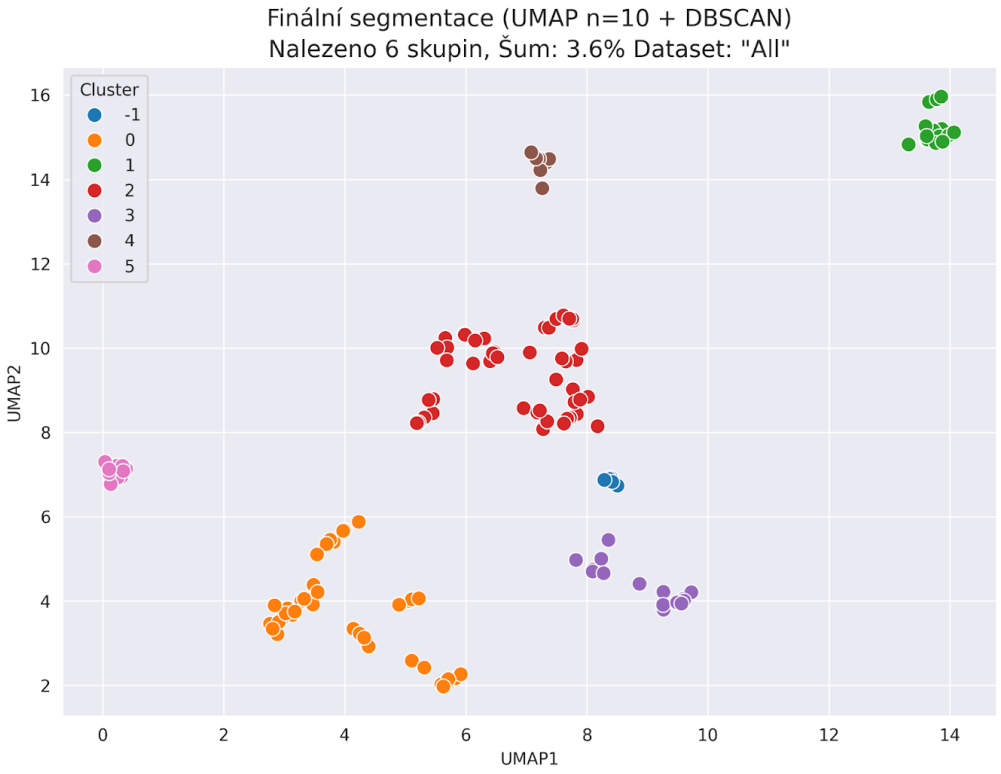
Cluster 2 je skupina se střední mírou stresu 3,61. Studenti mají kvalitní studijní prostředí a nemají špatné návyky, avšak pociťují vysokou míru akademické soutěživosti a tlaku z domova.

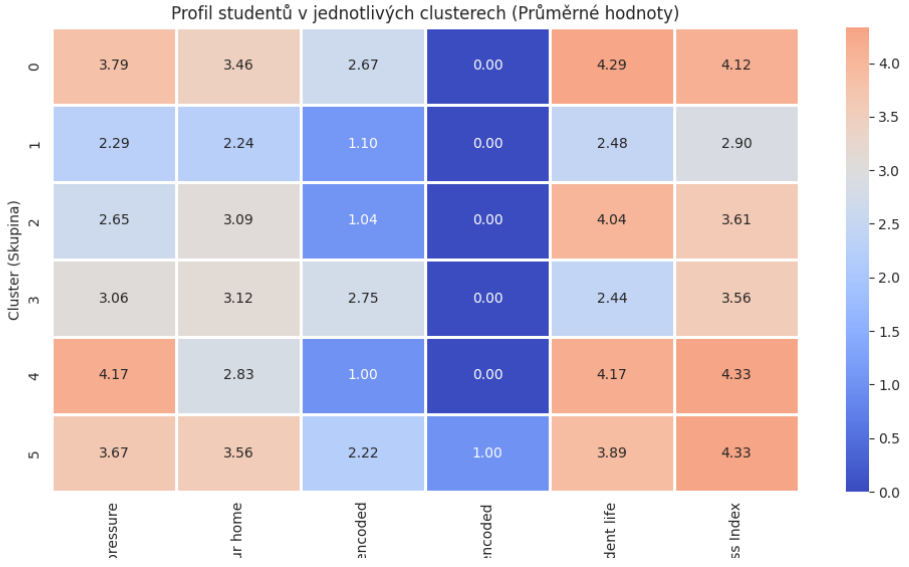
Cluster 3 je charakteristický narušeným studijním prostředím s hodnotou 2,75, což je hlavním zdrojem jejich nepohody. Ostatní faktory jako tlak vrstevníků jsou zde průměrné, což vede ke střední hladině stresu 3,56.

Cluster 0 sdružuje studenty s kombinací více nepříznivých faktorů. Trpí narušeným prostředím a zároveň velmi vysokou akademickou soutěživostí a tlakem vrstevníků. Tato kumulace stresorů vede k vysokému indexu stresu 4,12.

Cluster 4 je vysoce riziková skupina s nejvyšším stresem v celém datasetu o hodnotě 4,33. Stejně jako u vysokoškoláků je zde hlavním spouštěčem extrémní tlak vrstevníků a soutěživost, navzdory tomu, že mají dobré studijní prostředí.

Cluster 5 kopíruje vzorec z podskupiny vysokoškoláků. Jde o studenty se špatnými návyky a horším prostředím, což vede ke shodně vysoké hladině stresu 4,33





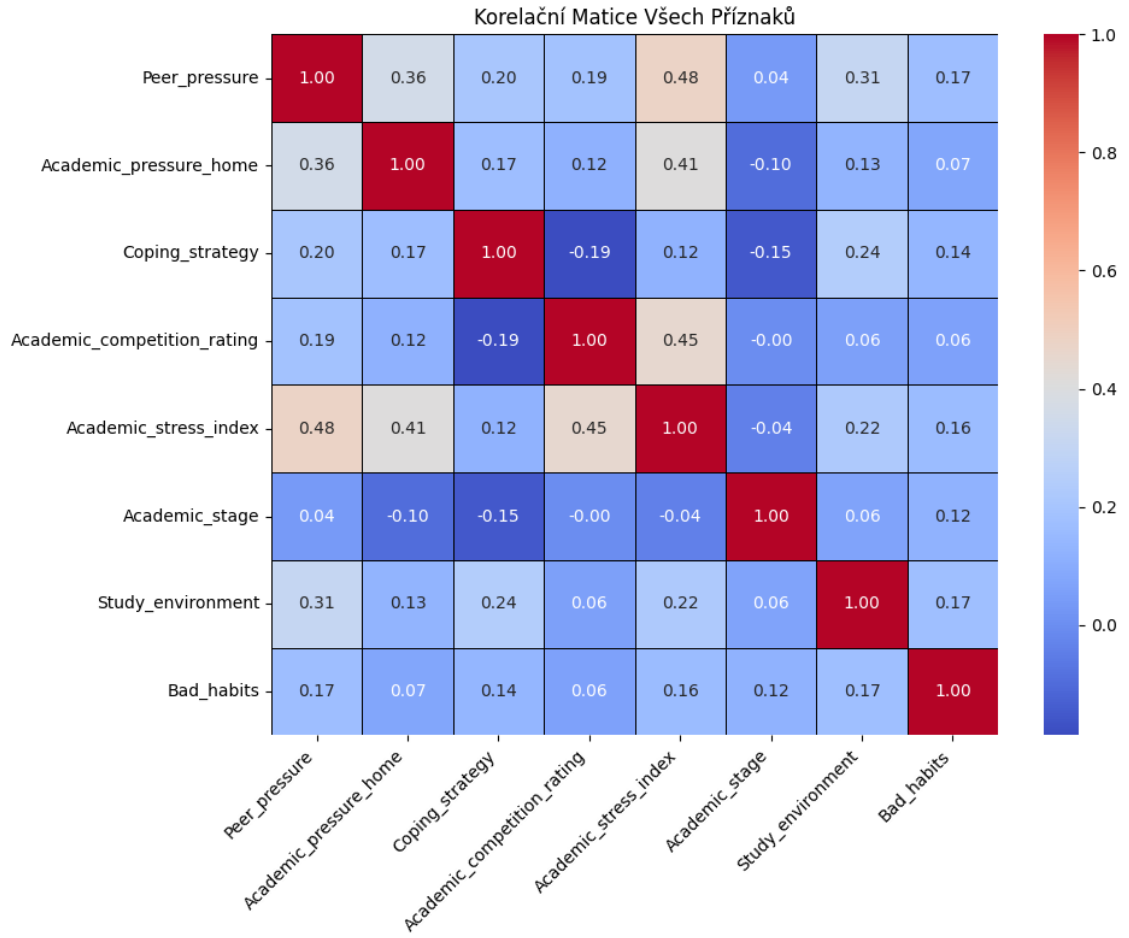
Porovnání shluků

Při porovnání výsledků segmentace pro podskupinu vysokoškoláků a kompletní dataset je patrná vysoká míra shody v identifikovaných profilech. Struktura shluků zůstává v obou případech stabilní, což potvrzuje, že mechanismy vzniku stresu jsou v rámci zkoumané populace konzistentní.

Analýza odhalila, že nejrizikovějším faktorem pro vznik extrémního stresu není špatné studijní zázemí, ale psychosociální tlak. V obou analýzách vykazují nejvyšší míru stresu skupiny definované vysokým tlakem vrstevníků a akademickou soutěživostí (Cluster 4). Druhým nejvýznamnějším faktorem je životní styl, kdy přítomnost špatných návyků spolehlivě indikuje vysokou psychickou zátěž (Cluster 5). Materiální podmínky, jako je hluk či nekvalitní prostředí, sice stres zvyšují, ale nedosahují takových extrémů jako sociální faktory.

Predikce chování při vystavení stresu

Nejprve došlo k ověření toho, jak spolu parametry korelují (Pearson):



podle výstupu zde není žádná kritická korelace. Pro predikci je třeba provést klasifikaci s určováním parametru *academic stress index*. Nejprve bylo třeba vyřadit nevalidní vzorky (chybějící hodnoty - NaN). Načítání dat pro různé konfigurace:

- pro Random forest bylo dosaženo nejlepších výsledků s načtením parametrů *Peer pressure, Academic pressure, Academic competition* jako hodnoty integer 1-5 (tedy ne jako one-hot encoded parametry)

- u SVC, k-NN i neuronové sítě docházelo po změně načítání všech atributů jako *str* (tj. všechny by byly kodovány jako one-hot) k rapidnímu snížení jak trénovací, tak validační přesnosti, takže nastavení zůstalo stejné jako pro random forest

Sloupec dat *Timestamp* byl odstraněn. Zbylé hodnoty byly převedeny pomocí one-hot encoding s *get_dummies(drop_first = True)*, tím se dá zabránit multikolinearitě.
Dataset byl rozdělen na dvě části: na validační a trénovací + testovací množinu. Validační 25% část je použita na vytvoření confusion matice, zbytek je určen k trénování a cross-validaci, také je upsamplován pomocí SMOTE/ADASYN.
Protože se jedná o skutečně malý dataset, a vzhledem k potřebě detekovat převážně problémové případy (stres), byl *Academic stress index* rozdělen dvě úrovně, vysoký (hodnoty 4, 5) a nízký (1, 2, 3). Rovněž vzhledem k velikosti datasetu byla použita stratifikace (*StratifiedKFold* pro účely cross-validace) s 5x opakováním, tj. dělením na 5 částí s *shuffle=True*. Vnitřně je díky *folds=5* poměr stanoven na 80% vzorků trénovací sada, 20% vzorků testovací sada.
Protože je rovněž potřeba upsamplovat dataset, bylo nutné použít *ImbPipeline* z modulu *imbalanced-learn*, která umožňuje do *Pipeline* (nástroj ze *scikit* k přehlednější synchronizaci procesu tvorby modelu) přidat speciální objekty, jako je např. *SMOTE.* Po SMOTE byl aplikován standartní scaler (*StandardScaler*) k aplikaci z-normalizace (hlavně pro k-NN a SVM, např. u stromů a lesů k ničemu moc nebude, ale nijak neuškodila). Vyzkoušeli jsme i ADASYN, který přestože na omezené validační sadě dával lehce horší výsledek, tak na testovací sadě vycházel lepší *recall*, takže tam zůstal. Následující vrstva *FeatureSelection* slouží pro detekci vhodných atributů, na základě kterých model sestavit. Byly použity následující:

- u *random forest* nejlepší výsledky poskytoval:

```
('feature_selection', SelectFromModel(
    RandomForestClassifier(
        n_estimators=100,
        random_state=42,
        class_weight='balanced',
       ),
))
```

tj. *RandomForestClassifier* tu působí jinak než u cílové klasifikace (jeho úkolem není predikovat cílovou třídu, ale určit jak moc každý příznak přispívá ke správnému rozdělení dat). Dobré výsledky poskytovaly i dva následující:

```
('feature_selection', RFE(
    estimator=LogisticRegression(random_state=42, solver='lbfgs',
```

```
max_iter=1000),
        n_features_to_select=4,
    )),

    ('feature_selection', SelectFromModel(LinearSVC(penalty="l1",
dual="auto", random_state=42))),
```

tj. logistická regrese s *lbfgs* solverem a *LinearSVC* (vychází z SVM, jen určeno pro klasifikaci).

- u SVC *(SVC = SVM* varianta pro klasifikaci*)* nejlepší výsledky poskytoval:

```
clf = ImbPipeline([
    ('adasyn', ADASYN(random_state=42)),
    ('scaler', StandardScaler()),
    ('feature_selection', RFE(
        estimator=LogisticRegression(random_state=42, max_iter=1000),
    )),
    ('classification', LinearSVC(
        random_state=42,
        penalty='l2',
        dual=True,
        class_weight='balanced'
    )),
])
```

tj. pro feature selection metodu RFE (*recursive feature elimination*) s logistickou regresí. V tomto případě logistická regrese přiřazovala váhy jednotlivým features, a pomocí RFE byly postupně testovány menší a menší množiny atributů, až do cílového počtu *n*. Tomuto předcházela normalizace (SVC ji vyžaduje).

- u k-NN *(k Nearest neighbours)* poskytnul výsledky:

```
clf = ImbPipeline([
    ('adasyn', ADASYN(random_state=42)),
    ('scaler', StandardScaler()),
    ('feature_selection', SelectKBest(
        score_func=mutual_info_classif,
    )),
    ('classification', KNeighborsClassifier(n_jobs=-1)),
])
```

výběr atributů tedy zajistil výběr *n* nejlepších atributů, v závislosti na skóre vzájemné informace mezi atributem a cílovou proměnnou (*stress indexem*).

- u MLP (*multi-layer perceptron*, verze pro klasifikaci) poskytoval nejlepší výsledky:

```
clf = ImbPipeline([
    ('adasyn', ADASYN(random_state=42)),
    ('scaler', StandardScaler()),
    ('feature_selection', SelectKBest(
        score_func=mutual_info_classif,
    )),
    ('classification', MLPClassifier(
        random_state=42,
        max_iter=1000,
        early_stopping=True,
        n_iter_no_change=20,
        solver="adam",
    )),
])
```

tedy výběr nejlepších atributů v závislosti na vzájemné vzdálenosti od cílového (aby trénování netrvalo moc dlouho), a klasifikace pomocí multi-layer perceptronu (s ADAM, s předčasným zastavením).

Následně bylo nad parametry prvků v pipeline provedeno globální *GridSearchCV* prohledávání v mřížce, s maximalizací F1 skóre. Později bylo změněno na maximalizaci *recall* hodnoty.
Vzhledem k tomu, že se snažíme co nejvíce předcházet stresu i za cenu vyšších FP (false positive), zajímá nás hodnota recall (tedy vyplatí se i skutečné případy nízkého stresu označit chybně jako vysoký stres, aby byla jistota že model chybně nezařadí vysoký stres jako nízký). U každého modelu jsme v rámci prohledávání na mřížce metodou pokus omyl postupně měnili ručně sady parametrů, dokud se neobjevily poměrně rozumné výsledky (viz. zdrojový kód, *grid_params*). Po dokončení prohledávání je vnitřně vybrán nejlepší model, parametry byly postupně redukovány na:

Typ modelu	**Parametry**
Random forest	{'adasyn__n_neighbors': 3, 'classification__max_depth': 9, 'classification__min_samples_leaf': 3, 'feature_selection__threshold': '0.7*mean'}
SVC	{'adasyn__n_neighbors': 5, 'classification__C': 0.1,

	'classification__class_weight': None, 'feature_selection__n_features_to_select': 5}
k-NN	{'adasyn__n_neighbors': 3, 'classification__n_neighbors': 13, 'classification__p': 1, 'classification__weights': 'uniform', 'feature_selection__k': 8}
MLP	{'adasyn__n_neighbors': 4, 'classification__activation': 'relu', 'classification__alpha': 0.0001, 'classification__hidden_layer_sizes': (19,), 'feature_selection__k': 10}

K ohodnocování modelu dochazí vnitřně, na základně již zmíněného rozdělení na trénovací a testovací sadu (80% a 20%). Díky cross-validaci dochází k věrohodnějšímu ohodnocení modelu:

Typ modelu	**Přízn. (před/po)**	**Příznaky**	**Prům. přesnost**
Random forest	11/6	*['Peer_pressure', 'Academic_pressure_home', 'Academic_competition_rating', 'Academic_stage_undergraduate', 'Study_environment_Peaceful', 'Study_environment_disrupted']*	Přesnost: 73.1% Odchylka: 7.66% F1 prům.: 71.45% F1 váh.: 73.25%
SVC	11/5	*['Peer_pressure', 'Academic_pressure_home', 'Academic_competition_rating', 'Study_environment_disrupted', 'Coping_strategy_Social support (friends, family)']*	Přesnost: 71.24% Odchylka: 8.31% F1 prům.: 69.42% F1 váh.: 71.29%

k-NN	11/8	*['Peer_pressure', 'Academic_pressure_home', 'Academic_competition_rating', 'Study_environment_disrupted', 'Coping_strategy_Emotional breakdown (crying a lot)', 'Coping_strategy_Social support (friends, family)', 'Bad_habits_Yes', 'Bad_habits_prefer not to say']*	Přesnost: 67.29% Odchylka: 8.26% F1 prům.: 66.46% F1 váh.: 67.74%
MLP	11/10	*['Peer_pressure', 'Academic_pressure_home', 'Academic_competition_rating', 'Academic_stage_undergraduate', 'Study_environment_Peaceful', 'Study_environment_disrupted', 'Coping_strategy_Emotional breakdown (crying a lot)', 'Coping_strategy_Social support (friends, family)', 'Bad_habits_Yes', 'Bad_habits_prefer not to say']*	Přesnost: 69.19% Odchylka: 4.11% F1 prům.: 58.25% F1 váh.: 63.99%

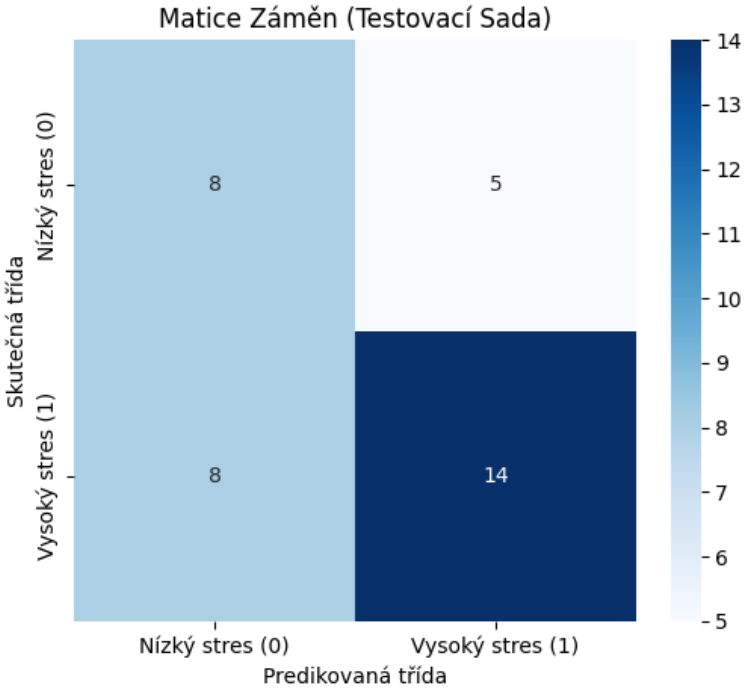
Pozn.: F1 váh. značí F1 pro každou kategorii zvlášť, odchylka je std odchylkou od Přesnosti (tj. kolikrát se model trefil do správného řešení).

Přestože u takto malého datasetu dojde ke zhoršení trénovacích výsledků (i když na tak malé validační sadě se toho moc nepozná), je dobré získat i *confusion matici*,

aby bylo zřejmé co se v modelu děje. Je získána z validačních dat, která nejsou použita při trénování ani testování. Pro validační množinu jsou rovněž vytvořeny přehledové statistiky.

Pro *random forest*:

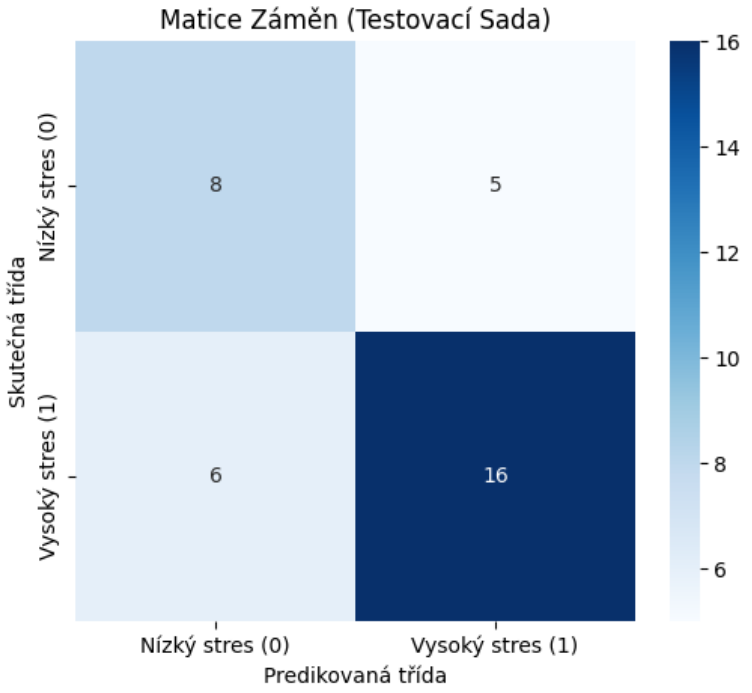
	precision	**recall**	**f1-score**	**support**
Nízký stres (0)	0.5	0.69	0.58	13
Vysoký stres (1)	0.76	0.59	0.67	22
přesnost			0.63	35
avg	0.63	0.64	0.62	35
weight. avg	0.67	0.63	0.63	35



Pro *SVC*:

	precision	**recall**	**f1-score**	**support**

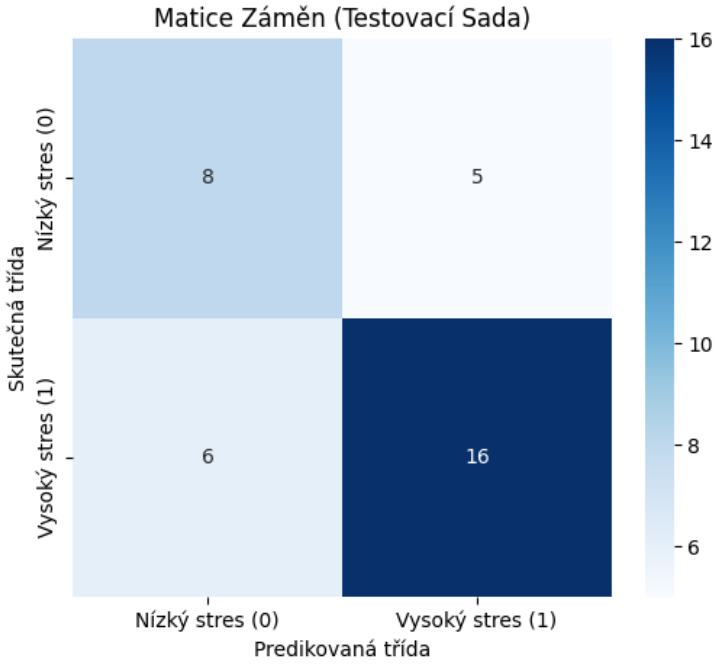
Nízký stres (0)	0.57	0.62	0.59	13
Vysoký stres (1)	0.76	0.73	0.74	22
přesnost			0.69	35
avg	0.67	0.67	0.67	35
weight. avg	0.69	0.69	0.69	35



Pro *k-NN*:

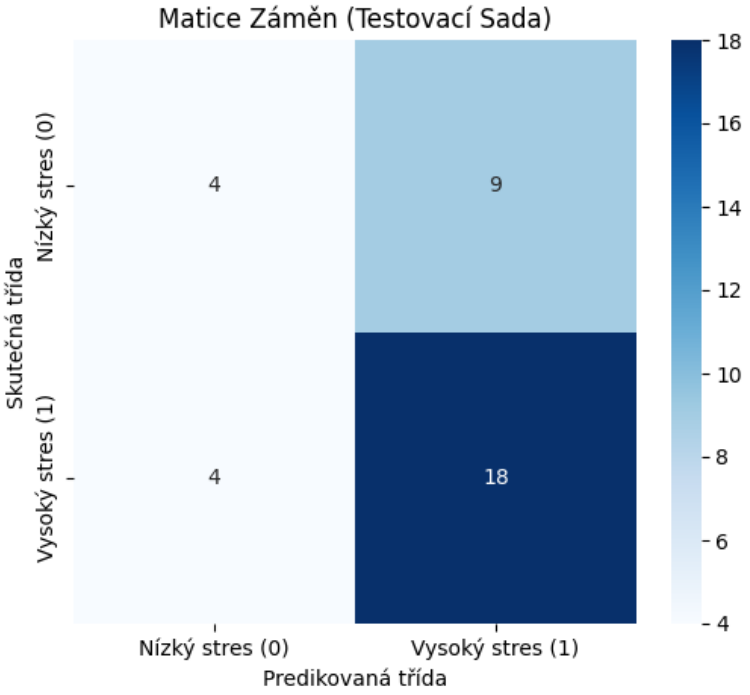
	precision	**recall**	**f1-score**	**support**
Nízký stres (0)	0.57	0.62	0.59	13
Vysoký stres (1)	0.73	0.76	0.74	22
přesnost			0.69	35
avg	0.67	0.67	0.67	35

weight. avg	0.69	0.69	0.69	35



Pro *MLP*:

	precision	**recall**	**f1-score**	**support**
Nízký stres (0)	0.50	0.31	0.38	13
Vysoký stres (1)	0.67	0.82	0.73	22
přesnost			0.63	35
avg	0.58	0.56	0.56	35
weight. avg	0.60	0.63	0.60	35



Závěry z predikce

Vytvoření korelační matice a podrobnějších statistik z modelu sice omezilo množství dat pro trénování (i po vyvážení), ale bylo nutné, protože bychom jinak nezjistili co se v modelu děje (muselo být na separátních datech). Neuronová síť, byť jednoduchá (skrytá vrstva s 19 parametry), si počínala dobře na validačním datasetu (vysoký recall pro *stress index*), otázkou je nakolik to bylo způsobeno množstvím dat.

SVC i náhodný les dosahovaly dobrých výsledků i s nízkým počtem parametrů. U náhodného lesa je dobrá hodnota F1, i váhovaná, model si tedy počíná dobře jak z hlediska *recall* tak *precision*. Na validační sadě zachytilo SVM i k-NN lepší množství FN (*false negative*), o což nám primárně jde. U MLP toto zachycení bylo ještě vyšší, na úkor falešně pozitivních (FP), které nám z hlediska účelu tak moc nevadí. Ve výsledku bych tedy použil bych tedy kombinaci náhodného lesa a MLP.

Pokud jde o rozptyl přesnosti, vzhledem ke cross-validaci s n=5 máme k dispozici jen 5 hodnot. U MLP je rozptyl dobrý (< 5%), u SVC a k-NN (8%) ukazuje na nestabilitu modelu, muselo by se vyzkoušet na více vzorcích.

Grid search, byť výpočetně náročnější, výrazně napomohl získání některých parametrů, který byly pak vzájemně ručně doladěny.