C10: Student stress factors
Mona Tolmats
Kirke Valt
Liisa-Lotte Pehter

Task 1

https://github.com/MonaTolmats/IDS-project

Task 2

Identifying business goals

This project aims to gather the concepts, principles and practice of data science learned in the course and to carry out a practical data science project. Our project aims to discover the main factors contributing to student stress through a comprehensive dataset analysis. Our primary objective is to delve into real-life factors such as sleep quality, study load, and environmental influences, to understand how these elements impact student stress.

Business Goals:

1. Understand the key aspects that affect students' stress levels.
2. Identify factors influencing academic performance and mental well-being.
3. The overall goal is to gain a deeper understanding of students' lives.

Business Success Criteria:

1. Increased awareness and understanding of student stress factors.
2. Improved academic performance and mental well-being among students.

Assessing the situation

Inventory of resources:

Our dataset from the Kaggle competitions contains information about 1100 students with around 20 features related to psychological, physiological, environmental, academic and social factors.

Requirements, assumptions and constraints:

The project must be completed by December 14. One requirement is that we have accurate and representative data. The data is gathered through a combination of online and offline surveys. The age group is 15 to 24 and data is gathered in Dharan, Nepal. We assume that the reported factors directly impact student stress.

Risks and contingencies:

We have a risk that there are external factors affecting student stress level that have yet to be captured in the dataset.

Terminology:

In our dataset the features are selected scientifically considering 5 major factors - Psychological, Physiological, Social, Environmental, and Academic Factors.
Psychological Factors are anxiety level, self-esteem, mental health history and depression.
Physiological Factors are headache, blood pressure, sleep quality and breathing problems.
Environmental Factors are noise level, living conditions, safety and basic needs.
Academic Factors are academic performance, study load, teacher-student relationship and future career concerns.
Social Factors are social support, peer pressure, extracurricular activities and bullying.

Features are rated on different scales, the mayor on a scale of 0 to 5.

Costs and benefits:

In our project data analysis is cost-effective and reporting. Benefits include improved student well-being, academic success and potential societal impact.

Defining data-mining goals

Data-mining goals:

1. Identify patterns and correlations between different factors and student stress.
2. Develop predictive models to anticipate stress levels based on key features.
3. Provide actionable recommendations for individuals and institutions.

Data-mining success criteria:

1. Accurate prediction of student stress levels.
2. Identification of modifiable factors for intervention.
3. Clear and interpretable presentation of findings.

Task 3

Gathering data

Outline data requirements

1. **Data source**
    - Kaggle dataset
    - Information from 1100 students
2. **Features**
    - Psychological Factors
        - Anxiety level(GAD-7 Score): 0 to 21
        - Self-Esteem (Rosenberg Scale): 0 to 30
        - Mental Health History: 0 (No history) or 1 (History present)
        - Depression (PHQ-9 Score): 0 to 27
    - Physiological Factors

        - Headache, Blood Pressure, Sleep Quality, Breathing Problems: 0 to 5
    - Environmental factors
        - Noise Level, Living Conditions: 0 to 5
    - Academic Factors
        - Academic Performance, Study Load, Teacher-Student Relationship, Future Career Concerns: 0 to 5
    - Social Factors
        - Social Support, Peer Pressure, Extracurricular Activities, Bullying: 0 to 5
    - Outcome Variable
        - Stress level: 1 to 2
3. **Age Group**
    - 15 to 24 years
4. **Location**
   Dharan, Nepal
5. **Data Collection Method**
   Combination of online and offline surveys
6. **Data Format**
   Diverse formats to capture the richness of students experiences
7. **Time Frame**
   Data gathered up to the project completion deadline (December 14)
8. **Scale Ratings**
    - Features rated on a scale of 0 to 5
    - Interpretation: 0-1 (Low), 2-3(Mid), 4-5(High)
9. **Project Goals Alignment**
    - Address the key aspects affecting student stress
    - Identify factors influencing academic performance and mental well-being
10. **Risk Considerations**
    Acknowledgement of the risk that external factors affecting student stress may not be fully captured in the dataset
11. **Data Exploration Verification**
    - Consistency observed with clear labels and counts for each variable
    - No missing values detected

Verify data availability:

The dataset, sourced from Kaggle competitions, is confirmed to be available in its entirety, comprising information from 1100 students with around 20 features related to psychological, physiological, environmental, academic, and social factors. This comprehensive dataset forms the foundation for our analysis of factors influencing student stress levels.

Define selection criteria

All of the features in the dataset are relevant to our project because we want to find which one affects the most. The inclusivity is based on the premise that each feature potentially contributes to understanding and predicting student stress levels. No feature is excluded at this stage as we aim to identify which factors exert the most significant influence on student stress.

Describing data

The Kaggle dataset contains information that was collected through a combination of online and offline surveys and 1100 students answered. The data encompasses various formats.

The following ranges have been used to describe the extent to which each factor affects a student:

The GAD-7 score has been used for anxiety, range of 0 to 21. GAD-7 score is a seven-item instrument that is used to measure or assess the severity of generalized anxiety disorder (GAD).

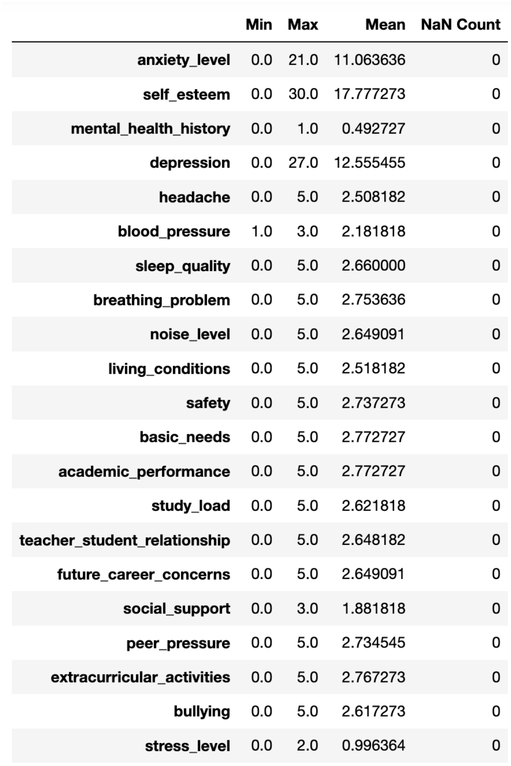
Rosenberg Self Esteem Scale has been used for self-esteem, range 0 to 30. The Rosenberg Self-Esteem Scale (RSES) is a 10-item, Likert scale, self-report measure originally developed to gather information about adolescent feelings of self-esteem and self-worth.

Mental Health History ranges 0 when there is no mental health history and 1 if there are.

Patient Health Questionnaire (PHQ-9) is used to assess depression, range 0 to 27. The PHQ-9 is the depression module, which scores each of the 9 DSM-IV criteria as “0” (not at all) to “3” (nearly every day).

Other features mostly range from 0 to 5 considering 0,1 to be low, 2,3 to be mid, and 4,5 to be high.

Exploring data



	Min	Max	Mean	NaN Count
anxiety_level	0.0	21.0	11.063636	0
self_esteem	0.0	30.0	17.777273	0
mental_health_history	0.0	1.0	0.492727	0
depression	0.0	27.0	12.555455	0
headache	0.0	5.0	2.508182	0
blood_pressure	1.0	3.0	2.181818	0
sleep_quality	0.0	5.0	2.660000	0
breathing_problem	0.0	5.0	2.753636	0
noise_level	0.0	5.0	2.649091	0
living_conditions	0.0	5.0	2.518182	0
safety	0.0	5.0	2.737273	0
basic_needs	0.0	5.0	2.772727	0
academic_performance	0.0	5.0	2.772727	0
study_load	0.0	5.0	2.621818	0
teacher_student_relationship	0.0	5.0	2.648182	0
future_career_concerns	0.0	5.0	2.649091	0
social_support	0.0	3.0	1.881818	0
peer_pressure	0.0	5.0	2.734545	0
extracurricular_activities	0.0	5.0	2.767273	0
bullying	0.0	5.0	2.617273	0
stress_level	0.0	2.0	0.996364	0

Verifying data quality

This dataset seems to be consistent with clear labels and counts for each variable. Also it seems that there are no missing values and it's good enough to support our goals.

Task 4

Task	Lotte	Kirke	Mona	Methods	Comments
Explore the dataset - Identify missing values - Assess data distribution - Clean the dataset by handling missing values, outliers, and inconsistencies	5	4	4	- Pandas for data exploration and cleaning. - Matplotlib and Seaborn for data visualization.	
Features - Select relevant features based on domain knowledge and initial exploration. - Engineer new features that could enhance the analysis.	5	5	6	- NumPy and SciPy for statistical analysis. - Matplotlib and Seaborn for visual representation.	
- Conduct descriptive statistics on key features. - Perform inferential statistics to identify correlations and patterns.	6	5	6	- NumPy and SciPy for statistical analysis. - Matplotlib and Seaborn for visual representation.	
- Investigate correlations	6	3	6	- Correlation analysis	

between anxiety levels, academic performance, and other factors. - Examine patterns to derive insights into the interplay of different stress factors.				using pandas. - Visualization with Matplotlib and Seaborn.	
- Compile a comprehensive report detailing findings, insights, and recommendations. - Prepare a presentation summarizing key results	8	6	8	- Jupyter Notebooks for documentation. - Illustrator for presentation creation.	