[Students' Social Media Addiction](https://www.kaggle.com/datasets/adilshamim8/social-media-addiction-vs-relationships)

**Real Data**

The **Student Social Media & Relationships** dataset contains anonymized records of students’ social‐media behaviors and related life outcomes. It spans multiple countries and academic levels, focusing on key dimensions such as usage intensity, platform preferences, and relationship dynamics.

* **Korrelationer/Prediktioner**: analysera hur användning/beroendegrad påverkar akademiska resultat, relationer, psykiskt mående samt sömnpåverkan.
* **Segmentering av studenter:** baserat på sociala medie-vanor.
* **Anomali-detektion:** för extremt högt eller lågt användande, finns det tröskelvärden som är “farliga” och som vi bör hålla utkik efter?.
* Väldokumenterat.

**Gruppuppgift/Diskussion**

Positivt med datasetet:

* Väldokumenterat
* Global representation
* Tillåter många perspektiv för orsak/verkan.

Tankar:

* Klassificera/Dela upp åldrar ( Grupp 1 ≤ 20, Grupp 2 ≥ 21).
* Identifiera/prediktera tröskelvärden.
* Hur ser plattformspreferenserna ut i olika åldersgrupper?
* Man kan jämföra antal sömntimmar med vetenskapliga rekommendationer för respektive åldersgrupp för att fördjupa analysen.
* Skulle kunna utföra binning på addicted\_score för bättre och tydligare riskanalys. Översätt korrekt utifrån Bergen Social Media Addiction Scale.
* Intressant för EDA att titta på addicted\_score för top-5 länder -> Inferenser

Kolumner som kan vara intressanta ur ett korrelationsperspektiv:

* Avg\_Daily\_Usage\_Hours
* Sleep\_Hours\_Per\_Night
* Mental\_Health\_Score,
* Relationship\_Status
* Conflicts\_Over\_Social\_Media
* Addicted\_Score

Kolumner som förmpdligen inte spelar så stor roll för en ML-modell:

* Student\_ID – Tas bort pga att den enda funktionen är att säkerställa att datasetet ej innehåller duplikat.
* Country – Har tittat på *Frequency Encoding*, men är ej ett alternativ då väldigt många länder innehar samma frekvens, och kan då ej skiljas från varandra. Eftersom den kategoriska kolumnen innehåller nominal data med ,110 unika värden, är det ej rimligt att encoda dessa till unika kolumner för modellen att tränas på.
* Academic\_Level – Kan ordinal-encodas för att granskas nämare (EDA?).
* Most\_Used\_Platform - Samma som Country
* Age – bra att ta bort för att minska risk för överanpassning. Håll modellen simpel, med de högst korrelerande variablerna intakta.

Övriga tankar:

Det kan vara idé att dela upp din dataframe i olika encoding-versioner beroende på syfte:

* Dummy-variable-encoding för linjär regressions-modell
* One-hot-encoding för övriga modeller

Tankar kring målbild:

* Båda har idéer på en Streamlit-applikation där en användare kan fylla i information/svara på frågor och få ut någon typ av resultat från en tränad ML-modell.
* Målbild kan istället vara en binär klassificering huruvida studieresultat riskerar att påverkas eller ej.