Prediktion av riskgrupper för socialt medieberoende bland studenter  
–

Hur sociala medievanor påverkar studenters hälsa, sömn, relationer och studieresultat



David Labett

NBI/Handelsakademin

Kunskapskontroll, AI Teori och Tillämpning, del 1

Höst 2025

# Abstract

# Förkortningar och Begrepp NSF: National Sleep Foundation

BSMAS: Bergen Social Media Addiction Scale

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Innehållsförteckning

[1 Inledning 1](#_Toc204036194)

[1.1 Underrubrik – Exempel 1](#_Toc204036195)

[2 Teori 2](#_Toc204036196)

[2.1 Exempel: Regressionsmodeller 2](#_Toc204036197)

[2.1.1 Exempel: Lasso 2](#_Toc204036198)

[2.1.2 Exempel: Ridge 2](#_Toc204036199)

[2.1.3 Exempel: Elastic Net 2](#_Toc204036200)

[2.2 Exempel: Neurala Nätverk 2](#_Toc204036201)

[3 Metod 3](#_Toc204036202)

[4 Resultat och Diskussion 4](#_Toc204036203)

[5 Slutsatser 5](#_Toc204036204)

[6 Teoretiska frågor 6](#_Toc204036205)

[7 Självutvärdering 7](#_Toc204036206)

[Appendix A 8](#_Toc204036207)

[Källförteckning 9](#_Toc204036208)

# Inledning

Intåget-, och influenserna av sociala medier i gemene mans liv är påtagligt. Vi finner en särskilt hög konsumtion i åldrarna 16 till 24. Enligt undersökningar i [en samlad global rapport](https://datareportal.com/reports/digital-2025-global-overview-report) hos Datareportal visar GWIs siffror att ”..on average around the world, people aged 16 to 24 spend more than three times longer using social media each day than people aged 65 and above do.”

Även 2025 är Sociala medier en snabbt växande industri – Kepios uppmäter att “..there are now **5.24 billion** active social media user identities around the world, with that total increasing by **4.1 percent** over the past 12 months”. Enligt rapporterade siffror innebär alltså en ökning på **206 miljoner** användaridentiteter bara det senaste året. Antalet användaridentiteter uppmäts till **63.9%** av den totala globala befolkningssiffran.

Detta väcker frågor kring hur den stigande adoptionen och användandet av sociala medier påverkar våra liv och relationerna till de runt omkring oss. Hur ser den dagliga användningen ut hos den åldersgrupp (16 till 24) som konsumerar mest? Hur präglar och påverkar dessa digitala vanor andra områden i personernas liv; exempelvis studieresultat, relationer, sömn och psykisk hälsa? Kan vi hitta någon form av tröskelvärde som kan ge indikation på sunt, kontra problematiskt, beteendemönster?

Denna rapport analyserar datasetet ***Student’s Social Media Addiction*** *– A Cross-Country Survey of Usage Patterns, Academic Impact, and Relationship* för att undersöka hur olika delar av studenters liv korrelerar med-, och påverkas av deras sociala medievanor.

Vidare tränas en Linjär Regressionsmodell som läses in i en Streamlit-applikation där användare kan besvara liknande, självskattade frågor som för informationen i datasetet. Målbilden är att användaren dels får reflektera över sina egna vanor genom frågeformuläret, men att hen även får en indikation om sin ”Addicted Score” baserat på svaren; predikterat av en tränad Maskininlärnings-modell. Till sist visualiseras även användarens egna svar i relation till datasetets medianer för berörda/relevanta features.

Ge en bred bakgrund till ditt arbete och varför det är relevant. Smalna successivt av och koppla bakgrunden till detta arbete. Fortsätt med syfte och frågeställning (syfte och frågeställning kan du ha i en egen underrubrik om du önskar det).

Syftet med denna rapport är \_\_\_\_\_\_\_\_\_\_\_\_, för att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. \_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_

# Teori

## Exploratorisk Data-Analys (EDA)

Datasetet *Student’s Social Media Addiction* innehåller i sitt original-skick 705 rader, och 13 kolumner; varav 7st är numeriska och 6st kategoriska.

Bland de kategoriska kolumnerna sticker *Country* samt *Most\_Used\_Platform* ut i antal unika värden. De innehåller 110- respektive 12 unika värden. Under rapportens EDA är dessa värden intressanta – men under bearbetningen av datan inför träning av makininlärningsmodeller behöver man ta ställning till hanterandet av dessa. Varje kategorisk kolumn med *nominal data* behöver hanteras via *dummy-variable-encoding* när en Linjär Regressions-modell ska tränas. För övrig modeller lämpar sig *one-hot-encoding* bättre (Prgomet et al. 2025).

### Begränsningar

### Alla mätningar i datasetet är självrapporterade och självskattade av varje enskild person. Brist på självinsikt och beroendebild kan bidra till icke sanningsenliga svar.

### Respondenterna skulle även kunna vara påverkade av social önskvärdhetsbias genom att de känner sig pressade att förneka beroendeproblematik, eller rationalisera den.

### Förståelse för varje kolumn och vad den representerar

* **Student\_ID**: Unik för varje student - vilket säkerställer att inga duplikat finns.
* **Age**: Åldersspannet för studien är mellan 18-24 år.
* **Gender**: Kategorisk kolumn med värdena *Male* eller *Female*.
* **Academic\_Level**: Kategorisk kolumn med personens högsta aktuella akademiska nivå (gäller även pågående).
* **Country**: Landet personen bor i vid svarstillfället.
* **Avg\_Daily\_Usage\_Hours**: Det genomsnittliga antal timmar per dag personen spenderar på sociala medie-plattformar.
* **Most\_Used\_Platform**: Kategorisk kolumn som innehåller värdet med den sociala medie-plattform personen mest frekvent använder.
* **Affects\_Academic\_Performance**: Kategorisk kolumn med värdena *Yes* och *No* baserat på frågan *”..whether the student perceives their social media use as having a negative impact on their academic performance”*.
* **Sleep\_Hours\_Per\_Night**: Det genomsnittliga antalet timmar som personen sover per natt.
* **Mental\_Health\_Score**: Skala mellan 1 = Dålig till 10 = Mycket bra, där personen själv skattar sin generella psykiska hälsa.
* **Relationship\_Status**: Kategorisk kolumn som innehåller *Single, In Relationship* och *Complicated*
* **Conflicts\_Over\_Social\_Media:** Antal konflikter eller meningsskiljaktigheter med familj, vänner eller partner över personens sociala medie-vanor.
* **Addicted\_Score**: Standardiserad skala där 1 = Låg till 10 = Hög. Baserad på [*Bergen Social Media Addiction Scale*](https://hub.salford.ac.uk/psytech/2021/08/10/bergen-social-media-addiction-scale/).

### 

### Fördelning av data

Genom att undersöka fördelningen av data kan vi identifiera eventuella bias, outliers eller brister.

* Könsfördelningen är nästan helt jämn i datasetet.
* De flesta respondenterna är *Undergraduate* eller *Graduate* - medan *High School* är underrepresenterad. Detta speglar också åldersfördelningen i datasetet.
* Instagram är den överlägset mest populära plattformen hos respondenterna. Följt av TikTok, Facebook, WhatsApp och Twitter osv.
* En stor andel av studenterna anser att deras Sociala Medie-vanor påverkar deras akademiska resultat.
* Relationsstatusen som var mest frekvent hos respondenterna var *Single*.
* Vi ser ett åldersspann mellan **18 och 24 år**.
* Medianåldern är **21 år**.
* Fördelningen av daglig användning mellan könen är tämligen likvärdig, men visar sig vara lite större hos kvinnor.

### Inferenser

* En jämn könsfördelning minskar risken för bias kön i analysen.
* Den höga andelen singlar kan spegla åldersgruppen - men skulle även kunna relateras till sociala mediers roll i relationer och socialt liv. (Det är svårt att undersöka detta vidare med den befintliga datan, därför stannar det vid hypotes-stadiet).
* Indien, USA, Canada är i toppen av länderna som representeras. Resursrika länder som dessa, där teknik är lättilgänglig för gemene man, skapar förutsättningar för att man också utsätts för mer/tyngre bruk av sociala medier - och i sin tur dess eventuella konsekvenser.
* Instagram och TikToks överlägsna popularitet skulle kunna bero på deras *Reel*-format som är konstruerat för *instant gratification* och dopamin-påslag som fångar användaren i ett oändligt scrollande.

### Korrelationer

*Addicted\_Score* är vår beroende variabel **y** – och vi vill undersöka hur de oberoende variablera **X** - f*eatures* - korrelerar med-, och påverkar **y**.

Nedan kan vi se en visualisering av korrelationer mellan våra numeriska *features och* ***y***.

A chart of different colors

AI-generated content may be incorrect.

I samband med hög *Addicted\_Score* sker en positiv korrelation, alltså en ökning, mellan *Avg\_Daily\_Usage\_Hours* och *Conflicts\_Over\_Social\_Media*.

Vid hög *Addicted\_Score* sker en negativ correlation, alltså en minskning, mellan *Sleep\_Hours\_Per\_Night* och *Mental\_Health\_Score*.

Vi kan teoretisera kring hur dessa korrelerande värden tillsammans bidrar till en negativ spiral som kan vara svår att bryta.

### Medianer

Genom att granska median-värdet för de största korrelationerna till **y** kan vi få en generell indikation på hur studenters vanor ser ut.

A group of graphs showing different sizes of graphs

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

Medianvärdet för *Avg\_Daily\_Usage\_Hours* är **4.92h/dag** vilket indikerar att sociala medier utgör en stor del av studenters dagliga vanor:

* Vidare kan detta möjligen korrelera med den stora andelen som känner påverkan på sina akademiska resultat då en significant del av potentiell studietid går förlorad.
* Datasetet belyser en generation och population med hög risk att utveckla problematiska- vanor och/eller beroende av sociala medier.

Samtidigt är medianvärdet för *Sleep\_Hours\_Per\_Night* **6.87h/natt** vilket visar på ett relativt högt antal sömntimmar.

Medianen för *Conflicts\_Over\_Social\_Media* är **2.85 Konflikter**. Detta tyder på att social friction kopplad till sociala medievanor är ganska vanligt förekommande för studenterna.

A graph of a graph with a line

AI-generated content may be incorrect.

*Addicted\_Score* är, likt resterande kolumner med skalor, mellan *1 - 10*. Ett medianvärde på **6.44/10** tyder på en konsekvens av ovanstående observationer.

Ett högt medianvärde över det största korrelerande spannet av variabler stärker bilden av att sociala medier är en central, och ibland problematisk del av den aktuella åldersgruppens liv.

### Sömn i relation till vetenskap

I ovanstående diagram har vi tittat närmare på medianvärdet för *Sleep\_Hours\_Per\_Night*. För att fördjupa förståelsen i analysen ställer jag detta värde i relation till [National Sleep Foundations vetenskapliga rekommendationer](https://www.sleephealthjournal.org/article/s2352-7218(15)00015-7/fulltext) för åldrarna **18-25**.

A graph of a sleep number

AI-generated content may be incorrect.

Vi ser alltså att medianvärdet i datasetet ligger utanför, och under, det minsta vetenskapligt rekommenderade värdet i det gröna spannet.

## Preprocessering av data under EDA

Under EDAn behölls samtliga kolumner i datasetet förutom *Student\_ID* – vars enda syfte är att säkerställa att inga duplikat finns i datan.

Vidare utfördes en uppdelning av numeriska- respektive kategoriska kolumner för vidare korrelationsanalys.

### Preprocessering & Feature Engineering inför modellträning

Under bearbetningen av datasetet inför modellträning läses datasetet in på nytt.

Eftersom jag tränar både en modell som använder linjär regression, och en som nyttjar klassificering krävs olika preprocesseringsrutiner för vardera modell. Förgreningen av vår dataframe sker via *.copy()* för att dessa inte ska påverka varandra.

Inför båda modellernas träning utförs en typ av *dimensionsreducering* via *.drop()*-metoden på följande kolumner, med motivering:

* *Student\_ID* – Syftet är detsamma som för EDAn, vilket nämns ovan.
* *Country* – En kolumn med nominal data som innehåller 110 unika värden. *Frequency Encoding* har undersökts, men inte visat sig optimal då många länder innehar samma frekvens, och kan därför inte särskiljas. Encoding utesluts då detta hade inneburit adderandet av alldeles för många unika kolumner – vilket i sin tur inte är rimligt.
* *Most\_Used\_Platform* – Ytterligare en kolumn med nominal data som lider av för många unika värden, och för låg korrelation med **y**, för att encodas.
* *Academic\_Level* – En kategorisk kolumn som skulle kunna ordinal-encodas. Dock har kolumnen indirekt samma värden som *Age*, fast i annat format.

Vad skulle kunna hända om vi istället behåller de features med stora antal unika värden?

Resultatet hade lett till *Curse of dimensionality; Ju fler variabler/features i ett dataset, desto högre dimension* (Prgomet et al. 2025). Högre dimensioner kräver större mängd data för att skapa bra modeller. Dessutom kan, i högre dimensioner, datan bete sig annorlunda.

Vidare hade vi kunnat hamna i en överanpassad modell som lär sig unika fall/länder, och som då får svårigheter att generalisera ny data.

### Preprocessering av data inför Linear Regression

För den linjära regressionsmodellen utför vi specifikt *dummy-variable-encoding*, och inte *one-hot-encoding* (som för vår kommande klassificeringsmodell). Båda tillvägagångssätten ger samma informationsinnehåll, men *..när vi ska använda nominal data i den linjära regressions-modellen så ska vi använda dummy-variable-encoding medan vi för övriga modeller använder one-hot-encoding* (Prgomet et al. 2025).

*Dummy-variable-encoding* (likväl som *one-hot-encoding*) används för att omvandla kategorisk data till numerisk data. Detta är nödvändigt för att våra ML-modeller ska kunna tränas, fungera och prediktera åt oss. Vi skiljer på *nominal* och *ordinal* data – där det förstnämnda innebär värden som saknar inbördes rangordning, medan det sistnämnda har en tydlig rangordning. För vår linjära regressionsmodell finns enbart nominal data. Därför utför vi *dummy-variable-encoding* på samtliga kategoriska kolumner, vilket ger oss resultatet *n – 1 kolumner*, och utesluter en *referenskategori* för varje kolumn som omvandlas (s.42 – 45, Prgomet et al. 2025.

### Preprocessering av data inför Random Forest Classifier

Processen för klassificeringsmodellen innebär lite mer arbete och ställningstaganden.

Målet är att prediktera en riskgrupp av beroendegraderna; *Low, Medium* och *High*. Precis som för Linear Regression behöver modellen numeriska värden för att tränas och fungera korrekt. För beroendegraderna uppnås detta genom *ordinal-encoding* eftersom det finns en tydlig rangordning bland dem (s. 45, Prgomet et al. 2025). Uppdelningen baserat på beroendegrad kommer att förenkla tolkningen av modellens resultat, jämförelse mellan grupper samt identifiering av just riskgrupper.

Vid själva uppdelningen - *binning* - är det viktigt att ta hänsyn till ett par olika faktorer:

* Fördelningen av datan i beroendegraderna behöver vara i interpreterbar storlek för att klassificeringsmodellen ska kunna tränas med gott prediktionsresultat.
* *Addiction\_Score* är en 10-gradig skala, men är baserad på [BSMAS](-%09https:/hub.salford.ac.uk/psytech/2021/08/10/bergen-social-media-addiction-scale/) i skalan 6 till 30. Eftersom målbilden är en *Streamlit*-applikation där BSMAS-skalan behöver transformeras till skalan för *Addicted\_Score* bör följande formel användas:

Beroendegradernas uppdelning sker i såkallad *Binning* – där ovanstående faktorer beaktas.

Efter att skaltranformationen utförts kan vi med ovanstående formel beräkna det avgörande *gränsvärde* som BSMAS-skalan har för sin bedömda hög-riskgrupp.

Gränsvärdet för hög-riskgrupp i BSMAS-skalan är **24**, vilket översätts till **7.75** efter transformering. Värdet avrundas uppåt till **7.8** - här skulle vi kunna spela lite med vår avrundning beroende på hur strikt gränsvärde vi vill ha för vår risk-prediktion. Vidare använder vi det avrundade värdet när vi delar upp våra grupper, eller *bins*, programmatiskt: bins = [1, 4.5, **7.8**, 10].

Hanteringen av övriga kategoriska variabler sker via *one-hot-encoding*. Till skillnad från *dummy-variable-encoding* lämnar denna metod alla kolumner *intakta*, och använder sig inte av ett *n-1 kolumn*-förhållande.

## Utvärderingsmått: RMSE

*RMSE är det mått som används i störst utsträckning i samband med att regressionsmodeller utvärderas* (s. 106, Prgomet et al. 2025). *Root Mean Squared Error* är ett mått på hur mycket en modells prediktioner avviker från de faktiska värdena, och beräknas med följande formel:

I sig ger dock inte *RMSE* oss hela bilden av hur *bra* en modell presterar. För att få förståelse för vårt *RMSE*-värde behöver vi skapa oss en referenspunkt – nämligen medelvärdet för **y** (våra faktiska värden).

Ekvationen nedan tar vår *RMSE* dividerat med medelvärdet för våra faktiska värden:

På så vis skapar vi oss en tydligare, *procentuell* bild för hur stort modellens prediktionsfel, *RMSE*, är i förhållande till medelvärdet för **y** (s. 82, Prgomet et al. 2025).

## 2.3.1 Utvärderingsmått: Confusion Matrix

*Confusion Matrix* (*Fig. 1)* används ofta vid utvärdering av klassificeringsmodeller. Detär ett verktyg som visualiserar en jämförelse av korrekta- eller felaktigt klassificerade i varje kategori. Det diagonala resultatet (från övre vänstra hörnet till nedre högra) visar samtliga korrekt klassificerade prediktioner (s. 155, Prgomet et al. 2025).

Eftersom vi har ett multiklass klassificeringsproblem skalas vår *confusion matrix* till antalet befintliga klasser – i vårt fall *Low, Medium* och *High.*

A diagram of positive negative

AI-generated content may be incorrect.

Fig 1. Schematisk representation av en Confusion Matrix

## 2.3.2 Utvärderingsmått: Classification Report

En *Classification Report* innehåller en samling av flera utvärderingsmått; *Accuracy, Precision, Recall* samt *F1-score*:

* ***Accuracy*** utgör den andel korrekta prediktioner av samtliga observationer.
* ***Precision*** avser de prediktioner som modellen klassificerar som positiva, som *faktiskt är positiva* (*TP*).
* ***Recall*** innefattar den andel *faktiskt positiva* (*TP*) som modellen hittar.
* ***F1-score*** är ett samlat, balanserat medelvärde för *precision* och *recall* enligt formeln:

(s. 162, Prgomet et al. 2025)

## Maskininlärningsmodell: LinearRegression

## 2.4.1 GridSearch

## Maskininlärningsmodell: RandomForestClassifier

## 2.5.1 Gini

## K-delad korsvalidering: Linjär Regression vs Klassificering

# Metod

Hur har du genomfört ditt arbete? Exempelvis, hur har datan erhållits?

# Resultat och Diskussion

|  |  |
| --- | --- |
| **RMSE för olika modeller** | |
| Enkel Linjär Regression | xx |
| Lasso | xx |
| Ridge | xx |

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

Figur 2: Hur man lägger in tabell eller figur nummer samt beskrivning.

# Slutsatser

Här besvarar du bl.a. frågeställningarna.

# Teoretiska frågor

# Självutvärdering

1. Vad har varit roligast i kunskapskontrollen?
2. Vilket betyg anser du att du ska ha och varför?
3. Vad har varit mest utmanande i arbetet och hur har du hanterat det?
4. Hur har grupparbetet gått?

# Appendix A

Om inget inkluderas här, ta bort avsnittet. Här kan man t.ex. lägga en GitHub länk till arbetet.

# Källförteckning

* <https://datareportal.com/reports/digital-2025-global-overview-report>

* <https://hub.salford.ac.uk/psytech/2021/08/10/bergen-social-media-addiction-scale/>
* Lär dig AI från grunden – Tillämpad maskininlärning med Python
* <https://www.sleephealthjournal.org/article/s2352-7218(15)00015-7/fulltext>