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

Arbetet undersöker sambandet mellan studenters användning av sociala medier och deras risk för att utveckla ett beroende. Särskild fokus har lagts på faktorer som daglig användning, sömn, mental hälsa, upplevd påverkan av studieresultat samt antal konflikter kopplade till sociala medievanor.

Den exploratoriska data-analysen är gjord på ett relevant och kontemporärt dataset. Tydliga mönster indikerar att högre daglig användning av sociala medier korrelerar med självskattade konsekvenser i form av ett högre beroende, sämre mental hälsa, sämre sömn, negativ påverkan på studieresultat samt ett högre antal konflikter kopplade till dessa medie-vanor.

Genom träning och utvärdering av maskininlärningsmodellerna *RandomForestClassifier* samt *Support Vector Classifier* har en prediktionsmodell utvecklats som med hög träffsäkerhet kan klassifiera studenters risknivå för socialt medieberoende. Resultaten visar att ovan nämnda faktorer är värdefulla verktyg för prediktion av beroendegrad.

Arbetets syfte är att belysa förståelse, och att förhoppningsvis kunna förebygga socialt medie-användande som får problematiska konsekvenser för unga vuxna och deras närstående. Vidare ges förslag på hur prediktionsmodeller kan användas i pedagogiskt syfte för tidig identifiering och självreflektion i form av en Streamlit-applikation.

# Förkortningar och Begrepp NSF: National Sleep Foundation

BSMAS: Bergen Social Media Addiction Scale

Table of contents

[Abstract 2](#_Toc209293608)

[Förkortningar och Begrepp NSF: National Sleep Foundation 3](#_Toc209293609)

[1 Inledning 1](#_Toc209293610)

[2 Teori 2](#_Toc209293611)

[2.1 Exploratorisk Data-Analys (EDA) 2](#_Toc209293612)

[2.1.1 Begränsningar 2](#_Toc209293613)

[2.1.2 Förståelse för varje kolumn och vad den representerar 2](#_Toc209293614)

[2.1.3 Fördelning av data 3](#_Toc209293615)

[2.1.4 Inferenser 3](#_Toc209293616)

[2.1.5 Korrelationer 4](#_Toc209293617)

[2.1.6 Medelvärden 5](#_Toc209293618)

[2.1.7 Sömn i relation till vetenskap 7](#_Toc209293619)

[2.2 Preprocessering av data under EDA 7](#_Toc209293620)

[2.3 Preprocessering och feature engineering inför modellträning 7](#_Toc209293621)

[2.4 Utvärderingsmått: Confusion Matrix 9](#_Toc209293622)

[2.4.1 Utvärderingsmått: Classification Report 9](#_Toc209293623)

[2.5 Klassificeringsmodeller 10](#_Toc209293624)

[2.5.1 Maskininlärningsmodell: RandomForestClassifier 10](#_Toc209293625)

[2.5.2 Gini 10](#_Toc209293626)

[2.5.3 Maskininlärningsmodell: Support Vector Classifier (SVC) 10](#_Toc209293627)

[2.6 K-delad Korsvalidering 11](#_Toc209293628)

[2.7 GridSearchCV 11](#_Toc209293629)

[3 Metod 12](#_Toc209293630)

[3.1 Insamling av data samt EDA 12](#_Toc209293631)

[3.2 Preprocessering av data inför modelträning 12](#_Toc209293632)

[3.3 RandomForestClassifier vs Support Vector Classifier 12](#_Toc209293633)

[3.4 Streamlit-applikation 13](#_Toc209293634)

[4 Resultat och Diskussion 14](#_Toc209293635)

[5 Slutsatser 16](#_Toc209293636)

[6 Självutvärdering 17](#_Toc209293637)

[Appendix A 18](#_Toc209293638)

[Källförteckning 19](#_Toc209293639)

# Inledning

Intåget-, och influenserna av sociala medier i gemene mans liv är påtagligt. Vi finner en särskilt hög konsumtion hos unga, i åldrarna 16 till 24. Enligt undersökningar i [en samlad global rapport](https://datareportal.com/reports/digital-2025-global-overview-report) hos *Datareportal* visar *GWI*s 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 år 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, välmående och relationerna till de runt omkring oss. Särskilt intressant är att undersöka hur den mest aktiva åldersgruppen präglas och påverkas i aspekter som studieresultat, relationer, sömn och psykisk hälsa.

Syftet med denna rapport är att analysera datasetet [***Student’s Social Media Addiction***](https://www.kaggle.com/datasets/adilshamim8/social-media-addiction-vs-relationships) *– 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 klassificeringsmodell som läses in i en *Streamlit*-applikation där användare besvarar liknande, självskattade frågor som för informationen i datasetet.

Förhoppningen är att *användaren* dels får reflektera över sina egna vanor genom frågeformuläret, men att hen även får en indikation om sin *Addicted\_*Grade - predikterad av en tränad maskininlärningsmodell.

Slutligen visualiseras *användarens* egna svar i relation till datasetets medelvärden för berörda/relevanta features, samt vetenskapliga sömnrekommendationer. Målet är att skapa ytterligare självreflektion genom att ge en ge *användaren* en visuell referenspunkt för hur dennes vanor ser ut jämfört med snittet.

Frågeställningar:

* Hur ser den dagliga användningen av sociala medier ut bland studenter?
* Vilka samband finns mellan sociala medievanor och faktorer som studieresultat, relationer, sömn och psykisk hälsa?
* Går det att hitta någon form av gränsvärde som signalerar övergången mellan sunt till problematiskt beteendemönster?
* Kan vi med hjälp av maskininlärningsmodeller förutsäga när en användare tillhör en riskgrupp och därigenom öka självinsikten om möjliga konsekvenser av deras sociala medievanor?

# 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 *one-hot-encoding*, då det lämpar sig bättre för klassificeringsmodeller än *dummy-variable-encoding.* Kategoriska kolumner som innehåller *ordinal data* hanteras via *ordinal-encoding* (s. 45, 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.

Det finns en risk för att en negativ spiral uppstår där *Mental\_Health\_Score*, *Sleep\_Hours\_Per\_Night*, *Avg\_Daily\_Usage\_Hours* samt *Conflicts\_Over\_Social\_Media* förstärker varandra. Eftersom alla dessa variabler påverkar varandra försvårar det bedömningen av orsak och verkan.

### Förståelse för varje kolumn och vad den representerar

* **Student\_ID**: Unik för varje student - vilket säkerställer att inga duplikat finns.
* **Age**: Åldersspannet i datasetet 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**.
* Medelåldern 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 könsbias 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.
* Kategorin *Academic\_Level* kan vara intressant att titta närmare på då lägre utbildning brukar upplevas som lättare, och *vice versa*. Kan en låg utbildningsnivå, dvs låg stimulans, bidra till högre *Avg\_Daily\_Usage\_Hours?*

### 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

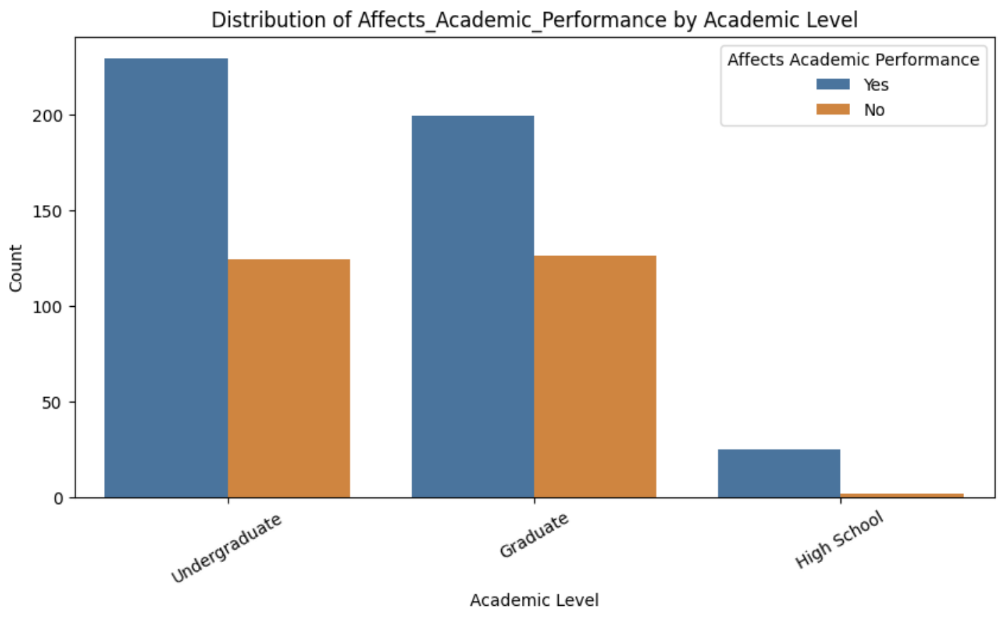
AI-generated content may be incorrect.

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.

| **Age** | **No (%)** | **Yes (%)** |
| --- | --- | --- |
| 18 | 14.29 | 85.71 |
| 19 | 31.29 | 68.71 |
| 20 | 33.33 | 66.67 |
| 21 | 29.49 | 70.51 |
| 22 | 44.90 | 55.10 |
| 23 | 58.82 | 41.18 |
| 24 | 46.15 | 53.85 |

  
Stapeldiagrammet och tabellen ovan visar andelen studenter som upplever störst negativ påverkan på sina studieresultat som konsekvens av sociala medievanor. Högst påverkan på studieresultat hittar vi hos **18-21**-åringar, där mellan 66-85% upplever ett negativt samband (beroende på ålder). Dessa åldrar befinner sig i *High School/Undergraduate* i *Academic\_Level*.

Datan tyder på att yngre studenter, som befinner sig i början av sina högre studier, är mer utsatta för negativ studiepåverkan. Vi kan teoretisera kring hur utbildningsnivåns svårighet och krav ökar gradvis i takt med varandra. I det lägre skiktet, High-School, utmanas elever inte lika mycket vilket ger utrymme för distraktioner som sociala medievanor att gro och påverka. I början av dessa högre studier har man ofta inte heller utvecklat gedigna studietekniker eller självdisciplin. Högre utbildningsnivåer kräver mer struktur och ansvar, vilket speglas något i datan.

Även om både *Age* och *Academic\_Level* korrelerar med *Affects\_Academic\_Performance* görs en avvägning där enbart *Age* behålls. Denna kolumn ger en mer kontinuerlig beskrivning av det beskrivna sambandet, och är en mer universellt tydlig. Det även ger mer precision, då den senare variabeln innehåller breda kategorier som överlappar åldrar.

### Medelvärden

Genom att granska medelvärdet 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.

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

Medelvärdet för *Conflicts\_Over\_Social\_Media* är **2.85 Konflikter**. Detta tyder på att social friktion 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 medelvärde på **6.44/10** tyder på en konsekvens av ovanstående observationer.

Ett högt medelvärde 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 medelvärdet 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 medelvärdet 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.

## 2.3 Preprocessering och feature engineering inför modellträning

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

Förgreningen av vår dataframe sker via *.copy()* för att dessa inte ska påverka varandra.

Inför 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 inte är rimligt.
* *Most\_Used\_Platform* – Ytterligare en kolumn med nominal data som lider av för många unika värden, 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. I min analys anser jag att kolumnen är intressant ur ett EDA-perspektiv, men väljer att behålla *Age* istället.

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 än förväntat.

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 för klassificeringsmodeller innebar lite mer arbete och ställningstaganden än för linjär regression.

*One-hot-encoding* (likväl som *dummy-variable-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.

Målet är att prediktera en riskgrupp av beroendegraderna; *Low, Medium* och *High*. Som nämnt tidigare behöver modellen numeriska värden för att tränas och fungera korrekt. För beroende-graderna 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* som lämpar sig bättre för klassifieringsmodeller. Till skillnad från *dummy-variable-encoding* lämnar denna metod alla kolumner *intakta*, och använder sig inte av ett *n-1 kolumn*-förhållande, (s.42 – 45, Prgomet et al. 2025).

## 2.4 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 . Schematisk representation av en Confusion Matrix

## 2.4.1 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*.

(s. 162, Prgomet et al. 2025)

## 2.5 Klassificeringsmodeller

*Klassificering* är en av fyra olika *problemkatgegorier* inom maskininlärning. Istället för att prekiktera kontinuerlig utdata som för *regression*, handlar målet istället om att prediktera *kategorisk* utdata med hjälp av indata (s. 151, Prgomet et al. 2025). För detta ändamål har jag valt att jämföra två olika klassificeringsmodeller: *RandomForestClassifier* och *Support Vector Classifier* (som är en *Support Vector Machine*).

Dessa två modeller har två olika angreppssätt. Genom att jämföra dem kan jag utvärdera om sambanden i datan är huvudsakligen linjära, där *Support Vector Classifier* presterar bäst, eller om det förekommer fler icke-linjära gränser som separerar klasserna - där *RandomForestClassifier* har fördel.

## 2.5.1 Maskininlärningsmodell: RandomForestClassifier

*RandomForestClassifier* är en optimerad implementation för *Random forest*-modeller, och fungerar genom ett såkallat *Beslutsträd*, (s. 204, Prgomet et al. 2025).

Modellen tillhör familjerna *ensemble* och *beslutsträd*. Dessa innefattar modeller som kan användas både för klassificerings- och regressionsproblem.

Ett *beslutsträd* byggs av data som delas upp i grenar av *noder*. Uppdelningen baseras på *tröskelvillkor* för de variabler som bäst separerar de olika klasserna. I toppen av trädet utgår vi från *rotnoden* – och stegar vidare nedåt genom *inre noder* för att slutligen nå en *lövnod*; en predikterad klass. Varje steg modellen går i trädet är baserat på den variabel och det tröskelvärde som gör att *MSE* blir så lågt som möjligt (s. 204, s.139-140, Prgomet et al. 2025).

*RandomForestClassifier* använder sig av en *ensemble*-metod; där flera modeller (i detta fall *beslutsträd*) kombineras för att uppnå bättre prediktionsförmåga (s. 190, 195-196, Prgomet et al. 2025)

## 2.5.2 Gini

*Gini* är ett index som mäter hur *oren* en nod i ett beslutsträd är. Lägre värde betyder renare nod. Ju längre ner vi stegar i ett beslutsträd, desto mindre blir värdet på *gini-koefficienten*, ända tills vi når vår predikterade klass/*lövnod* (s. 194-195, Prgomet et al. 2025).

Formel för *Gini*: 𝐺ᵢ = 1 − ∑ 𝑝ᵢ²

## 2.5.3 Maskininlärningsmodell: Support Vector Classifier (SVC)

*Support Vector Classifier* (*SVC*) är en klassificeringsmetod för *Support Vector Machines* (*SVM*). För regressionsproblem har SVM-modeller som mål att hitta en *väg* där så många datapunkter som möjligt från datasetet hamnar. För klassificeringsproblem försöker modell istället urskilja och separera datapunkterna som tillhör de olika klasserna med så stor marginal som möjligt. Därigenom hittas en optimal gräns som separerar klasserna i datan.

I de fall vi vill utvärdera om en linjär eller icke-linjär modell är bättre kan *SVC* användas där olika kernels - *linear* eller *poly* utvärderas (s. 190, Prgomet et al. 2025).

Valet av *SVC* är alltså för att modellen kan hantera både linjära och icke-linjära separationer i datan genom att använda olika *kernels* som *hyperparametrar*. *Linear* kernel försöker hitta en rak linje som separerar klasserna i datan. Medan *poly* kernel skapar mer komplexa relationer genom polynomer ([GeeksForGeeks](https://www.geeksforgeeks.org/machine-learning/major-kernel-functions-in-support-vector-machine-svm/)). Därav kan båda hyperparametrarna med fördel utforskas via *GridSearchCV*-optimering.

När vi använder oss av en *SVM*-modell är det god idé att standardisera vår data, då dessa modeller är känsliga för skala (s. 183-184, Prgomet et al. 2025). Detta utförs med hjälp av *StandardScaler()* placerad i en *pipeline*.

## 2.6 K-delad Korsvalidering

*K-delad korsvalidering* innebär att datan delas upp i tränings-, respektive testdata. Träningsdatan delas därefter upp i *k* lika stora delar, där modellen tränas (itererar) *k* gånger på *k-1* delar. Den återstående delen är valideringsdata för varje iteration, (s. 30-31, Prgomet et al. 2025).

I varje iteration utförs en validering och ett värde tilldelas utifrån det utvärderingsmått som används. Slutligen tas medelvärdet av alla utvärderingsmått; resultatet blir en siffra som bedömer hur bra modellen presterat. Motivering bakom användningen av denna metod är att datasetet är förhållandevis litet. Metoden lämpar sig väl när datamängden är begränsad, och ger då en robust bedömning.

När vi tittar närmre på fördelningen av data för våra *Addicted\_Grade*-klasser ser vi att den är ganska obalanserad. För att förbättra modellträningen ytterligare används *stratifierad* korsvalidering - [*StratifiedKFold*](https://scikit-learn.org/stable/modules/cross_validation.html). I praktiken motverkar detta obalansen mellan klasserna genom att försäkra att varje *fold* behåller samma klass-distribution som orginaldatan. Alltså förebygger detta *samples* som riskerar att sakna data från en minoritets-klass.

## 2.7 GridSearchCV

Vi kan finjustera våra inställningar för respektive modell via såkallade *hyperparametrar*. *GridSearchCV* är ett verktyg från *Scikit-learn* vars algoritm systematiskt tillämpar *K-delad korsvalidering* för varje unik kombination av hyperparametrar. Den optimala kombinationen väljs automatiskt av *GridSearchCV*, (s. 38, Prgomet et al. 2025).

# Metod

I detta avsnitt sammanfattas tillvägagångssättet för att genomföra analys, träning av maskininlärningsmodeller samt utvecklingen av en Streamlit-applikation som huserar modellen.

Arbetet har följt en systematisk process likt den vi följt repetitivt under kursens gång: från datainsamling, EDA, cleaning och feature engineering till modellval och utvärdering.

## Insamling av data samt EDA

* Nedladdning och inläsning av dataset.
* Övergripande granskning av data i form av struktur, fördelning, outliers och korrelationer för att bygga en informerad bild.
* Identifiering av *features* och *target*.
* Beslutstagande kring *feature engineering* där *Student\_ID* avlägsnades inför EDA.
* Djupare analys.

## Preprocessering av data inför modelträning

1. *Feature engineering* där *Country, Most\_Used\_Platform, Age, Gender, Student\_ID, Academic\_Level* avlägsnades.
2. *Binning* av *Addicted\_Score* till *Low, Medium, High* - kalkylerat med hänsyn till BSMAS.
3. Uppdelning av features *X* och target *y* som ska predikteras.
4. Train-Test-Split med Stratifiering. Datan delas upp i ett tränings- och ett testset med fördelningen 80/20. Detta utförs med stratifiering, vilket resulterar i en balanserad klass-distribution i varje fold under den kommande K-delade korsvalideringen.

## RandomForestClassifier vs Support Vector Classifier

1. Initiering av modeller RandomForestClassifier samt Support Vector Classifier.
2. Pipelines skapas för respektive modell för en strömlinjeformad och effektiviserad träning. StandardScaler initieras (så att alla features får samma skala) och adderas i pipeline för SVC då SVM-baserade modeller är känsliga. RandomForesClassifier är mindre känslig för skalning, men pipeline används ändå för enhetlighet.
3. Hyperparametrar samlas I ett grid för respektive modell inför testning och optimering med GridSearchCV. SVC kan hantera både linjära och icke-linjära beslutsgränser – hyperparametrar för kernel *linear* samt *poly* används för utvärdering. För RandomForestClassifier testas *n\_estimators* (antal träd) och *max\_depth*.
4. K-delad Korsvalidering används 5-faldigt. Datan delas alltså upp i fem delar, och testas iterativt. Detta minskar risken för slumpmässiga resultat som beror på en viss, tur- eller otursam uppdelning av datan.
5. Utvärdering och jämförelse av modeller utförs på testdatan efter träning och hyperparameteroptimisering har utförts. Resultaten presenteras genom *Accuracy, Classification Report* samt *Confusion Matrix*. Detta samlade underlag ger en tydlig bild av modellernas prestanda.
6. Val och export via *joblib* av bäst presterande modell för prediktioner i Streamlit-app.

## Streamlit-applikation

Streamlit-applikationens syfte är att ge användare en möjlighet att reflektera över sina sociala medievanor genom självreflektion. Applikationen samlar in användarens svar på en rad frågor om sociala medievanor, sömn, psykisk hälsa och andra relevanta, betydande faktorer.

Därefter får användaren en prediktion från en tränad maskininlärningsmodell som bedömer graden av eventuellt beroende.

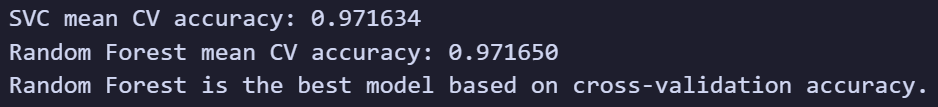
Användarens svar visualiseras i relation till medelvärdena i datasetet, samt mot vetenskapligt förankrade sömnrekommendationer. Förhoppningen är att detta ska förstärka reflektioner och insikter hos användaren – och bidra till ökad medvetenhet kring sociala mediers påverkan.

Streamlit-applikationen utvecklades i Visual Studio Code. Eventuell preprocessering och majoriteten av visualiseringarna är hämtade från EDAn skriven i en Jupyter Notebook. Funktioner i applikationen har delvis utvecklats självständigt med hjälp av Streamlits dokumentation, och delvis konsulterats med stöd av Claude Sonnet ([Anthropic](https://www.anthropic.com/), 2025), ([Streamlit documentation](https://docs.streamlit.io/), 2025).

# Resultat och Diskussion

Efter träning och utvärdering av modellerna *RandomForestClassifier* och *Support Vector Classifier* visade resultaten att båda modellena preserade mycket väl på testdatan. Båda hade hög *accuracy* och ett balanserat klassificeringsresultat mellan de tre klasserna *Low, Medium, High*.

RandomForestClassifier uppmätte en öerhört marginell skillnad med något högre accuracy än SVC:



Skillnaden var 0.000016 för att vara mer exakt.

Eftersom hyperparametrarna hos SVC medvetet utforskat både linjär och icke-linjär separation kan vi ta en närmre titt på vilken hyperparameter som presterade bäst:



Vi ser att den linjära visade sig vara starkast. Trots det linjära sambandet presterade ändå RandomForestClassifier något bättre. Detta tyder på att datan till stor del ter sig att vara linjärt separerbar – men RandomForestClassifier har möjligheten att även fånga upp små variationer och icke-linjära mönster.

A chart of different colored squares

AI-generated content may be incorrect.A chart of different colored squares

AI-generated content may be incorrect.Jämnheten mellan modellerna speglas i resterande utvärderingsmått, där både Confusion Matrix och Classification Report är identiska eftersom det hade behövts högre antal decimaler för att beskriva de minimala skillnaderna.

**SVC – Classification Report RFC – Classification Report**

| **Klass** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| 1 | 0.90 | 0.90 | 0.90 | 20 |
| 2 | 0.98 | 0.98 | 0.98 | 81 |
| 3 | 1.00 | 1.00 | 1.00 | 40 |
| **Accuracy** |  |  | 0.97 | 141 |
| **Macro avg** | 0.96 | 0.96 | 0.96 | 141 |
| **Weighted avg** | 0.97 | 0.97 | 0.97 | 141 |

| Klass | Precision | Recall | F1-score | Support |
| --- | --- | --- | --- | --- |
| 1 | 0.90 | 0.90 | 0.90 | 20 |
| 2 | 0.98 | 0.98 | 0.98 | 81 |
| 3 | 1.00 | 1.00 | 1.00 | 40 |
| Accuracy |  |  | 0.97 | 141 |
| Macro avg | 0.96 | 0.96 | 0.96 | 141 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 141 |

Tolkningen av resultaten visar att det är möjligt att - med relativt hög precision - förutsäga grad av socialt medieberoende utifrån den självskattade data som modellen tränats på. Bland klasserna finner vi högst träffsäkerhet hos klass 3 (som representerar *High*). Denna gradering är viktigast att prediktera korrekt vid riskbedömning.

Valet mellan modellerna verkar ha mindre betydelse – kvaliteten på datasetet och betydande features väger tyngre. Återigen vill jag dock nämna begränsningen gällande självskattade värden. Denna typen av data är känslig för osäkerhet och diverse bias.

## Förbättringar

Generellt är jag nöjd med både modellernas höga resultat. Modellerna hade kunnat utvärderats med ytterligare mått, så som *ROC-AUC* för att ge en mer nyanserad bild av prestandan.

Det vore en stor förbättring att samla in mer data. Dels hade detta minskat risken för överanpassning hos modellerna, men det hade även skapat mer användbarhet hos features som *Country* – som hade hela 79st länder med enbart 1st respondent.

# Slutsatser

Utforskandet av datan visade att den dagliga användningen av sociala medier bland studenter är hög. Vidare fanns det samband som indikerade att hög användning:

* Leder till högre självskattat beroende.
* Kan påverka sömn negativt.
* Kan påverka psykiskt mående negativt.
* Kan bidra till fler konflikter i relationer.
* Kan försämra studieresultat.

Däremot är det nämnvärt att alla dessa faktorer också korrelerar med varandra. Tillsammans kan de bilda en negativ spiral som gör det svårt att få en helhetsbild och att dra tydliga slutsatser.

Gränvärden kunde slutligen identifieras med hälp av Bergen Social Media Addiction Scale och binning av *Addicted\_Score*. Genom analys av fördelningen kunde vi särskilja *Low, Medium* och *High* risk för beroende eller problematiskt användande. Dock är dessa gränsvärden beroende av självskattade värden, vilket begränsar legitimiteten.

Båda maskininlärningsmodellerna visade att det är möjligt att med hög precision prediktera vilken riskgrupp en person tillhör – baserat på deras svar i formuläret i Streamlit-applikationen. Användare kan få en direkt indikation på hur deras sociala medie-vanor påverkar faktorer i vardagen. Genom detta har de chansen att erhålla ökad självinsikt genom reflektion och information.

# Självutvärdering

1. **Vad har varit roligast i kunskapskontrollen?**

Det roligaste i kunskapskontrollen har varit att få dedikera ordentligt med tid för att kunna förkovra mig i ett ML-projekt från början till slut.

1. **Vad har varit mest utmanande i arbetet och hur har du hanterat det?**Det mest utmanande för mig har varit mitt breda intresse – ett dubbeleggat svärd som ibland kan försvåra uppgifter som kräver snävhet. Stundvis har det känts som att jag befinner mig mitt ute i öppet hav, utan land i sikte på något håll. Och det är min förmåga att finna intresse för *allt*, samtidigt som min hjärna är en idéspruta, som försätter mig där. Ibland har det lett till famlande, och ibland till sidospår där jag sedan fått sortera bort eller göra om arbete.

Jag hanterar detta genom checklistor och prioriteringar som håller mig på rätt spår. Samtidigt har det också varit väldigt värdefullt att bolla idéer och tankar med andra.

1. **Hur har grupparbetet gått?**

Grupperbetet har gått ypperligt bra. Malin och jag har arbetat tidigare. Vi är olika på ett bra sätt. Malin är duktig på att hitta den simplaste vägen eller lösningen, medan jag tillför bredden och är ganska bra på att överkomplicera saker. Tillsammans brukar vi hitta kompromisser för våra idéer där både bredden med olika perspektiv vägs in, samtidigt som tillvägagångssättet skalas ner och hålls inom rimliga gränser.

# Appendix A

https://github.com/DavidLabett/ML-Social-Media-Addiction

# Källförteckning

* https://www.kaggle.com/datasets/adilshamim8/social-media-addiction-vs-relationships
* <https://datareportal.com/reports/digital-2025-global-overview-report>

* <https://hub.salford.ac.uk/psytech/2021/08/10/bergen-social-media-addiction-scale/>
* Prgomet, Antonio, Terese Johnson, Amanda Solberg, Linus Rundberg Streuli (2025) Lär dig AI från grunden – Tillämpad maskininlärning med Python (Pedagogicus Publishing)
* <https://www.sleephealthjournal.org/article/s2352-7218(15)00015-7/fulltext>
* <https://scikit-learn.org/stable/modules/cross_validation.html>
* <https://www.geeksforgeeks.org/machine-learning/major-kernel-functions-in-support-vector-machine-svm/>
* <https://docs.streamlit.io/>
* https://www.anthropic.com/