Hälsoanalys baserad på wearables

En studie av stressnivåer och hälsotrender med data från smartklockor

Alia Atawna

EC Utbildning

Kunskapskontroll 2– Projekt i Data science

2024-10

Abstract

This study examines the use of wearable technology data to predict stress levels and identify health patterns through machine learning. Using data from devices like Apple Watch and Fitbit, which monitor heart rate, sleep, and activity, we aimed to develop predictive models for stress. We applied Logistic Regression and RandomForest classifiers, finding RandomForest particularly effective in identifying key stress predictors such as step count and calorie expenditure. Feature engineering improved model accuracy, and results indicated that wearable data can offer valuable insights into individual stress patterns. These findings suggest wearables have potential in preventive healthcare, enabling real time monitoring and proactive stress management. Future work could incorporate more physiological markers and temporal data for enhanced model performance.

Innehållsförteckning

1 Inledning

I dagens samhälle har hälsa och välbefinnande blivit en central fråga, särskilt med ökande stressnivåer och den snabba takten i det moderna livet. Wearable teknologier som smartklockor och fitnessarmband, såsom Apple Watch och Fitbit, har gjort det möjligt för människor att övervaka sina hälsovärden i realtid. Dessa enheter genererar stora mängder datarelaterade till fysisk aktivitet, sömnmönster och hjärtfrekvens, vilket erbjuder värdefulla insikter om individers hälsa och livsstil. (Piwek, Ellis, Andrews, & Joinson, 2016).

Med den ökande tillgången till data från wearables har möjligheten att analysera och dra slutsatser om hälsotrender och riskfaktorer ökat avsevärt. Genom att använda datavetenskap och maskininlärning kan vi inte bara förstå dessa trender, utan också förutsäga potentiella hälsoproblem. Sömn, till exempel, har visat sig spela en avgörande roll för både mental och fysisk hälsa, vilket gör sömnmönster till en viktig variabel vid hälsodataanalys. (Piwek, Ellis, Andrews, & Joinson, 2016).

I detta projekt syftar vi till att analysera data från wearables för att identifiera mönster som är relaterade till stress och andra hälsoutfall, med hjälp av variabler som hjärtfrekvens, sömntid och aktivitetsnivåer. Forskning visar att hjärtfrekvens och andra fysiologiska mätvärden kan användas för att uppskatta stress och känslomässiga tillstånd, vilket gör dessa variabler viktiga för analysen.

Syftet med denna rapport är att undersöka hur data från wearables kan användas för att förutsäga höga stressnivåer och andra hälsorelaterade utfall, vilket kan hjälpa individer att få bättre insikt i sin hälsa och vidta förebyggande åtgärder vid behov. För att uppfylla syftet kommer följande frågeställningar att besvaras:

1. Hur kan daglig hjärtfrekvens och sömnmönster användas för att förutsäga höga stressnivåer?
2. Vilka variabler från wearables har störst påverkan på individens stressnivå och allmänna hälsa?

Genom att analysera dessa frågeställningar hoppas vi kunna bidra till förståelsen av hur wearables kan användas som verktyg för förebyggande hälsovård och ökad medvetenhet kring individuella hälsotrender.

2 Teori

För att analysera hälsodata från wearables och förutsäga stressnivåer har vi använt flera datavetenskapliga tekniker och maskininlärningsmetoder. Dessa metoder inkluderar explorativ dataanalys (EDA) för att förstå mönster i hälsodata, feature engineering för att skapa nya variabler som förbättrar modellens prestanda, samt klassificeringsmodeller för att förutsäga perioder med hög stress baserat på faktorer som hjärtfrekvens, steg och sömn.

2.1 Explorativ dataanalys (EDA)

EDA är en viktig metod inom datavetenskap som syftar till att ge en första förståelse för datasetets struktur och identifiera mönster, samband och eventuella avvikelser. I denna studie utförde vi EDA på variabler såsom hjärtfrekvens, steg, kaloriförbrukning och sömn för att undersöka hur dessa faktorer varierar över tid och relaterar till varandra. Genom att använda visualiseringar som linjediagram och boxplotter kunde vi få en tydlig bild av hur variablerna förändrades över tid och hur de kunde vara relevanta för stressanalys.

Exempel på EDA:

- **Histogram och tidsserier** för att visa fördelningen av hjärtfrekvens och antalet steg per dag.
- **Korrelationsmatris** som visade relationer mellan variabler och hjälpte oss att identifiera vilka variabler som var starkt kopplade till stress.

EDA bidrog till att identifiera vilka variabler som är mest betydelsefulla för vår analys av stress och hälsa. (U.S. Environmental Protection Agency, 2023).

2.2 Feature engineering

Feature Engineering innebär att skapa nya variabler eller "features" baserade på existerande data för att förbättra maskininlärningsmodellens prestanda och tolkningsbarhet. I vårt projekt använde vi Feature Engineering för att utveckla variabler som mer direkt relaterar till stressnivåer och aktivitetsmönster. (Domino Data Lab, n.d) Exempel på sådana features är:

- **Tidsbaserade features**: Genom att dela upp dagen i tidsperioder som morgon, eftermiddag och kväll kunde vi analysera variationer i stressnivåer beroende på tidpunkt.
- **Rullande medelvärden**: Vi skapade rullande medelvärden av hjärtfrekvens och steg över en veckas tid. Dessa trender hjälpte till att jämna ut dagliga fluktuationer och gav oss en mer långsiktig bild av aktivitets- och stressnivåer.
- **Sömnkvalitet**: Vi beräknade en sömnkvalitetsvariabel baserat på total sömntid i relation till den tid som spenderats i sängen. Sömnkvaliteten kan kopplas till stressnivåer och fungerar som en indikator på återhämtning.

Feature Engineering förbättrade modellens förmåga att förutsäga stressnivåer genom att skapa variabler som representerar relevanta mönster och trender i datan.

2.3 Klassificeringsmodeller

För att förutsäga perioder med hög stress använde vi olika klassificeringsmodeller inom maskininlärning. Syftet var att identifiera perioder som klassificerades som stressiga eller icke stressiga baserat på de hälsovariabler vi samlat in. De klassificeringsmodeller vi använde var bland annat:

- **Logistisk regression**: En enkel och effektiv modell för binär klassificering, som fungerar bra för att förutsäga kategorier som stressig/icke stressig. Logistisk regression använder sig av sannolikheter för att bestämma till vilken kategori en viss datapunkt tillhör.
- **RandomForest**: RandomForest är en ensemblemodell som bygger på flera beslutsträd för att förbättra förutsägelsens noggrannhet och stabilitet. I vårt fall hjälpte RandomForest oss att visualisera vilka faktorer som har störst inverkan på stressnivåer, såsom hjärtfrekvens och sömnkvalitet. RandomForest har fördelen att den minskar risken för överanpassning, vilket kan vara ett problem med enskilda beslutsträd, och ger dessutom en tydlig bild av variabelviktighet.

2.4 Utvärdering av modeller

För att säkerställa att våra modeller var effektiva utvärderade vi dem med hjälp av prestandamått som noggrannhet, precision och återkallning. Dessa mått hjälpte oss att bedöma modellens förmåga att korrekt klassificera stressiga perioder och att identifiera hur pålitlig modellen är. Genom att justera hyperparametrar och utvärdera modellens resultat på olika tränings- och testuppsättningar kunde vi förbättra noggrannheten i våra förutsägelser.

3 Metod

Denna del beskriver de metoder vi använt för att samla in, bearbeta och analysera data från wearables, med målet att förstå och förutsäga stressnivåer. Arbetet genomfördes i flera steg där varje gruppmedlem ansvarade för specifika uppgifter för att effektivt fördela arbetet.

3.1 Agil arbetsmetodik

Projektet genomfördes enligt en agil arbetsmetodik där gruppen regelbundet hade möten via teams för att diskutera framsteg och justera arbetet utifrån de nya insikter som uppstod. Den agila metoden främjar samarbete och flexibilitet, vilket underlättar när förändringar eller nya idéer behöver integreras i projektet (Agile Alliance, 2001).
För att organisera arbetet använde vi Trello för att dela in uppgifterna i mindre steg och hålla koll på vilka delar som var avklarade och vilka som återstod. Gruppmedlemmarna tog ansvar för olika delar av analysen och modellen, men samarbetade även över olika moment när det behövdes. Till exempel bidrog två medlemmar till Explorativ dataanalys och feature engineering, vilket gjorde att arbetet fortskred smidigt och att fler perspektiv kunde inkluderas i processen.
Den agila metoden innebar också att vi efter varje större delmoment, som t.ex. EDA eller modellutvärdering, reflekterade över vad som fungerade väl och vad som kunde förbättras. Genom veckovisa möten och återkoppling kunde vi snabbt justera arbetet för att möta projektets behov och säkerställa att vi nådde våra mål i tid.

3.2 Datainsamling

Data som analyserades i detta projekt samlades in från wearables, inklusive enheter som Apple Watch och Fitbit. Datamängden inkluderade variabler som hjärtfrekvens, steg, kaloriförbrukning och sömnmönster, vilka är relevanta indikatorer för stress och aktivitetsnivåer.

Arbetet med att hämta data leddes av en i gruppen som ansvarade för att konfigurera en Google Colab miljö där de insamlade datan strukturerades och laddades upp till en Google Drive-mapp. På så sätt kunde hela gruppen enkelt komma åt och bearbeta datan. Data från varje mätning sparades i separata CSV filer, vilket gav möjlighet att utföra Explorativ Dataanalys (EDA) och identifiera mönster. En kombinerad dataset skapades för att underlätta vidare analys och modellträning.

	date	Id	mean	max	min	TotalSleepRecords	SleepHours	TimeInBedHours	Calories	StepTotal
2	2016-04-12	4020332650	83.4990136324824	133	49	1	8.35	9.016666666666667	3654	8539
3	2016-04-12	5553957443	64.36511350059737	106	50	1	7.35	7.733333333333333	2026	11596
4	2016-04-12	5577150313	65.65607369027059	154	41	1	6.983333333333333	7.3	3405	8135
5	2016-04-12	6962181067	85.03632013919095	176	47	1	6.1	6.45	1994	10199
6	2016-04-12	8792009665	68.92157638065112	135	48	1	7.633333333333334	8.216666666666667	2044	2564
7	2016-04-13	2347167796	73.81290461804058	158	55	1	7.783333333333333	8.85	2038	10352
8	2016-04-13	5553957443	61.483418208918145	95	47	2	7.583333333333333	8.133333333333333	1718	4832
9	2016-04-13	5577150313	59.7035779576913	121	44	1	7.2	7.633333333333334	2551	5077
10	2016-04-13	6775888955	82.71974376572867	135	55	1	3.9166666666666665	4.333333333333333	2400	4053

Figur 1: Visualisering av datainsamling och strukturering

3.3 Dataförberedelse

För att säkerställa att datan var ren och konsekvent genomfördes flera steg för dataförberedelse innan analys och modellering. Det första steget var att ladda in de insamlade CSV filerna i analysmiljön och utföra initial datarensning. Detta innebar att identifiera och hantera saknade värden, samt att justera datatyperna för att matcha analysens krav.

Datan omvandlades och strukturerades i Google Colab för att underlätta bearbetning. Viktiga variabler såsom hjärtfrekvens (medelvärde, max och min), antal steg, kaloriförbrukning och sömntid standardiserades och organiserades i en enda kombinerad dataset. Genom att integrera dessa variabler skapade vi en sammanhängande datamängd som lade grunden för analys och modellträning. Vid behov skapade vi även nya variabler genom feature engineering för att representera relevanta mönster, exempelvis rullande medelvärden och sömnkvalitet, vilket förbättrade vår analys.

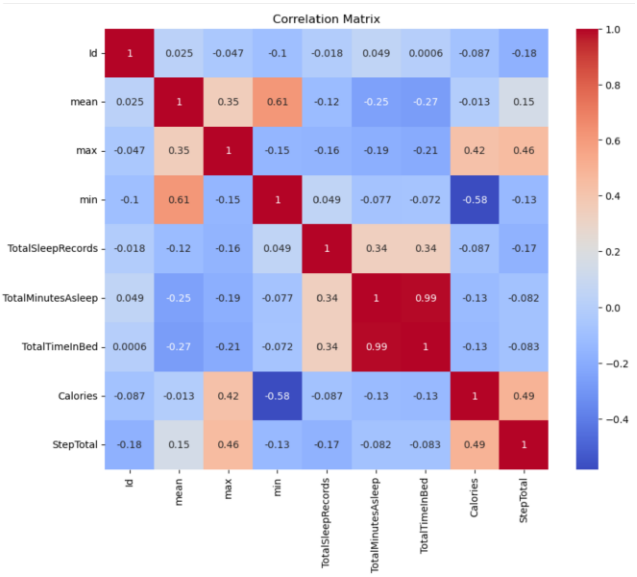
3.4 Explorativ dataanalys (EDA)

En Explorativ Dataanalys (EDA) utfördes för att undersöka datasetets struktur och identifiera mönster och avvikelser. Målet var att förstå hur variablerna, som hjärtfrekvens, sömn och antal steg, relaterade till varandra och om de hade potentiella samband med stressnivåer.

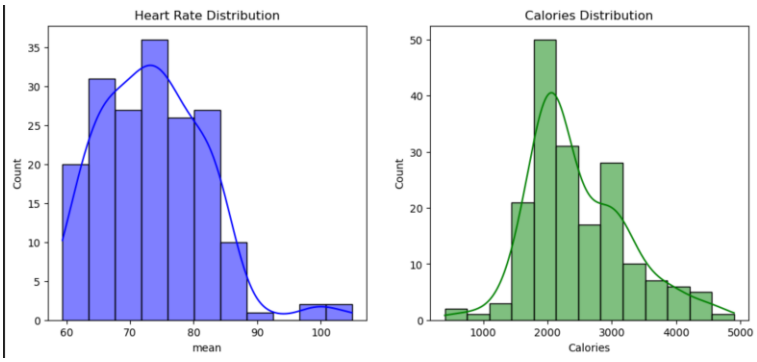
Under EDA skapade vi olika visualiseringar:

- **Histogram** för att visa distributionen av steg och kalorier.
- **Linjediagram** för att observera hur hjärtfrekvens och aktivitetsnivåer förändrades över tid.
- **Korrelationsmatris** för att förstå sambanden mellan variabler som hjärtfrekvens, sömnkvalitet och kaloriförbrukning.

Dessa visualiseringar hjälpte oss att identifiera variabler som kan påverka stressnivåer och gav värdefulla insikter som användes för att förbättra våra modeller.



Figur 2: Korrelationsmatris från explorativ dataanalys



Figur 3: Histogram visar distributionen av steg och kalorier.

3.5 Modellering och utvärdering

För att förutsäga stressnivåer använde vi flera klassificeringsmodeller, inklusive logistisk regression och RandomForest, vilka är väl lämpade för binära klassificeringsproblem, som att avgöra om en period är stressig eller inte.

- **Logistisk regression**: Denna modell användes som en baslinje för att förutsäga sannolikheten för höga stressnivåer baserat på variabler som hjärtfrekvens och sömnkvalitet.
- **RandomForest**: RandomForest modellen är en metod som består av flera beslutsträd. Den tillåter oss att se hur olika variabler bidrar till förutsägelserna och ger insikt i vilka faktorer som mest påverkar stressnivåerna, såsom höga värden på hjärtfrekvens eller låg sömnkvalitet.

Modellerna utvärderades genom prestandamått som noggrannhet, precision och återkallning för att säkerställa att våra förutsägelser var korrekta och pålitliga. Vi använde även F1-score för att bedöma balansen mellan precision och återkallning. RandomForest modellen tillät dessutom visualisering av variabelviktighet, vilket gav en tydlig bild av de mest betydelsefulla variablerna för stressprediktion.

3.6 Modelljustering och hyperparameteroptimering

För att förbättra modellens prestanda justerades hyperparametrar och vi genomförde korsvalidering för att säkerställa att modellen var robust och generaliserbar. Genom att justera hyperparametrar, såsom beslutsträdets maxdjup eller regulariseringsparametern i logistisk regression, kunde vi hitta en optimal balans mellan modellens noggrannhet och komplexitet. Detta hjälpte oss att förbättra modellens prediktiva förmåga och minska risken för överanpassning till träningsdatan.

Genom att noggrant utvärdera modellerna och anpassa dem till data från wearables kunde vi utveckla en pålitlig modell för att förutsäga perioder av hög stress, vilket kan vara ett användbart verktyg för hälsoövervakning.

4 Resultat och Diskussion

I detta projekt analyserade vi data från wearables för att förutsäga stressnivåer och förstå mönster i hälsorelaterade variabler såsom hjärtfrekvens, antal steg och sömnkvalitet. Genom att tillämpa olika maskininlärningsmodeller och jämföra deras prestanda kunde vi identifiera de metoder som bäst passade för att göra dessa förutsägelser.

4.1 Feature engineering och variabelval

Vi skapade en binär variabel "HighStress" baserad på en kombination av höga värden för hjärtfrekvens, kaloriförbrukning och låg sömn. Denna etikett definierades som "1" när hjärtfrekvensen var över 90, kalorier var över 2500 och sömntimmarna var färre än 7, annars "0". Detta var avgörande för att förbättra modellens förmåga att förutsäga stressnivåer genom att tydligt identifiera perioder som potentiellt stressiga.

4.2 Modellutvärdering

För att förutsäga stressnivåer testade vi två modeller: Logistic Regression och RandomForest. Båda modellerna tränades och testades, och prestandan utvärderades med mått som precision, recall och accuracy.

1. **Logistic regression**: Modellen uppnådde en hög noggrannhet med balanserad precision och recall. En femfaldig cross validation gjordes för att bekräfta modellens stabilitet. Resultaten från Logistic Regression visade att modellen konsekvent kunde förutsäga höga stressnivåer med hög noggrannhet.

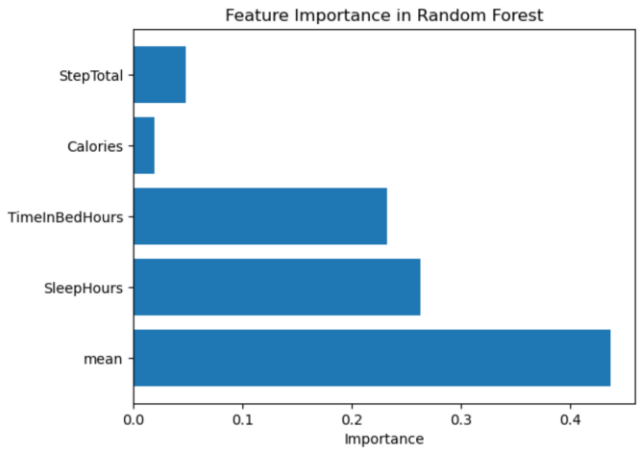
2. **RandomForest**: RandomForest modellen presterade utmärkt och visade en förmåga att identifiera de mest betydelsefulla variablerna för att förutsäga stress. Modellen uppnådde högsta möjliga accuracy på testdatan efter att vi tillämpade oversampling för att hantera klassobalansen. Cross validation accuracy var också konsekvent hög.

Modell	Noggrannhet	Precision	Recall	F1-score
Logistic Regression	95.8%	96%	92%	96%
RandomForest	100%	100%	100%	100%

Tabell 1: Modellprestanda för Logistic Regression och RandomForest.

4.3 Viktiga prediktorer för stressnivåer

För RandomForest modellen analyserades variabelbetydelse för att identifiera vilka variabler som mest påverkar modellens beslut. Stegantal och Kalorier var de mest betydelsefulla variablerna för att prediktera höga stressnivåer. Detta tyder på att fysisk aktivitet och energiförbrukning har starka kopplingar till stressnivåer, vilket är värdefullt för vidare forskning om hälsa och stresshantering.



Figur 4: Viktighet av variabler för stressprediktion i RandomForest modellen.

4.4 Diskussion

Våra resultat visar att wearables kan ge värdefull information för att förutsäga och övervaka stressnivåer. RandomForest modellen visade sig vara särskilt effektiv och kunde tydligt identifiera mönster kopplade till hög stress. Det är intressant att se hur fysisk aktivitet (antal steg) och kaloriförbränning spelar en roll i stressnivåerna, vilket stödjer hypotesen att regelbunden fysisk aktivitet kan bidra till att hantera stress.

Genom att använda oversampling kunde vi hantera klassobalansen, vilket förbättrade modellens förmåga att korrekt identifiera högstressperioder. Denna teknik visade sig vara avgörande för att uppnå tillförlitliga prediktioner och kan rekommenderas för liknande projekt med obalanserade dataset.

5 Slutsatser

I denna studie analyserades data från wearables för att undersöka mönster relaterade till stressnivåer och hälsotrends. Målet var att förutsäga höga stressnivåer baserat på variabler som hjärtfrekvens, sömn och fysisk aktivitet, samt att identifiera vilka faktorer som har störst inverkan på en individs stress och allmänna hälsa.

1. **Hur kan daglig hjärtfrekvens och sömnmönster användas för att förutsäga höga stressnivåer?**

Resultaten visar att hjärtfrekvens och sömnmönster är viktiga indikatorer på en individs stressnivå. Genom att använda en kombination av höga värden för hjärtfrekvens, hög kaloriförbrukning och låg sömntid kunde vi skapa en binär etikett för höga stressnivåer, vilket förbättrade modellens prediktiva förmåga. Feature engineering, såsom rullande medelvärden och beräkning av sömnkvalitet, gjorde det möjligt att fånga långsiktiga trender som hjälper till att identifiera perioder av hög stress.

2. **Vilka variabler från wearables har störst påverkan på individens stressnivå och allmänna hälsa?**

Analysen av variabelviktighet i RandomForest modellen visade att antalet steg och kaloriförbrukning var de viktigaste variablerna för att förutsäga höga stressnivåer. Detta antyder att fysisk aktivitet och energiförbrukning har starka kopplingar till stress, och kan potentiellt användas som förebyggande indikatorer. Även sömnkvalitet och hjärtfrekvens spelade en viktig roll i modellen, vilket stödjer tidigare forskning om deras betydelse för hälsa och välmående.

5.1 Sammanfattning

Sammanfattningsvis visar denna studie att wearables erbjuder värdefulla data för att övervaka och förutsäga stressnivåer. RandomForest modellen visade sig vara särskilt effektiv för detta ändamål och gav insikter i vilka faktorer som är mest kopplade till stress. Genom att använda maskininlärningstekniker kunde vi utveckla en robust modell för stressprediktion, vilket har potential att användas som ett verktyg för förebyggande hälsovård och bättre medvetenhet om individens hälsotrender.

Dessa resultat kan bidra till utvecklingen av personanpassade hälsotjänster som använder data från wearables för att identifiera riskfaktorer i realtid. Det fortsatta arbetet skulle kunna fokusera på att vidareutveckla modellerna genom att inkludera fler fysiologiska variabler och undersöka hur stressnivåer påverkas av faktorer som sömnkvalitet över längre perioder.

6 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Jag stötte på svårigheter med obalanserad data och modellval, som jag löste genom oversampling och omfattande EDA för att förbättra analysen och modellprestandan.

2. Vilket betyg du anser att du skall ha och varför.
   Jag anser att betyget G är lämpligt för mig, då jag tror arbetet uppfyller de grundläggande kriterierna.

3. Något du vill lyfta fram till Antonio?
   Jag har fått värdefull insikt i hälsodataanalys och maskininlärning och uppskattar handledningen som hjälpt mig förstå metodologin bättre.

Källförteckning

Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The Rise of Consumer Health Wearables: Promises and Barriers. *PLOS Medicine*, 13(2), e1001953. Hämtad 28 oktober, 2024 från https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001953

U.S. Environmental Protection Agency. (2023). *Exploratory Data Analysis*. Hämtad 28 oktober, 2024 från https://www.epa.gov/caddis/exploratory-data-analysis