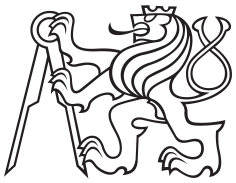
Bakalárska práca



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická

Zber a analýza dát získaných pomocou mobilných platforiem pre zber behaviorálnych dát

Klasifikácia a detekcia pracovných a voľných dní na základe dát z akcelerometru

Ondrej Sakači

Školiteľ: doc. Ing. Daniel Novák, Ph.D.
Školiteľ–specialista: Ing. Jakub Schneider, Ph.D.
Január 2022

Poďakovanie

Formálne by som sa rád poďakoval Fakulte elektrotechnickej za poskytnutie zdrojov pre náš projekt.

Ďalej by som sa chcel nesmierne poďakovať profesorovi Danielovi Novákovi, za vynikajúce vedenie tejto Záverečnej práce, kvelé nápady a rady k riešeniu.

Doktorovi Jakubovi Schneiderovi, za asistenciu pri vedení práce, spätnú väzbu a cenné pripomienky, v neposlednom rade za trpezlivosť.

Za pomoc pri deploymente a údržbe LAMP platformy na školskom serveri patrí moja vďaka server administrátorovi Jiřímu Wildovi.

Ďakujem všetkým svojim kolegom zo semestrálneho projektu a týmu pod vedením profesora D. Nováka - Michal Kubina, Lukáš Sláma, Eric Žíla - za spoluprácu a pomoc.

Na koniec patrí moja vďaka aj týmu projektu mindLAMP za aktívny a spoľahlivý support.

Prehlásenie

Čestne prehlasujem, že som predloženú prácu vypracoval samostatne, a že som uviedol všetku použitú literatúru.

V Prahe, 4. januára 2022

I declare that the following document was written solely by me and that I had cited all the sources I had used in the bibliography.

Prague, January 4, 2022

Abstrakt

Záverečná bakalárska práca je finálnym výstupom projektu, ktorý sa začal pred viac ako rokom a zavŕšením bakalárskej etapy štúdia autora.

Poskytuje úvod do vedného oboru digitálneho fenotypovania a aktigrafie.

Predstavuje predpoklady a hypotézy pre záverečnú prácu a náhľad do metodológie práce.

Hlavná úloha autora bola klasifikácia dní na základe pohybovej aktivity. Riešenie tejto úlohy je v práci opísané od implementácie, cez opis dátových štruktúr, postup analýzy dát až po výsledky.

Mimo hlavnú úlohu pojednáva aj o sekundárnej úlohe autora, ktorou bolo nasadenie a údržba platformy mindLAMP na školskom serveri.

Kľúčové slová: digitálny fenotyping, zber dát, analýza dát, mobilné zariadenia, akcelerometer, GPS, klasifikácia dát, strojové učenie, nasadenie platformy

Školiteľ: doc. Ing. Daniel Novák, Ph.D.
Ústav Analýzy a interpretace biomedicínských dat,
Na Zderaze 269/4,
Praha

Abstract

The Final bachelor thesis is the final output of the project which had started more than a year ago and it is a completion of the bachelor study phase of the author.

Thesis provides introduction to the science field of digital phenotyping and actigraphy.

It introduces postulates and hyphoteses for the Bachelor thesis and a peek into the methodology of the project.

Main task of the author was classification of the days based on the movement activity. Solution to this task is described in the thesis from implementation, through data sctructure descriptions, process of data analysis to the results.

Beyond the main task it discusses the secondary task of the author which was deployment and maintainance of the mindLAMP platform on the school server.

Keywords: digital phenotyping, data collection, data analysis, mobile devices, accelerometer, GPS, data classification, machine learning, platform deployment

Title translation: Data collection and analysis from behavioral mobile platforms — Classification and detection of work days and free days based on the accelerometer data

Obsah

Obrázky

Tabuľky

Kapitola 1
Úvod

Táto práca je zhrnutím vyše ročnej práce štvorčlenného tímu a výstupov štúdie, ktorá bola prevedená pod vedením doc. Ing. Daniel Novák, Ph.D. s odborným dohľadom a konzultáciami Ing. Jakub Schneider, Ph.D..

Na nasledujúcich stránkach sa nachádzajú hlavné motivácie a ciele bakalárskej práce, teoretický úvod, implementáciu a samotné výsledky, ku ktorým autor dospel.

Hoci sa na štúdií podieľal viac-členný tým, každý mal vlastné zadanie toho, ako má dáta analyzovať. Táto záverečná práca je teda výstupom autora samotného.

1.1 Motivácia

V dnešnom svete je jedným z výrazných, avšak stále veľmi bagatelizovaných problémov úpadok duševného aj telesného zdravia. Či už sa jedná o psychické alebo telesné problémy spôsobené nezdravým životným štýlom (sedavá práca, nezdravé stravovanie, nedostatok pohybu, atď.) alebo prostredím v ktorom sa človek nachádza a žije.

Okrem toho aké podmienky životné alebo pracovné vplývajú na psychický a fyzický stav jednotlivca, nemožno opomänúť dopad aký mala a v dobe písania tejto práce stále má celosvetová pandémia vírusu Covid-19 na ľudí. Nie len negatívny dopad vo forme strachu z nákazy. Veľa ľudí prišlo počas lockdownov o živobitie a niektorí aj o blízkych. Mimo to ani ľudia, ktorý neboli zasiahnutý až tak drasticky často pociťovali dopady lockdownu inou formou. Práca z domova alebo štúdium na diaľku boli jedny z miernejších dopadov ale okrem socializácie ovplyvnili aj obecný pohyb jednotlivca (samozrejme sa nejedná o obecné pravidlo).

Ako tým veríme, a nie sme jediní, že duševné aj fyzické zdravie je možno monitorovať a korigovať za pomoci zberu dát z mobilných zariadení. Tieto dáta sa po spracovaní dajú využiť ako informačný doplnok pre špecialistu (doktora). Existuje taktiež možnosť, že môžu existovať modely, ktoré by boli schopné nie len klasifikovať a rozpoznávať nálady a aktivity, ale predvídať budúce relpasy pacientov alebo nálady zdravých ľudí. Ak by takéto modeli existovali, dali by sa utilizovať za účelom prevencie zhoršenia psychického ale

aj fyzického stavu jedinca[1].

Téma tejto práce vznikla ešte pred vypuknutím pandémie a bola rovnako veľmi zameraná na pomoc pri diagnóze, kontrole a liečbe duševných porúch (menovite sa malo jednať o pacientov s bipolárnou poruchou), avšak v súčasných podmienkach by mohla pomôcť aj všeobecne zdravým jedincom.

Ak už niekto pociťuje úzkosť, stres alebo depresívne stavy z tých alebo onakých dôvodov, obmedzovanie ľudského kontaktu muselo mať dopady aj na získavanie odbornej pomoci. Za pomoci zberu a spracovania dát z mobilného zariadenia, by bolo možné mať lepší dohľad na pacienta bez nutnosti častého fyzického kontaktu. To je mimo predpokladané pozitíva ďalšie, ktoré má potenciál byť podstatným prvkom pri pomoci psychicky trpiacim ľuďom v čase akejkoľvek súčasnej alebo budúcej celosvetovej/lokálnej zdravotníckej alebo inej krízy.

Viac informácií o problematike zberu a spracovania dát sa nachádza v Kapitole 2. Konkrétne sa pozrieme na disciplínu Digitálneho fenotypovania a takzvané aktigrafické dáta.

1.2 Ciele bakalárskej práce

Hlavnými cieľmi tejto bakalárkej práce sú:

- Zoznámiť sa s aktuálnym stavom a dostupnými mobilnými platformami pre zber a analýzu dát.
- Úspešne nasadiť patformu na zber dát z mobilných zariadení.
- Nájsť minimálne 10 dobrovoľníkov, ktorí sa zúčastnia pilotného testovania platformy a samotnej štúdie.
- Zozbierať aspoň mesiac valídnych dát od väčšiny účastníkov.
- Úspešne klasifikovať dni na voľné a pracovné na základe dát z náramkového akcelerometru a mobilných zariadení.

[1] Nejedná sa len o zhoršenie nálady alebo nedostatok pohybu, ale aj o kombinované dopady ako je sebapoškodzovanie alebo sebevražedné sklony, ktorým by bolo možné zabrániť.

Kapitola 2
Teoretický úvod

V tejto kapitole sa nachádza kompletné teoretické zhrnutie bakalárskej práce. Jedná sa primárne o vysvetlenie dôležitých termínov (digitálny fenotyping, aktigrafia), stručný prehľad platforiem na zber dát z mobilných zariadení, ich porovnanie a prehľad štúdie.

2.1 Digital Phenotyping

Digitálny fenotyping (Digital phenotyping, ďalej DF) je vedná disciplína, ktorá sa zaoberá možnosťami využitia pasívnych a aktívnych dát zozbieraných z mobilných zariadení za účelom zlepšenia zdravotnej starostlivosti a/alebo zdravotného stavu užívateľa[1].

2.1.1 Druhy dát

Dáta, ako už bolo naznačené, sa delia na aktívne a pasívne:

- Aktívne dáta - Sú dáta, ktoré užívateľ musel vedome a aktívne zozbierať a odoslať. Patria sem napríklad dotazníky, hlasové správy alebo fotografie. Akékoľvek dáta, ktoré sa nedajú vygenerovať/zosnímať automaticky a vyžadujú aktívnu účasť užívateľa mobilného zariadenia.
- Pasívne dáta - Sú dáta, ktoré sa dajú zbierať bez aktívnej účasť používateľa mobilného zariadenia. Patria sem dáta z mobilných senzorov ako sú napríklad mikrofón, akcelerometer, magnetometer, gyroskop alebo senzory tepla a vlhkosti vzduchu.

Aktívne dáta môžu byť zbierané aj osobnou formou, napríklad prehliadkou u doktora/psychiatra alebo vyplnením dotazníku osobne. Pasívne dáta na druhú stranu potrebujú prístroj na ich snímanie, zpracovanie a následný prenos, či už formou synchronizácie s dokovacou stanicou na mieste zpracovania dát, alebo odosielaním týchto dát cez internet. Ideálne by sa pasívne dáta mali odosielať cez internet alebo inou cestou bez nútenia užívateľa aktívne sa podieľať na ich odosielaní.

Z týchto dôvodov ideálne zariadenie pre DF musí byť schopné ukladať dáta a pripojiť sa cez internet k serveru, ktorý zbiera a prípadne aj spracováva

zozbierané dáta. Smart telefóny sú pre toto skvelým kandidátom aj z dôvodu, že dnes väčšina populácie smart telefón už vlastní.

Zber pasívnych dát pritom môže prebiehať priamo na samotnom zariadení, alebo ak má bluetooth (ďalej BT) môže byť zber dát outsourceovaný na periférne BT zariadenie ako sú napríklad smart hodinky. Periférne BT zariadenia zozbierané dáta odosielajú do mobilného zariadenia počas synchronizácie, ktorá môže byť kontinuálna, alebo v prípade, že periférne zariadenie obsahuje vlastnú pamäť, môže nastávať raz za určitý časový interval, pričom si zariadenie medzi synchronizáciami dáta ukladá do svojej pamäte.

Jednou z výhod využitia perifémych zariadení ktoré sú nosené na tele je to, že snímajú pohyb osoby presnejšie ako mobilný telefón, ktorý môže často zostať v taške alebo položený na stole. Tiež je výhodou, že náramky na snímanie aktivity nie sú ľuďom neprirodzené nosiť, keďže sme už zvyknutý na hodinky a môžu plniť aj iné praktické účely (ako napríklad ukazovať čas). Náramky nosené na menej dominantnej ruke majú menšiu presnosť správnej klasifikácie činností/aktivity ako senzory pripevnené na boku, ale ich presnosť bola už viac krát dokázaná ako dostačujúca. Jedna štúdia zrovnávajúca náramkové senzory a senzory upevnené na boku uvádza najsilnejší súhlas detekcie dát medzi modelmi náramkov a senzorov na bokoch až 95%[2].

Medzi známych výrobcov smart hodiniek patria napríklad fitbit[3] a Xiaomi[4].

2.1.2 Využitie

Digitálny fenotyping sa najčastejšie využíva pre výskum a zlepšenie mentálneho zdravia, no nie je to pravidlo. Teoreticky môže byť aplikovaný k monitoringu fyzickej aktivity za účelom zlepšenia fyzického zdravia.

Tak ako aj pri akomkoľvek inom zbere dát uživateľov existujú obavy z využitia týchto dát pre marketingové účely alebo ich úniku a následnému zneužitiu.

Marketing

V súčasnosti sa digitálny fenotyping nevyužíva pre marketingové účely a zatiaľ nič nenasvedčuje tomu, že by sa tento stav mal meniť v blízkej budúcnosti. DF vyžaduje využitie senzorov mobilného zariadenia ako napríklad mikrofón alebo akcelerometer, čo znamená zásadný zásah do súkromia používateľa a možný únik citlivých dát.

Dnes sa najbližšie k digitálnemu fenotypovaniu za marketingovými účelmi blížia personalizované reklamy, ktoré profilujú uživateľa na základe online aktivity v aplikáciách alebo na základe cookies z webového prehliadača.

Medzi najväčších gigantov na poli personalizovaných reklamných služieb, ktoré uživateľom na základe ich aktivity na internete zobrazujú relevantné reklamy patria Google (Google AdSense[5]) a Facebook (Facebook Ads[6]). Tieto služby sú implementované primárne na platformách, ktoré tieto spoločnosti vlastnia. V prípade Google AdSense sa jedná napríklad o platformy Google a Youtube. Facebook Ads sa zas dajú pre reklamu využiť na platfor-

mách Facebook, Instagram a v blízkej budúcnosti by mali byť implementované aj do niektorých aplikácií platformy Oculus.

Virtuálna realita (ďalej len VR) je jedna zo zatiaľ málo preskúmaných oblastí, čo sa týka zberu dát uživateľov. V štúdií z roku 2020[7] sa podarilo identifikovať jednotlivca používajúceho VR headset z celkového počtu 551 účastníkov na základe dát nazbieraných počas 276 sekundového intervalu záznamu činnosti s presnosťou 95%. Z tejto štúdie vyplýva, že VR je prekvapivo dobrý zdroj dát na profilovanie uživateľov, počnúc od výšky headsetu od zeme, končiac pohybmi rúk a hlavy, kde je VR schopné zachytiť aj veľmi malé a špecifické pohyby pre jednotlivca.

Hlavný problém, ktorý tu vyvstáva je možnosť zberu dát len ak uživateľ headset aktívne používa. Bez headsetu je zbytok dňa netrackovaný. Teda vyvstáva možnosť využitia VR len ako doplnkový prameň dát, no určite by nebolo rozumné sa naň spoliehať.

VR technológie sa avšak pre digitálny fenotyping používať v dohľadnej dobe nebudú, či už sa jedná o dostupnosť, alebo potrebu zbierať dáta priebežne a nie len počas voľného času uživateľa tráveného vo Virtuálnom prostredí. Facebook akožto prevádzkovateľ platformy Oculus a konkrétne modelov VR headsetov Oculus Quest, by mohol takto nazbierané dáta spájať s Facebookovými kontami uživateľov a tak zlepšovať svoj dataset na identifikáciu a profilovanie uživateľov. Obavy o takéto praktiky vyvstávajú aj z faktu, že posledný model Oculus Quest 2[8] vyžaduje prihlásenie cez Facebook účet a bez tohoto prihlásenia nie je schopný fungovať.

Zber osobných dát z dôvodov personalizácie reklám sa považuje za kontroverznú oblasť aj ak sa jedná len o dáta užívateľa na stránke vlastníka. Facebook v roku 2008 začal riešiť otázky týkajúce sa marketingu pre farmaceutické firmy[9], avšak nechcel personalizovať reklamu na základe uživateľských diagnóz a presonálnych údajov kvôli potenciálu zlého PR B.1.

Za účelom regulácie zberu osobných údajov a prevencií voči prípadným únikom alebo zneužitiu dát užívateľov Európska Únia zaviedla General Data Protection Regulation (ďalej GDPR).

Zdravie

DF sa stále viac a viac využíva v oblasti mentálneho zdravia. Medzi najviac preskúmavané poruchy patria ASD (Autism Spectrum Disorder), schizofrénia, poruchy úzkosti, ADHD (Attention deficit hyperactivity disorder) a depresia.

Záujem o využitie DF v medicínskych oblastiach stále narastá. Tento fakt dokazuje aj neustále narastajúci počet ročne publikovaných štúdií a článkov zaoberajúcich sa DF (Obrázok 2.1).

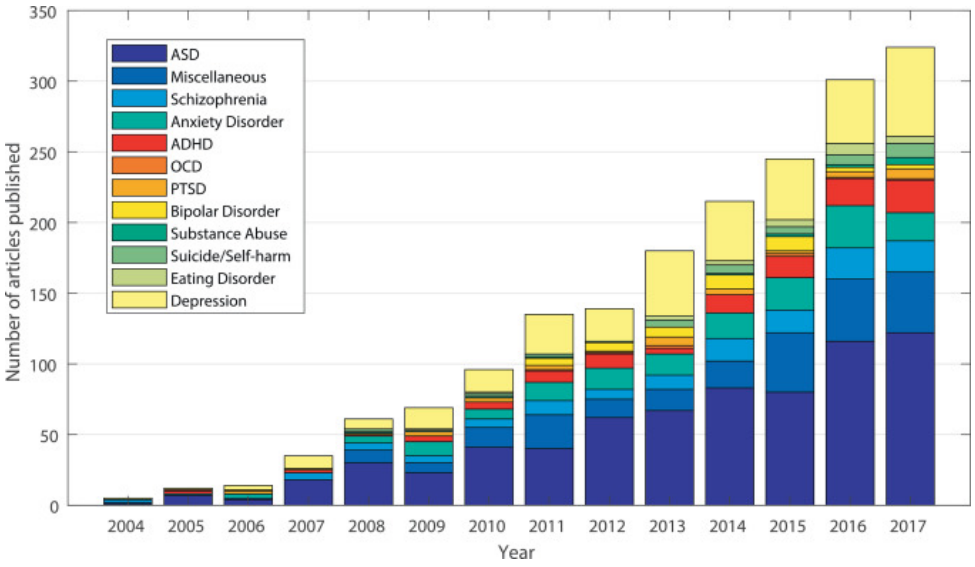
DF sa snaží nachádzať korelácie medzi aktuálnym zdravotným stavom a dennou aktivitou uživateľa/pacienta. V prípade, že je určitá korelácia zachytená, tak nasledujúci krok je často-krát snaha o vytvorenie prediktívneho modelu, aby sa prípadné relapsy zachytili skôr ako naplno prepuknú a bolo im možné predchádzať preventívne.

V prípade pacientov s depresiami sa DF snaží klasifikovať depresívne epizódy, v prípade bipolárnej poruchy (ďalej BP) je snaha identifikovať akožto

depresívne tak aj manické epizódy.

Dobrý prediktívny model by mohol zrýchliť a ulahčiť liečbu pacientov s rozličnými mentálnymi poruchami a predchádzať nežiaducim stavom, ktoré by mohli viesť kľudne až k ublíženiu na zdraví sebe alebo iným osobám.

Čo sa prevencie na základe detekovaných začínajúcich relapsov alebo predikcie budúcich týka, môže byť praktikovaná buďto intervencia lekára alebo teoreticky aj samotnej zdravotnej aplikácie, ktorá by mohla pacienta naviesť určitými neškodnými radami k lepšej starostlivosti o seba.



Obrázok 2.1: Graf znázorňujúci počet publikovaných článkov v jednotlivých rokoch zaoberajúcich sa tematikou digitálneho fenotypovania.[1]

Mimo mentálne problémy môže DF napomáhať aj vo forme sledovania pohybovej aktivity za účelom zlepšenia fyzického zdravia užívateľa. Môže sa jednať o formu aplikácie ktorá funguje ako obyčajný krokomer, no plný potenciál by bol v počítaní učivateľovej sedavej aktivity, kedy by bolo možné užívateľa nemotivovať počtom neodchodených krokov k aktivite, ale počtom hodín presedených za PC, v práci, škole alebo preležaných v posteli.

Tieto dáta sa ale nedajú získať spoľahlivo len z mobilného zariadenia (ako už bolo spomenuté mobil sa nenachádza stále v blízkosti tela aby dokázal správne detekovať pohyb) a preto by najlepším prístupom malo byť čo najmenej prerušované nosenie senzoru aktivity.

2.1.3 Etické problémy

Zber dát zo senzorov mobilného telefónu so sebou prináša isté obavy o ochranu súkromia užívateľa. Keďže smart telefóny majú dnes mimo mikrofón a akcelerometer aj iné senzory. Medzi tieto senzory patria mimo iné: magnetometer, gyroskop, bluetooth, wifi, kamera/y, teplomer, senzor vlhkosti a iné. Tieto senzory majú rôznu úroveň možného narušenia súkromia uživateľa, no sú zneužiteľné všetky. Je tiež pravda, že nie všetky zariadenia majú všetky druhy

senzorov a líšia sa aj ich kvalitou.

Aby sa predchádzalo možnému zneužitiu dáta uživateľov sú anonymizované a často kvôli výskumom sú ešte niekoľko násobne "zaslepené", aby sa predchádzalo možným predsudkom pri spracovaní analytikmi.

Ďalšie možnosti ako obmedziť možné úniky citlivých dát je ich vôbec nezbierať a sústrediť sa na menej invazívne druhy dát ako su napríklad aktigrafické dáta (Kapitola 2.2).

Etikou a šedými zónami DF sa neustále zaoberajú odborníci[10].

V tejto práci etické otázky zberu a analýzy dát nie sú prvoradé. Napriek tomu bola štúdia organizovaná a prevádzkovaná spôsobom, ktorý umožňoval diskrétnosť na dostatočnej úrovni. Viac o organizácií štúdie sa nachádza v Kapitole 3.2.

2.2 Aktigrafia

Aktigrafia je neinvazívna metóda zberu dát. Takzvané aktigrafické dáta sú dáta zozbierané akcelerometrom noseným na tele. Tieto dáta sa od surových dát z akcelerometru líšia tým, že vyjadrujú celkovú úroveň pohybu. Zjednodušene sa jedná o metódu zberu dát kde sa hodnoty z troch osí akcelerometru sčítajú do určitej hodnoty aktivity. Táto suma sa následne zaznamenáva v určitých intervaloch a po zbere sa spracováva.

Touto metódou sa teda nedá monitorovať presný pohyb časti tela na ktorej je aktigraf upevnený, ale je možné sledovať celkovú pohybovú aktivitu za určitý čas.

Jeden z projektov, ktorý aktigrafiu využíva a umožnil nám použitie ich platformy a aktigrafických náramkov, je Mindpax[11].

2.3 Predpoklady a hypotézy

V grafe 2.1 je vidieť, že digitálny fenotyping je aktívne preskúmavaná oblasť a vďaka stále rastúcemu množstvu výskumov je možné stavať našu prácu na už objavených skutočnostiach. Hypotézy, na ktorých stavia táto práca sú vyvodené z viacerých už existujúcich štúdií a výskumov. V tejto sekcií sa nachádzajú predpoklady a hypotézy hypotézy, s ktorými sa pracuje pri implementácií v kapitole 5.

2.3.1 Predpoklady

Predpoklady na ktorých stavia táto štúdia sú rozdelené do niekoľkých podkategórií. Všetky predpoklady sú doložené už dokončenými štúdiami.

Náramok s akcelerometrom sa dá využiť ako presný zdroj dát dennej aktivity

Už v minulosti sa používali zariadenia s akcelerometrom pripevnené k boku sledovanej osoby s dobrou presnosťou určenia aktivity jedinca. Tieto zariade-

nia nie sú však najpraktickejšie. Počas štúdie sme mali k dispozícií aktigrafické náramky a mobilné zariadenia sledovaných subjektov. Aby ale dáta z aktigrafických náramkov boli použiteľné na detekciu aktivity musíme predpokladať, že pohyb menej dominantnej ruky počas dňa dostatočne koreluje s pohybom tela. Funkčnosť a presnosť akcelerometrov nosených na zápästí bola našťastie už dokázaná viacerými štúdiami.

Štúdia z roku 2019 „Wrist-specific accelerometry methods for estimating free-living physical activity"[2] sa zaoberá porovnaním 9 prediktívnych modelov pre akcelerometer nosený na zápästí so špecifickou referenčnou metódou akcelerometra noseného na boku. Z 9044h dát, ktoré boli zozbierané mimo laboratórne podmienky a následne zanalyzované, sa im podarilo namerať senzitivitu a špecificitu nad 60%. Zároveň sa ale odkazujú na štúdiu[12], ktorej sa podarilo dosiahnuť koreláciu medzi náramkovým a telovým akcelerometrom až 96–99%..

Ďalšia štúdia z roku 2016 „Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification"[13] porovnávajúca náramkový akcelerometer s akcelerometrom noseným na boku mimo laboratórne podmienky. Dataset pozostával zo 40 žien s nadváhou alebo obezitou, ktoré nosili oba akcelerometre a prenosné kamery na záznam aktivity pre kontrolu presnosti, po dobu 2 dní. Bol vytvorený model pre detekciu štyroch obecných aktivít (sedenie, státie, chodenie/beh a jazda vo vozidle). Pri detekcií aktivít sa im podarilo zaznamenať priemernú presnosť 89.4% akcelerometru na boku a 84.6% akcelerometra noseného na zápästí.

Taktiež existuje niekoľko štúdií porovnávajúcich rôzne druhy náramkov medzi sebou. Tieto štúdie okrem modelov ktoré využívajú nemajú pre túto prácu priveľký význam v rámci zrovnania náramkov.

T. G. Pavey[1] s týmom pri porovnaní GENEActiv náramkového akcelerometra s akcelerometrom noseným na stehne zistil $ICC^2 = 0.80$ vo svojej práci v roku 2016[15].

Na základe týchto štúdií môžeme predpokladať vysokú presnosť a užitočnosť dát nazbieranych z aktigrafických náramkov nosených na zápästí menej dominantnej ruky.

Senzory z mobilných zariadení môžu byť dobrým zdrojom dát dennej aktivity

Jedna z hlavných obecných tém projektu bol zber dát z mobilných zariadení za účelom analýzy dennej aktivity a nálady uživateľa. Mimo to, že samotný

[1] O rok neskôr, v roku 2017, vydal článok[14] porovnávajúci Criterion krokomer s 13 ďalšími krokomermi. Zrovnanie krokomerov nemá na prvý pohľad veľký prínos pre túto prácu, avšak implikuje veľkú presnosť algoritmov na počítanie krokov z mobilných zariadení. Meranie krokov mobilnými zariadeniami sa dnes využíva vo veľkom. Nie je často tak presné ako špecializované zariadenia, no je dostatočne presné pre sledovanie pohybovej aktivity. V tejto štúdií dosiahli random forest algoritmom nacvičeným v laboratórnych podmienkach priemerné presnosti správnej detekcie aktivít nasledovne: 80.1% sedavá činnosť, 95.7% státie+, 91.7% chôdza and 93.7% beh. Výsledky boli porovnávané s activPAL senzorom a všetky pedometre dosahovali zhodu nad 90%.

[2] Interclass correlation

zber a spoľahlivosť dát zozbieraných z mobilných zariadení bude porovnaná s dátami zozbieranými aktigrafom nosenom na zápästí, je dobré predpokladať aspoň určitú schopnosť týchto dát byť dobrým zdrojom dát.

Dvojica Kangjae Lee a Mei-Po Kwan v roku 2018 sledovali 36 osôb v laboratórnych podmienkach s mobilným telefónom umiestneným vo vrecku nohavíc. Z mobilných zariadení zbierali akcelerometrické a GPS (Global Positioning System) dáta. S použitím prediktívneho random forest klasifikátoru dosiahli presnosť 95.1%. Avšak s použitím gradient boostingu[3] (ďalej GB) zaznamenali presnosť až 99.1%[16].

V roku 2020 tým pod vedením I. M. Pires porovnával niekoľko metód na klasifikáciu aktivity z akcelerometrických dát z mobilu noseného vo vrecku nohavíc. Najhoršie výsledky pre nich dosiahla k-najbližších susedov (k-Nearest neighbors, ďalej kNN) metóda 65.55%. Naopak najlepšie výsledky dosiahli využitím decision trees (ďalej DT) a to konkrétne 85.22%[17].

Napriek faktu, že štúdia K. Lee prebiehala v laboratórnych podmienkach, tak štúdia I. M. Pires dokazuje, že dáta z mobilných telefónov môžu byť použité na prekvapivo presné rozlíšenie aktivít ako sú chodenie do schodov, zo schodov, beh, chôdza, státie/ležanie a sedenie[17].

Zároveň ale nemôžeme prehliadať fakt, že senzory nosené na tele (či už sa jedná o zápästie alebo bok), majú vyššiu presnosť ako mobilné zariadenie.

Toto tvrdenie by sme teoreticky mohli stavať na štúdií „Activity classification using realistic data from wearable sensors“ od J. Parakku z roku 2006[18].

Sú ale dva dôvody prečo to nemusí byť pravda:

1. Štúdia J. Parakku nemala tak dobré výsledky ako iné štúdie.
2. Technológia mobilných telefónov aj prediktívne modely od roku 2006 sa zlepšili.

Na druhú stranu, môžeme predpokladať vyššiu nepresnosť mobilných dát kvôli častejším odloženiam (tj. úsekom, kedy sa mobilný telefón nie je v kontakte s telom ale leží napríklad na stole), než u náramkov. Jeden z dôvodov je obecne vyššia výdrž batérie špecializovaných náramkových senzorov a často bývajú vodotesné, čo umožňuje uživateľovi nosiť náramok väčšinu času.

Z tohoto dôvodu predpokladáme, že dáta z mobilných zariadení, hoci dostatočne presné vždy keď sú v kontakte s uživateľom, môžu byť nespolahlivý zdroj dát. Preto sa v tejto práci zameriame na analýzu aktigrafických dát z náramku a mobilné dáta budú viac-menej fungovať ako doplňujúce dáta na spresnenie aktivity.

Rozloženie práce môže ovplyvniť produktivitu a náladu

Tento predpoklad neznamená, že je potreba mať efektívny rozvrh dňa pre zvýšenie produktivity. Samozrejme mať usporiadané aktivity je vždy výhoda, no v tomto bode ide skôr o vplyv nesprávne rozloženej aktivity.

[3]Jedná sa o machine learning techniku pre regresiu, klasifikáciu a iné úlohy. Produkuje prediktívny model kombináciou viac slabích prediktívnych modelov.

V roku 2007 prebehla štúdia „Workdays, in-between workdays and the weekend: A diary study on effort and recovery“[19]. Štúdia sa snažila objaviť koreláciu medzi vyťaženosťou v práci, mentálnym zdravím pracovníkov a voľnočasovými aktivitami.

Účastníci štúdie boli rozdelený do low-effort a high-effort skupín podľa vyťaženosti v práci. Šúdie sa zúčastnilo 676 pracovníkov stredne veľkej univerzity v Dánsku. Validných pracovníkov bolo však len 93. Štúdie sa totiž mohli účastniť len ľudia splňujúci nasledujúce podmienky:

1. Nemali prácu mimo univerzitu na ktorej výskum prebiehal (z dôvodov zníženia variacie pracovných aktivít vrámci akceptovateľných limitov).
2. Žili s partnerom ktorý pracoval aspoň 2.5 dňa týždenne (kvôli zvýšeniu pravdepodobnosti, že účastníci vykonávali aspoň nejaké domáce povinnosti vo voľnom čase).

Všetci účastníci vypĺňovali v priebehu 7 dní každodenný dotazník.

Výsledkom štúdie boli 3 zistenia. V porovnaní s low-effor skupinou, high-effort skupina:

1. menej často praktikovala aktívne voľnočasové aktivity počas týždňa a mali častejšie cezčasy cez víkend.
2. považovali obe, pracovné aj domáce aktivity, za viac náročné ale menej uspokojivé.
3. hlásili častejšie problémy so spánkom počas pracovného týždňa, zvýšenú únavu počas dňa, viac zaoberania sa prácou (počas pracovného týždňa) a nižšiu motiváciu začať nasledujúci pracovný týždeň počas víkendu.

Aj keď samotný dataset bol primalý na vyvodzovanie záverov, môžeme predpokladať, že vyššia vyťaženosť v práci môže spôsobovať demotiváciu a nižšiu výkonnosť v práci.

Ak by tento predpoklad platil, dal by sa DF využiť aj na zvýšenie efektivity rozloženia práce a tým pádom aj produktivity v práci.

Existuje korelácia medzi psychickým stavom a dennou aktivitou jedinca

Predpoklad, že na základe pohybových dát sme schopný predikovať psychický stav jedninca je dobre podložený štúdiou „Comparison of Night, Day and 24 h Motor Activity Data for the Classification of Depressive Episodes“[20] z roku 2020. Z datasetu čisto aktigrafických dát sa podarilo predikovať depresívne epizódy počas dňa s obrovskou presnosťou. Senzitivita sa pohybovala medzi hodnotami 98.24% a 99.37% a rozsah špecifickosti, klasifikácie depresívnych epizód s presnosťou 99.72%, sa pohyboval medzi hodnotami 98.08% a 99.31%.

V štúdií sledovali okrem dennej aktivity aj nočnú aktivitu ako napríklad nekľudný spánok.

Na základe týchto zistení môžeme tvrdiť, že je možné predikovať psychický stav na základe aktigrafických dát z náramkového akcelerometru.

2.3.2 Hypotézy

Pred vypracovaním práce bolo treba vytvoriť set hypotéz, ktoré sa budeme snažiť dokázať. Tieto hypotézy sú postavené na predchádzajúcich predpokladoch.

Síce na prvý pohľad vyzerajú ako jednoduché úlohy, je treba k možným výsledkom pristupovať skepticky. Jedným z dôvodov je fakt, že počas štúdie prebiehala globálna covid pandémia. To znamená, že veľkú časť času, keď boli dáta zbierané, bolo veľa študentov a zamestnancov na homeoffice (ďalej HO). Taktiež veľa ľudí opúšťalo svoje domovy minimálne, či už kvôli obmedzeniam, strachu alebo chorobe.

Je teda veľká pravdepodobnosť, že dáta zaznamenávajúce voľný čas a prácu/štúdium budú veľmi podobné pre väčšinu účastníkov štúdie. Predpoklad, že bude dosiahnutý štatisticky významný výsledok v akejkoľvek z hypotéz je až priveľmi optimistický a preto nie je uvažovaný.

Je možné rozlíšiť pracovné a voľné dni na základe aktigrafických dát

Témou rozlišovania pracovných a voľných dní na základe aktigrafických dát, sa pred touto prácou nikto nezaoberal.

Ale z predošlých zmienených štúdií vieme, že sa obecne dajú denné aktivity klasifikovať s pomocou aktigrafických dát s vysokou presnosťou, v prípade, že je využitý správny model. Tieto aktivity musia byť klasifikované v obecných kategóriach ako sú nízko energetické aktivity ako sedenie, ležanie, fyzicky nenáročné aktivity ako chôdza a fyzicky náročnejšie aktivity ako športové aktivity. Toto obecné rozdelenie nie je len z dôvodu zjednodušenia klasifikácie ale aj z dôvodu charakteristiky aktigrafických dát. Keďže sa nejedná a raw triaxiálne dáta ale o ich sumu za určitý časový úsek, tak je možnosť sledovať presné pohyby ruky za určením konkrétnej činnosti prakticky nemožné.

Naša hypotéza nepracuje priamo s faktom, že budeme schopný rozlíšiť rôzne aktivity, ale primárne s faktom, že budeme schopný s využím správnych príznakov vytvoriť individuálne klasifikačné modely, ktoré budú schopné identifikovať a rozlíšiť u jedinca vzory pracovných a voľných dní.

Je možné rozlíšiť pracovné a voľné dni na základe dát zozbieraných z mobilných zariadení

Podobne ako z aktigrafických dát, tak predpokladáme, že aj z dát z mobilného telefónu budeme schopný rozlišovať voľné a pracovné dni.

Najväčší rozdiel medzi aktigrafickým náramkom a mobilným telefónom je možný počet odložení a výpadkov dát mobilného zariadenia.

Na druhú stranu nám ale mobilné zariadenia dávajú možnosť sledovať aktivitu obrazovky a iné senzory, ktoré môžu poskytnúť lepší kontext a viac dát na hľadanie vzorov (viac v Kapitole 3).

Rozlíšenie pracovných a voľných dní bude mať najlepšie výsledky pri kombinácií aktigrafických dát s dátami z mobilného zariadenia

Hypotéza, že využitím kombinácie aktigrafických dát a dát z mobilného zariadenia dosiahneme najlepšie výsledky, by mala byť najjednoduchšia na dokázanie. Aj napriek faktu, že kvôli okolnostiam, v ktorých sa všetci nachádzali počas zberu dát (HO, izolácia od ľudí), je pravdepodobné, že dosiahne táto metóda najlepšie výsledky.

Môžeme totiž predpokladať, že počet odomknutí a uzamknutia mobilného zariadenia bude korelovať s pracovnou, voľnočasovou alebo školskou aktivitou. Či už z dôvodov, že je mobilné zariadenie odložené alebo používane pri danej činnosti. Využívanie mobilného zariadenia počas rôznych činností bude s najväčšou pravdepodobnosťou veľmi individuálne. Niektorí jedinci môžu kľudne prokrastinovať počas práce/štúdia a iný zas opačne.

Ďalšie dáta, ktoré by mohli byť na klasifikáciu pracovných a voľných dní nápomocné sú GPS dáta. Hoci predpokladáme, že väčšina subjektov pracovala/študovala doma, nemôžeme vylúčiť fakt, že ľudia chodili minimálne vo voľnom čase na nákupy, prechádzky poprípade športovať mimo miesto bydliska.

Kapitola 3
Metodológia

V tejto kapitole sa nachádza popis použitých platforiem, ich porovnanie a organizácia štúdie vrátane výčtu platforiem, ktoré boli použité po pilotnom testovaní. Samotné metódy použité pri analýze sa nachádzajú v kapitole 6.

3.1 Prehľad platforiem

Pred začiatkom štúdie boli vybrané tri platformy na zber DF dát. Jedna z nich bola platforma Mindpax M0, ktorá už bola odskúšaná ako spoľahlivý zdroj aktigrafických dát.

Ďalšie dve platformy boli Beiwe[21] vyvynutá týmom Onnela lab[22] a platforma mindLAMP[23] od tímu lokalizovanom v nemocnici Beth Israel Deaconess Medical Center (ďalej len BIDMC). Tieto platformy sa špecializujú na zber dát z mobilných zariadení.

Beiwe aj mindLAMP boli založené Samom Onnelom.

3.1.1 Beiwe

Beiwe je projekt tímu Onnela Lab[22] zo zdravotníckej univerzity Harvard T. H. Chan. Tým Onnela Lab vedie professor bioštatistiky Dr. JP Onnela.

Platforma Beiwe je už relatívne stará a jej podpora/vývoj nie je v najlepšej forme, zato je stabilná a dobre zdokumentovaná.

Samotná platforma sa pyšní zberom dát z Apple aj Android mobilných zariadení a možnosťou vytvárania viacerých druhov dotazníkov pre pacientov.

Okrem klasických textových dotazníkov, kde si človek vyberá odpovede na otázky alebo ich sám píše, aplikácia ponúka aj odosielanie zvukových záznamov.

Aplikácia pre iOS má veľmi estetický dizajn zatiaľ čo Android verzia je pomerne dosť strohá.

Nasadenie

Nasadenie a údržbu platformy Beiwe mal na starosti M. Kubina. Preto v tejto práci bude nasadenie opísané iba stručne.

Platforma Beiwe sa dá nasadiť na ön-premises"server alebo na Amazon Web Services (ďalej len AWS). Avšak ön-premises"deployment nie je zdokumentovaný, takže bola platforma Beiwe nasadená na AWS servery.

Dáta

S pomocou beiwe2 mobilnej aplikácie je možné zbierať rôzne druhy dát. Dotazníky (aktívne dáta) majú niekoľko rôznych foriem. Senzorické (pasívne) dáta, je možné vyberať z niekoľkých rôznych možností, avšak niektoré druhy pasívnych dát sú platformovo závislé, alebo sa líši ich formát medzi Android a iOS zariadeniami.

Aktívne[24] a pasívne[25] dáta, ktoré je možné zbierať za pomoci aplikácie Beiwe2 sa nachádzajú v tabuľke 3.1.

Aktivita	Pasívne dáta
Voice survey	Accelerometer
Multiple choice survey	GPS
Free text survey	Power state
Slider survey	Reachability (iOS)
Checkbox survey	Magnetometer (iOS)
	Device motion (iOS)
	iOS log (iOS)
	Bluetooth (A)
	Wifi (A)
	Android log (A)

Tabuľka 3.1: Tabuľka aktivít a pasívnych dát, ktoré je možné zbierať s pomocou aplikácie Beiwe2. Legenda: (iOS) = iOS-only funkcia, (A) = Android-only funkcia.

Je možné vybrať, ktoré druhy dát sa zbierajú a nastaviť vzorkovaciu frekvenciu zberu dát.

Aplikácia (aspoň do určitej verzie) podporovala zber dát na pozadí aj keď aplikácia nebežala explicitne. Tento exploit starších SDK využitých pre vývoj Beiwe2 nám dovoloval získavať viac kompletné dáta bez toho aby bolo treba väčšinu užívateľov často upozorňovať na fakt, že majú vypnutú aplikáciu.

Ak nie je zariadenie využívajúce Beiwe2 aplikáciu pripojené k internetu, aplikácia je schopná krátkodobého ukladania dát, dokým sa zariadenie opäť nepripojí k internetu.

Zhrnutie

Platforma Beiwe spolu s aplikáciou Beiwe 2 nie je najprívetivejšia pre uživateľov operačného systému Android (ďalej len Android/Android OS). Napriek tomu je veľmi stabilná a umožňuje nastaviť presnú frekvenciu zberu dát (vzorkovania), čím je možné prispôsobiť zber rýchlejších zariadení ostatným. Táto charakteristika robí z beiwe platformu, ktorá značne zjednodušuje preprocessing dát (keďže ich nie je potreba normalizovať naprieč zariadeniami).

3.1.2 mindLAMP

Projekt mindLAMP je nová platforma, ktorá vznikla pod krídlami Department of Digital Psychiatry v BIDMC, ktoré je spolupracovníkom Harvard Medical School. Toto oddelenie bolo založené v roku 2018[26].

Platforma mindLAMP je neustále vo vývoji a má veľký potenciál, ale stále nie je dokončená a jej dokumentácia nie je kompletná a obsahuje chyby. Napriek tomu má veľmi aktívnu a nápomocnú podporu a vývoj sa posúva dopredu rýchlym tempom.

Platforma ponúka niekoľko rôznych druhov aktivít, z ktorých mnohé sú vo vývoji a budú vypustené v budúcnosti. Narozdiel od Beiwe sa bežné dotazníky nedelia na niekoľko druhov, ale priamo otázky dotazníkov. Medzi tieto otázky patria otázky s multiple choice, slidermi, výberom času alebo vlastnými odpoveďami.

Podobne ako Beiwe poskytuje platforma zber mnohých pasívnych dát a v dokumentácií sa pyšní širokou škálou senzorov a dát. Dokonca podporuje third party smart hodiniky a trackeri, pokiaľ je na mobile nainštalovaný požadovaný software.

Nasadenie

Nasadenie platformy mindLAMP mal na starosti autor tejto práce, Ondrej Sakači.

Platforma bola nasadená „on-premises“ na univerzitnom servery.

Detaily nasadenia platformy sa nachádzajú v Doplnku 4.

Dáta

Aktívne dáta platformy mindLAMP sú veľmi pestré a neustále sa menia.

Každý dotazník sa dá zložiť z niekoľkých rôznych otázok zahŕňajúcich napríklad: výber času, slider, výber jednej z niekoľkých odpovedí, výber viacerých odpovedí a pod.

Okrem dotazníkov medzi aktivity patria ale aj rôzne minihry.

Zdrojové kódy niektorých z nich sa dajú nájsť na githube[27], no existuje aj možnosť vytvoriť si vlastné aktivity a nasadiť ich na platforme.

Mimo vyššie uvedené aktívne dáta je platforma schopná zbierať pasívne senzorické dáta.

Prehľad pasívnych dát zbieraných platformou sa nachádza v tabuľke 3.2. Je ale treba poznamenať fakt, že vďaka neustálemu a dosť rýchlemu vývoju sa zoznam pasívnych dát neustále mení spolu s dokumentáciou[28].

Pasívne dáta	Špecifikácie	Pasívne dáta	Špecifikácie
accelerometer	-	respirator rate	(E)
blood pressure	(E)	screen state	-
bluetooth	(A)	sleep	?
calls	-	sms	?
device motion	-	steps	?
distance	?	weight	(M)
flights climbed	?	height	(M)
GPS	-	wifi	(A)
gyroscope	-	workout segment	?
heart rate	(E)	magnetometer	-

Tabuľka 3.2: Tabuľka pasívnych dát, ktoré je možné zbierať s pomocou aplikácie mindLAMP 2. Legenda: (iOS) = iOS-only funkcia, (A) = Android-only funkcia, (E) = je potrebný externý monitor, (m) = manuálne zadaná hodnota, - = dostupné pre obe platformy, ? = neúplná dokumentácia.

Narozdiel od platformy Beiwe2 neumožňuje[1] mindLAMP 2 nastavovať frekvenciu zbierania pasívnych dát (vzorkovania). To bolo v čase výberu platformy na zber mobilných dát vnímané ako veľký nedostatok.

Defaultne nastavená frekvencia zberu dát zo senzorov na najvyššiu možnú, sa drastický líši medzi mobilnými zariadeniami a pri novších zariadeniach mohla byť jednou z príčin vyššej spotreby energie[2].

Aplikácia má tiež nespoľahlivé cachovanie dát. V priebehu pilotného testovania sme neboli schopný odhaliť príčinu prečo sa dal uložiť len zanedbateľný úsek dát. Samotná dokumentácia o inštalácií aplikácií na iOS a Android OS[29] sa zmieňuje o tom, že pre spoľahlivý chod aplikácie je potrebné mať prístup k neustálemu internetovému pripojeniu, ideálne wifi.

Na druhú stranu platforma má výhodu vo vysokej personalizovateľnosti aktivít pre užívateľov a je schopná zbierať viac druhov dát vrátane dát z externých monitorov.

Zhrnutie

Platforma mindLAMP 2 je nová a rýchlo sa rozvíjajúca platforma. Má veľa funkcií a možností využitia. Napriek svojím výhodám má stále veľa nedostatkov, ktoré by mali v blízkej budúcnosti byť eradikované.

Podpora zberu dát z externých monitorov a third-party náramkov ako napríklad smart hodinky Mi Band[4] je veľkou výhodou oproti platforme beiwe2.

3.1.3 Mindpax M0

Platforma M0 od českej firmy Mindpax[11] sa zameriava na zber aktigrafických dát zo špecializovaných aktigrafických náramkov na zápästie.

[1] V čase konca štúdie a písania tejto práce.

[2] Toto tvrdenie nebolo možné v pilotnej fáze testovania platformy potvrdiť ani vyvrátiť, keďže neskoršie verzie nám neumožnili meniť vzorkovaciu frekvenciu.

M0 zatiaľ slúži primárne k výzkumnej činnosti s pacientami s bipolárnou poruchou.

Hlavným cieľom je pomôcť pri predikcií depresívnych a manických epizód na základe pohybovej aktivity a byť doplnkovým zdrojom informácií pre odborného lekára pacienta.

Nasadenie

Platforma je prevádzkovaná firmou Mindpax. Jediná inštancia backendu aj prístup k dátam je možný jedine na serveroch vlastnených touto firmou. Preto táto platforma nebola nasadená vrámci semestrálneho projektu ani počas práce na bakalráskej práci na žiaden školský alebo iný server.

Dáta

Dáta zbierané s pomocou platformy M0 sú taktiež aktívne aj pasívne.

Aktívne dáta sú dotazníky, manuálne alebo odborníkom zadaná medikácia spolu s denným dávkovaním, udalosti a rôzne iné manuálne vložiteľné dáta či už užívateľom alebo jeho doktorom.

Pasívne dáta sú iba raw aktigrafické dáta zo špeciálnych náramkov.

Zhrnutie

Platforma M0 je výborným zdrojom aktigrafických dát a umožňuje zber dotazníkov, pri ktorých sme nezaznamenali problémy s notifikáciami tak časté ako u ostatných platforiem.

3.2 Štúdia

V tejto časti sa nachádza stručné zhrnutie organizácie štúdie. Zhrnutie účasti, formy štúdie, použitých platforiem a ďalších informácií.

3.2.1 Použité platformy

Po pilotnom nasadení platforiem mindLAMP 2 a Beiwe2 sme sa rozhodli využiť v štúdií iba platformu Beiwe2.

Hlavné dôvody prečo staršia a menej podporovaná platforma vyhrala nad novou a rýchlo sa rozvíjajúcou boli:

- Možnosť nastavenia vzorkovacej frekvencie (v čase priebehu štúdie platforma mindLAMP 2 túto možnosť nepodporovala).
- Väčšia stabilita na väčšine zariadení.
- Schopnosť zbierania dáta aj keď aplikácia nebeží na pozadí.
- Lepší caching dát (u platformy mindLAMP 2 nespolahlivý caching).

- Aj vďaka cachingu nepotrebuje platforma Beiwe2 neustále pripojenie k internetu.

Druhá využitá platforma bola M0.

Táto platforma slúžila na zber aktigrafických dát. Na štúdiu bolo zapožičaných cca 21 aktigrafických náramkov.

3.2.2 Forma štúdie

Štúdia bola plánovaná na viac ako jednomesačný zber dát, no v priebehu sme sa rozhodli predĺžiť zber o ďalší mesiac (len s účastníkmi, ktorí boli ochotný pokračovať dlhšie ako bolo dohodnuté).

Štúdia bola anonymizovaná aby sa predišlo zneužitiu dát sledovaných subjektov.

Každý účastník používal neustále aplikáciu M0 a Beiwe2. Účastníci boli tiež povinný nosiť aktigrafický náramok po celý čas štúdie s výnimkami len pri sprchovaní[3].

Okrem zbierania aktigrafických a pasívnych dát mali účastníci za úlohu vyplňovať každý deň denný dotazník nálady skladajúci sa z 7 otázok[4] a každý týždeň týždenný dotazník nálady.

Týždenný dotazník bol dotazník ASERT[30], ktorý má za úlohu detekovať relapsy depresívne a manické. Tento dotazník sa skladá z desiatich otázok na celkovú náladu za uplynulý týždeň.

Denný dotazník bol navrhnutý týmom pracujúcim na tomto projekte.

Obsahuje otázky[5] o začiatku a konci nočného spánku, nálada, energia a stress v daný deň, dĺžka športovej aktivity[6] počas predchádzajúceho dňa a druh predchádzajúceho dňa.

Druh predchádzajúceho dňa môže byť: Voľný deň (ďalej VD), pracovný deň[7] (ďalej PD), pracovný deň doma (čiže homeoffice, ďalej HO) a mix homeofficu a normálneho pracovného dňa (ďalej PD/HO).

3.2.3 Účastníci štúdie

Prvotnými účastníkmi štúdie boli členovia týmu, zvyšok účastníkov bolo naverbovaných na ústne.

Pred začiatkom štúdie bolo cieľom nájsť dostatok dobrovoľníkov, aby celkový počet fungujúcich účastníkov bol aspoň 20.

Účastníci štúdie museli byť mentálne zdraví a svojprávni jedinci.

[3] Náramky sú vodotesné, takže sa jedná skôr o QoLB.1 možnosť, než povinnosť.

[4] Formát sa opravoval po začatí štúdie ale ustálil sa asi po 2 týždňoch od začiatku štúdie.

[5] Konečná verzia dotazníku.

[6] Športová aktivita bola definovaná ako akákoľvek fyzická aktivita, ktorú subjekt vykonáva čisto za účelom jej vykonávania a/alebo za účelom zlepšenia zdravia. To znamená, že prechádzky sa dajú brať ako športová aktivita ak ich vykonáva osoba cielene za účelom ich vykonávania kvôli zdraviu a pohybu, nie ak sa jedná o prechádzku za účelom nákupu a podobne. Tým pádom utekanie za MHD sa tiež ako športová aktivita neráta.

[7] Alebo študijný deň.

Vstup do štúdie

Vstúpiť do štúdie mohli iba jedinci netrpiaci diagnostikovanou depresiou, bipolárnou poruchou a pod.

Za účelom kontroly duševného zdravotného stavu bol každý záujemca podrobený pred prijatím do štúdie upravenou a preloženou verziou štandardného dotazníku M.I.N.I.[31]. V tejto verzií chýbala väčšina pre nás nepodstatných sekcií a zostali len sekcie týkajúce sa depresií, mánie a suicidality. Takisto sme pridali vlastné personálne otázky vzťahujúce sa na zamestanie/štúdium, vek, užívanie návykových látok, pravidelnú športovú aktivitu a pod.

Zo všetkých zájemcov mali všetci dostatočné výsledky pre prijatie do štúdie.

Následne všetci prijatý účastníci dostali na podpísanie písomný informovaný súhlas o zbere dát, aktigrafický náramok, prihlasovacie údaje do aplikácie M0 a Beiwe2 a bol inštruovaný ako si tieto aplikácie sprevádzkovať.

Štúdie sa celkovo zúčastnilo 20 účastníkov.

Kapitola 4

Nasadenie platformy LAMP

Prieskum možností nasadenia, využitia a vývoja platformy mindLamp, spolu s jej nasledovným pilotným nasadením mal na starosti autor tejto práce. V tejto Kapitole 4 sa budeme zaoberať práve nasadením a celkovým prehľadom platformy LAMP v čase vzniku tejto záverečnej práce.

Platforma mindLAMP je stále v aktíávnom vývoji. Vďaka tomu je veľká pravdepodobnosť, že prerekvizity, postup a aj práca s platformou sa od doby napísania tejto práce zmenili.

4.1 Možnosti nasadenia

Platformu LAMP je možné v súčasnosti nasadiť dvoma spôsobmi: on-premises[1] alebo na platforme AWS[2]. Náš tým sa rozhodol z finančných, praktických a bezpečnostných dôvodov využiť deployment on-premises.

4.2 Nasadenie „On-Premises“

Nasadenie on-premises bolo uskutočnené na školskom serveri s odbornou asistenciou serverového administrátora Ing. Jiřího Wilda, Ph.D.. V čase pilotnej prevádzky sa na nasadení a dopomoci s prevádzkou platformy čiastočne podieľali aj kolegovia L. Sláma a E. Žíla.

Primárne dôvody rozhodnutia nasadania on-premises boli:

1. Financie - Deployment na platforme AWS nie je najlacnejší. Snažili sme sa teda ceny tohoto projektu držať na minime a ideálne sme chceli všetko bez zbytočne veľkých výdavkov.[3].

2. Bezpečnosť - Ochrana osobných a citlivých údajov zbieraných z osobných mobilných zariadení subjektov. Síce veríme, že dáta udržované na AWS serveroch by boli dobre chránené, no zároveň by boli v podstate mimo náš dosah a v držaní inej spoločnosti. Spolu s týmom sme sa zhodli na

[1] Deployment na vlastnom serveri)
[2] Amazon Web Services[32].
[3] Z technických dôvodov nebolo možné platformu Beiwe nasadiť inak než na AWS.

tom, že radšej budeme mať dáta na školských serveroch než riskovať ich vlastníctvo inou spoločnosťou.

3. Praktickosť - V tomto prípade sa jedná o komplikovanejší bod. AWS a nasadenie na ňom je jednoduchšie, hoci Docker nasadzovanie a údržbu servera on premises veľmi uľahčuje. Teoreticky aj údržba na AWS by mohla byť jednoduchšia, ale už len fakt, že mal prístup k on-premises serveru každý člen tímu a ľudia mimo tým (ako pán Wild), značne uľahčila riešenie problémov týkajúcich sa deploymentu.

Po rozhodnutí vyskúšať on-premises deployment nasledovalo samotné nasadenie a pilotná prevádzka.

4.2.1 Prerekvizity

Medzi hlavné prerekvizity pre správne fungujúci „On-Premises“ server patria:

1. Dostačujúce systémové požiadavky[4] - Tieto požiadavky sa menia podľa počtu užívateľov. Minimálne požiadavky uvedené v dokumentácií[33] BIDMC tímu sú nasledovné:
   - Dvojjadrový procesor
   - 2 GB RAM
   - 250 GB HDD – 4 TB SSD[5]
   - Privátny sieťový endpoint s priepustnosťou 1 Gbps – 10 Gbps
2. SSL certifikát - Potrebný pre používanie https protokolu. Aplikácia je funkčná aj bez SSL certifikátu, a je možné do nej pristupovať, avšak nie je možné sa na server pripojiť mobilnou aplikáciou. Tzn. nie je možné zbierať dáta.
3. Správne nastavenie firewallu a routing dát cez porty 80 (http) a 433 (https).
4. Nainštalovaný Docker.
5. Konfigurovaný Docker Swarm cluster.

Všetky potrebné a aktualizované prerekvizity sú uvedené v oficiálnej dokumentácií[34].

[4] Počas pilotného testovania neboli robené záťažové testy na zistenie presných doporučených špecifikácií. Nie len z časových dôvodov, ale aj preto, že odhad pri našom nízkom počte testujúcich uživateľov by nemusel byť dostatočne presný.

[5] Podľa počtu užívateľov. Podľa dokumentácie sa cena za prevádzkovanie vlastného serveru môže pohybovať medzi 35 – 1500 $/mesiac.

4.2.2 Postup

Nasadenie platformy je jednoduché s niekoľkými nepovinnými krokmi, ktoré majú však za účel zlepšiť „Quality-of-Life“ pre správcu aplikácie (napr. Cloud Mesh Router).

Anekdotálne skúsenosti ale dokazujú, že nie všetky nepovinné kroky by sa mali preskakovať. Pri nasadzovaní sme čelili problémom so spustením a prevádzkou aplikácie až dokým sa nenainštaloval Portainer, ktorý je v dokumentácií opisovaný primárne ako „troubleshooting diagnostic tool“.

Veľa prvotných pokusov o nasadenie platformy skončilo neúspešne. Nakoniec sa deployment oficiálneho image serveru s pomocou yaml súboru ukázal ako najjednoduchšia a najefektívnejšia cesta k funkčnému backendu.

Celé nasadenie je opísané v oficiálnej dokumentácií[35].

4.2.3 Prevádzka

Deployment spolu so všetkými komplikáciami zabral asi dva mesiace. Server bol úspešne nasadený okolo 20. decembra 2020.

Pilotná prevádzka trvala pár týždňov. Po rozhodnutí využiť Beiwe ako primárnu platformu nebol ale tento server úplne odstavený a bol udržovaný z dvoch prípadov.

Zaprvé bolo v najväčšom záujme sledovať updaty a novinky LAMP týmu. Primárne za účelom prehľadu o stave platformy a možnej implementácií featúr, ktoré nám pri rozhodovaní medzi platformami chýbali.

Druhý dôvod prevádzky aj po ukončení pilotného testovania, bola mitigácia prípadného zlyhania platformy Beiwe.

4.3 Práca s platformou LAMP

Samotná plateforma sa skladá zo serverového backendu, databázy a mobilných aplikácií pre iOS a Android zariadenia. Ku všetkým trom miestam sa dá pristupovať rôznymi spôsobmi.

Webová aplikácia - Dashboard

Prístup k aplikácií a databázy za administrtívnymi účelmi pacientských účtov, účtov výskumníkov, výbere senzorov, ktoré budú zbierať dáta, tvorbe aktivít a kontrole aktivity pacientov zabezpečuje dashboard[36], front-endová webová aplikácia. Hlavná stránka dashboardu ponúka, mimo pripojenia sa na konkrétny server hostujúci platformu mindLAMP, možnosť spustiteľného dema aplikácie. V tomto deme je možné si prezrieť aplikáciu a zároveň vyskúšať si navigáciu po aplikácií a niektoré základné funkcie, ktoré ponúka.

Dashboard umožňuje prístup k serveru aj v prípade, že nemá platný SSL certifikát, no za týmto účelom je potrebné v prehliadači povoliť navštevovanie danej http domény.

LAMP-API

Prístup k či už pasívnym alebo aktívnym dátam uložených v databázy je možný použitím špecializovaného LAMP API[37] (Application Programming Interface). S využitím tohoto API je, po nadviazaní autorizovaného pripojenia, možné odosielať serveru querries za účelom získavania rôznych druhov dát od ľubovoľných[6] pacientov.

Od pilotnej prevádzky bola dokumentácia API zlepšená vo veľa smeroch. Asi najvýraznejšia zmena sa týka dokumentácie dátových typov[38].

Okrem stručnej inštruktáže pre začiatok používania programovacieho jazyka Python[39] na analýzu dát existuje aj obšírnejšia dokumentácia pre programovací jazyk R[40].

V prípade chýb alebo chýbajúcich informácií (ktorých počas pilotnej prevádzky bolo veľa) sa dá vždy obrátiť na aktívne a nápomocné fórum projektu[41].

Mobilná aplikácia

Mobilná aplikácia (v čase písania tejto práce vo verzií 2) má, na rozdiel od aplikácie beiwe2, rovnaký dizajn pre iOS aj Android zariadenia. Taktiež zbiera viac druhov dát a umožňuje vytváranie rozmanitých aktivít a dotazníkov.

Zber dát je avšak nastavený na maximálnu možnú frekvenciu zariadenia, čím sa zvyšuje záťaž mobilného zariadenia, znižuje výdrž batérie a dáta sú naprieč rôznymi zariadeniami neuniformné.[7] Táto neuniformita vytvára potrebu pre zložitejší preprocessing dát kde je potreba zjednotiť vzorkovanie dát pre následnú analýzu.

Taktiež v čase testovania mala aplikácia takmer mizivé cacheovacie možnosti a musela byť neustále pripojená na internet. Okrem dodatočnej záťaži zariadenia a zníženia výdrže batérie, je tento fakt negatívny aj z dôvodu, že nemôžeme predpokladať schopnosť všetkých uživateľov byť neustále na celulárenj sieti, v prípade, že sa dostanú mimo dosah WiFi pripojenia.

Čo sa samotnej záťaže batérie týka, presné testovanie rozdielov spotreby neboli namerané počas pilotnej prevádzky, no dopady boli znateľné pre všetkých testujúcich. Samotná spotreba a vyťaženie mobilného zariadenia bolo sledované po niekoľko dní, no bez exaktných meraní.

Anekdotálny príklad: Autorovi sa na svojom mobilnom zariadení Xiaomi MI Max 2 (Andriod) po vypnutí všetkých obmedzení aplikácie, ktoré aplikácií bránia v správnom chode, podarilo dostať mindLAMP na prvé miesto v rebríčku spotreby dát. Tu prekonala spotrebou batérie obrazovku štvornásobne (obrazovka ≈11%, mindLAMP ≈44%). Priemerné hodnoty spotreby batérie mindLAMP aplikáciou v čase pilotného testovania sa pohybovali v rozmedzí 25% – 57%.

[6] Ku ktorým má daná autorizovaná osoba prístup v systéme.

[7] Featura na nastavenie konkrétnej vzorkovacej frekvencie pre všetky zariadenia je v pláne už dlho. V čase písania tejto práce funguje aspoň možnosť obmedziť plošne množinu senzorov, ktoré zbierajú dáta, čím sa dá aspoň mierne regulovať záťaž zariadení.

Kapitola 5
Implementácia

V tejto kapitole sa nachádzajú všetky podstatné údaje o implementácií preprocessingu, processingu a vizualizácie dát. Bloky ukážkového kódu budú referencované z Doplnku A.

5.1 Základné informácie o implemntácií

Celé riešenie bolo implementované v programovacom jazyku Python 3.9 s použitím prostredia Jupyter-Lab[1].

Python3.9 bol vybraný pretože autor práce s ním má najväčšie skúsenosti[2] a pretože existujú balíčky scikit-learn[42] na analýzu a processing dát.

Prostredie Jupyter-Lab bolo vybrané primárne kvôli formátu Jupyter-Notebook (ipynb), ktorý umožňuje spúšťať menšie skripty nezávisle na sebe a v rôznom poradí, ale aj priamo celý notebook chronologicky. Vďaka možnosti vytvárať v notebooku bunky s Python kódom, Markdawnom alebo Raw textovým formátom zjednodušuje dokumentáciu a prehľadnosť jednotlivých skriptov.

5.1.1 Prerekvizity a použité balíčky

Okrem inštalácie Python3.9 a Jupyter-Lab boli použité nasledujúce Python balíčky.

Zo štandardnej knihovny: datetime, gc, json, IPython.display, os, glob, enum, csv, math.

Knihovny, ktoré je treba si nainštalovať[3] pre správne fungovanie implementačnej časti: scikit-learn, numpy, pandas, jupyter-lab.

[1] Primárne v textovom editore Visual Studio Code s využitím Jupyter-Lab rozšírenia, z dôvodu využitia menej pamäti RAM, než má priemerný webový prehliadač, v ktorom Jupyter-Lab beží natívne.

[2] V porovnaní s jazykom R a prostredím Matlab.

[3] Napríklad s pomocou utility pip A.5

5.2 Špecifikácie implementácie

Implementácia bola rozdelená do 2 primárnych Jupyter-Notebookov: data_preprocessing.ipynb a data_processing.ipynb. Na výpomoc pri prechádzaní už získaných výsledkov bol vytvorený results.ipynb, avšak tento notebook je v súčasnosti nefunkčný (problémy s parsingom súborov).

Väčšina skriptov a definície tried boli umiestnené v priečinku utils. Tieto skripty a dátové štruktúry boli rozdelené do python súborov: consts.py, enums.py, general_classes.py, gps.py, preprocessor.py a processor.py.

5.2.1 Dáta preprocessing

Dáta z M0 aj Beiwe2 boli uložené v csv formátoch. Tieto dáta bolo treba načítať z csv formátu a predzpracovať pred samotnou analýzou dát. Detailnejšie formáty dát M0 aj Beiwe2 sú opísané v Kapitole 6.1.

Dáta z M0 boli pre každý subjekt uložené v zložke s jeho čitateľným id[4] v štyroch súboroch. Aktidáta, dáta profilu[5], medikácie zadané subjektom[6] a dotazníky.

Dáta z Beiwe2 boli uložené v priečinkoch, z ktorých každý obsahoval dáta za jeden týždeň. V každom priečinku sa nachádzali dáta subjektu pod adresárom s názvom rovnakým ako je id účtu. Podadresáre v každom týždni sa nachádzali posledné adresáre podľa druhu dát a v každom z nich bolo veľa csv súborov.

Všetky dáta museli byť aspoň do určitej miery zpredprocesované, či už kvôli uniformite dát pre každý subjekt, alebo z dôvodu zníženia výpočtovej náročnosti pri samotnom počítaní výsledkov.

Preprocessing M0 dát

Aktidata sú už zpredprocesované dáta z trojosého akcelerometra v aktigrafickom náramku. Druhotný preprocessing M0 dát mal primárny dôvod zníženie záťaže na výpočet a RAM počítaču pri finálnom počítaní výsledkov (aktidáta).

Počas preprocessingu sa aktidáta zclustrovali do záznamov dôležitého zhrnutia dát z 15 minútových úsekov. O každých 15tich minútach sa zaznamenali základné údaje ako začiatok a koniec úseku, počet záznamov v úseku a rôzne dáta o nameraných aktigrafických hodnotách ako: suma celkovej aktivity, priemerná hodnota, medián, maximum, minimum, kurtóza[43].

Rozhodnutie zlučovať dáta na 15 minútových úsekoch, malo primárny dôvod obmedzený výpočetný výkon autorovho počítača. V ideálnych podmienkach by asi predprocessing dát bol menej drastický, no dosť možno s podobnými koncovými dátami, ktoré sa vkladali do klasifikačných modelov (Kapitola **??**).

Druhotná potreba preprocessingu prišla vo forme zpracovania dotazníkov, v ktorých subjekty denne odpovedali o svojhom dni. Nie len že bolo treba

[4] Anonymizované id.
[5] Výška, váha, id, a neinvazívne dáta, ktoré nenarušujú anonymitu dát.
[6] Drvivá väčšina súborov prázdna.

oklasifikovať dni na základe ich náplne, no aj napraviť nekonzistencie. Začiatok štúdie priniesol niekoľko rôznych opráv z zmien.

Medzi tieto zmeny patrila aj zmena formátu denných dotazníkov. Okrem dopĺňania a posúvania odpovedí v dotazníku, sa menila aj samotná náplň otázok. Po pár týždňoch sa tým rozhodol, že bude najlepšie aby otázka na náplň dňa bola zameraná na deň predchádzajúci.

Nanešťastie okrem zmien v organizácií štúdie a technických zmien nastalo aj niekoľko nečakaných technických komplikácií na serverovej strane. Niekoľko dní uprostred štúdie boli dotazníky takmer úplne znehodnotené a nesprávne uložené na servery.

Všetky zmeny v parsingu a všetky výpadky dotazníkov bolo treba v predprocesingovej klasifikácií dní na základe denných dotazníkov brať do úvahy.

Preprocessing Beiwe2 dát

Dáta nazbierané platformou Beiwe2 bolo treba prvotne načítať pre každého pacienta, zoradiť, odfiltrovať nesprávne hodnoty a spojiť do väčších súborov pre budúce jednoduchšie načítanie.

Za týmto účelom bolo treba vytvoriť špecializovaný loader súborov, ktorý prešiel všetky podadresáre patriace jednému pacientovi ak existovali a načítal z neho všetky dáta.

Dáta ktoré boli špecificky predprocesované po tomto kroku boli primárne GPS dáta. Napriek tomu, že Beiwe2 umožňuje nastaviť uniformnú vzorkovaciu frekvenciu pre zber dát na zariadení, nefunguje rovnako dobre na každom zariadení. Najznateľnejší rozdiel je pozorovateľný na rozdiely kvality a formátu dát medzi Android a iOS zariadeniami.

V GPS dátach boli najprv nájdené diery. Podľa chýbajúcich úsekov dát sa validovala kompletnosť nazbieraných dát pre každý deň samostatne. Maximum chýbajúcich dát bolo stanovené na hodnotu 25%.

Následne boli dáta unifikované. Záznamy nazbierané počas jednej sekundy boli orezané na stredný záznam z podmnožiny.

Kapitola 6
Analýza dát

V tejto kapitole sa nachádza zhrnutie formátu zpracovávaných dát a ich zpracovanie. Nachádza sa tu aj zhrnutie jednotlivých použitých klasifikačných modelov.

6.1 Formát dát

V tejto sekcií sa budeme zaoberať formátom dát aktigrafických z náramku, formát denných dotazníkov a GPS dát z mobilného zariadenia. Ostatné dáta sa v tejto kapitole nenachádzajú, pretože pre výsledky neboli využité.

Nebude sa jednať len o originálny formát dát, ale aj o formát predspracovaných dát, ktoré sa vkladali do klasifikačných modelov.

6.1.1 Actidata

V prípade aktigrafických dát boli počas preprocessingu z dát vytiahnuté charakteristické informácie o určitých množinách dát. Tým pádom sa originálny aj nový formát líši už na prvý pohľad.

Originálny formát

Formát, v ktorom sme aktidáta dostávali nie je surový, prebehol predspracovaním a záznamy o pohybe v troch osách akcelerometra boli prevedené na hodnotu activity. Vďaka tomu jediné poptrebné preprocesovanie je získanie príznakov na hodnotenie a vyradenie dní s nedostatkom nazbieraných dát.

Originálny formát aktigrafických dát sa nachádza v Tabuľke 6.1

Názov stĺpca	Datový typ	Poznámky
utc_time	string / datetime64ns	UTC čas
activity	float	Hodnota aktivity, nemá jednotky
local_time	string / datetime64ns	Lokálny čas

Tabuľka 6.1: Originálny formát aktigrafických dát.

Nový formát

Tieto dáta boli následne spracované do formátu viac pripomínajúceho príznaky, až na to, že sa jednalo o príznaky pre krátke časové úseky (15 minút).

Názov stĺpca	Datový typ	Poznámky
start_time	string / datetime64ns	Začiatok úseku
end_time	string / datetime64ns	Koniec úseku
number_of_entries	int	Počet záznamov v úseku
sum	float	Suma hodnôt activity
mean	float	Priemerná hodnota
median	float	Stredná hodnota
max	float	Maximálna hodnota
min	float	Minimálna hodnota
kurtosis	float	Kurtóza[43] hodnôt

Tabuľka 6.2: Formát aktidát po zpracovaní.

Záznamy formátu ukázaného v Tabuľke 6.2 boli následne uložené pre konečné spracovanie a výpočty.

Príznaky ktoré sa použili pre klasifikáciu dát sú často len sumy a priemery hodnút z nových záznamov.

6.1.2 Dotazníky

Dotazníky menili formát počas štúdie, tým pádom počas prípravy dát na zpracovanie nebolo treba uniformovať dáta ale správne ich prečítať.

Formáty dotazníkov sú v Tabuľke 6.3. V týchto dotazníkoch sa vyskytujú otázky na náladu (šťastný – smutný), vnútorný pocit (zrelaxovaný – úzkostný), energiu (vyčerpaný – energický).

Mimo to sa vo V2 dotazníku objavujú 2 otázky na druh dňa. Tieto otázky sa odpovedali sliderom, ktorý ukazoval na hodnotu na spektre medzi dvoma hodnotami. Tieto slidere boli vymenené za jednoduchšiu otázku so 4mi exaktnými odpoveďami (druh dňa).

Číslo otázky	Typ otázky V1	Typ otázky V2	Typ otázky V3
1	Počet hodín spánku	<*	Začiatok spánku
2	Nálada	<*	Koniec spánku
3	Pocit	<*	Nálada
4	Energia	<*	Pocit
5	-	Deň (VD – PD)	Energia
6	-	Deň (HO – PD/HO)	Druh dňa
7	-	Počet hodín športu	<*

Tabuľka 6.3: Formáty dotazníkov. Vysvetlivky: -chýbajúca otázka, «*-rovnaká otázka ako otázka naľavo.

Rozparsované dotazníky sa následne pre každý subjekt uložili do slovníkového súboru, kde kľúče boli dátumy dní a hodnotami boli číselné označenia typu dní. Tieto hodnoty sa nachádzajú v Tabuľke 6.4.

Číselná hodnota	Typ dňa
-1	Neklasifikovaný deň
0	Voľný deň
1	Pracovný deň (normálny)
2	Pracovný deň (homeoffice)
3	Pracovný deň (mix)

Tabuľka 6.4: Typy dňov

6.1.3 GPS dáta

Dáta z GPS svoj formát (Tabuľka 6.5) pri predspracovaní nemenili.

Na druhej strane ale počas predprocessingu prešli sclustrovaním na menšie a uniformnejšie datasety a prešli kontrolou kompletnosti dát.

Názov stĺpca	Datový formát	Poznámky
timestamp	timestamp / int	-
UTC time	string / datetime64ns	UTC čas
latitude	float	Zemepisná šírka
longitude	float	Zemepisná dĺžka
altitude	float	Nadmorská výška
accuracy	float	Presnosť merania

Tabuľka 6.5: Formát GPS dát.

Clustrovanie prebehlo tak, že pre každú sekundu boli vybrané všetky záznamy z tohoto úseku. Z týchto záznamov bol následne vybraný ekvivalent ich mediánu (záznam uprostred záznamov chronologicky zoradených).

6.2 Algoritmy

Na samotnú klasifikáciu boli použité tri rôzne algoritmy z balíčku sklearn[42]. Konkrétne sa jednalo o Decision tree, SVC (C-Support Vector Classification) a Gaussian Naive Bayes.

Všetky vyššie zmienené a aj mnoho ďalších algoritmov a utilít je opísaných na sklearn API Reference stránke[44].

Každý algoritmus bol prepočítaný sto-krát pre každý subjekt. Pred každým výpočtom, boli dni s príznakmi a so správnou klasifikáciou premiešané náhodne. Randomizácia a vysoký počet opakovaní zaručuje čo najväčšiu presnosť.

Pri samotnom počítaní každého z modelov, sú cross-validované (premiešané dáta podrobené niekoľko násobnou zmenou kontrolnej podmnožiny a trénovacieho setu).

6.2.1 Decision tree

Algoritmus využívajúci tzv. decision trees (rozhodovacie stromy). Jedná sa o acyklické grafy.

Na základe rôznych výsledkov (ktoré končia v listoch stromu), je vytvorení rozhodovací strom. V tomto strome odpoveď na každú otázku ohľadom datasetu rozhoduje na ktorú stranu podstromu bude pri klasifikácií pokračovať.

Nakoniec rozhodovací proces končí v jednom z listov, ktorý je predpokladaný výsledok. V tomto prípade sa jedná o klasifikačné rozhodovacie stromy.

6.2.2 SVM - Support Vector Machine

Pri použití SVM (v tomto prípade konkrétne SVC - C-Support Vector Classifiacation) sa priestor možných N-dimenzionálnych bodov predelí niekoľkými hyperplochami. Tieto body sú určené súradnicami, ktoré dostaneme z datasetu príznakov.

Následne sa pre každý bod hľadá kolmica na každú z týchto hyperplôch a smer vektoru ktorý vznikne od najbližšej plochy k bodu, ktorý je klasifikovaný, určuje v priestore množinu do ktorej tento bod spadá.

6.2.3 Naive Bayes

Posledný z použitých machine learning algoritmov je Naive Bayes. Ako už názov napovedá jedná sa o jeden z najprimitívnejších algoritmov, no zároveň najrýchlejších a prekvapivo efektívnych.

Naive Bayes prikladá každému príznaku v datasete rovnakú váhu pri rozhodovaní a zároveň ich vníma ako nezávislé (jeden príznak svojou prítomnosťou neovplyvňuje príznaky ostatné).

6.3 Spracovanie dát a výpočty

Dáta boli spracovávané niekoľkými spôsobmi pre porovnanie účinnosti rozhodovacích algoritmov nad rôznymi množinami príznakov.

Výpočty boli prevádzkované nad datasetmi príznakov vytiahnutých z aktidát, následne nad príznakmi z GPS dát a nakoniec nad kombináciami týchto príznakov.

Príznaky aktidát

Aktidáta používali ako príznaky väčšinou hodnoty ktoré sa nachádzali v predprocesovaných záznamoch. Jedná sa teda o hodnoty vypočítané každých 15 minút (suma, priemer, medián, maximum, minimum, kurtóza).

K týmto príznakom sa pridávali avšak aj celkové sumy, priemeruy, maximá a minimá hodnôt za určité denné úseky. Tieto úseky by sa mali líšiť aktivitou počas pracovného dňa a voľného.

Raňajšie hodiny by mali byť väčšinou kľudnejšie a zahrnovať z väčšej časti neaktivitu.

Denné sa budú líšiť podľa pracovnej doby/rozvrhu jedinca a náplne jeho práce[1].

Večerné hodiny by mali byť charakteristické oddychom, domácimi prácami alebo voľnočasovými aktivitami u väčšiny subjektov.

Tieto úseky boli teda nasledovné:

1. Ráno - 0:00 – 7:00
2. Pracovný deň - 7:00 – 19:00
3. Večer - 19:00 – 24:00

Maximálny počet príznakov získaných z aktigrafických dát bol za deň 588[2] a najmenší počet príznakov bol 16[3].

Ako ale bolo spomínané prebehlo niekoľko rôznych výpočtov nad rôznymi datasetmi príznakov.

1. Hlavné denné úseky
2. Pätnásť minútové úseky
3. Pätnásť minútové úseky (iba kurtóza a priemer)
4. Pätnásť minútové úseky (iba kurtóza a priemer) + Hlavné denné úseky
5. Všetky príznaky napočítané za deň

Posledné 4 druhy mali ešte variantu, v ktorej sa 15 minútové úseky spojili do hodinových úsekov čím zmenšili počet príznakov o štvrtinu (ak sa nepočítajú 3 hlavné denné podmnožiny).

Príznaky GPS dát

Príznaky zo spracovaných GPS dát bolo 24. Jednalo sa o 6 príznakov napočítaných na 4 podmnožinách GPS dát nazbieraných v daný deň.

Tieto príznaky sú: prejdená vzdialenosť, zmena výšky, maximálna výška a priemerný bod (priemerná zemepisná šírka, zemepisná dĺžka a nadmorská výška).

Podmnožiny, na ktorých sa tieto príznaky počítali boli:

1. Celý deň
2. Ráno - 0:00 – 7:00
3. Pracovný deň - 7:00 – 19:00
4. Večer - 19:00 – 24:00

[1] Kvôli nevedomosti pracovných a študijných dôb jednotlivých subjektov sú denné hodiny dlhšie než 8 hodín.

[2] 15 minútové úseky so 6 príznakmi každý, 3 denné úseky so 4 príznakmi každý.

[3] 3 hlavné denné úseky, každý 4 príznaky.

Klasifikácia

Triedy, ktoré boli použité na klasifikáciu dní, boli 4: voľný deň, pracovný deň, homeoffice, mix pracovného dňa a homeoffice.

Bola odskúšaná možnosť zlúčenia niektorých tried dokopy, tým vznikli len 2 triedy, čo by mohlo výsledky zlepšiť alebo zhoršiť. Tieto dve triedy sú nasledujúce:

1. voľný deň + homeoffice
2. pracovný deň + mix pracovného dňa a homeoffice

Validnosť datasetu príznakov

Obecne sa dataset považoval za dostatočný ak mal aspoň 30 valídnych dní. To znamená: dostatok aktidát a/alebo dostatok GPS dát, zaznamenaný druh dňa v dennom dotazníku.

No bola odskúšaná aj možnosť vylúčiť z datasetu všetky subjekty ktoré mali počet valídnych dní pod 60.

Výpočty

Všetky vyššie zmiené varianty datasetov a druhy príznakov boli kombinované medzi sebou a porovnávané.

Samotné výsledky výpočtov budú uvedené v Kapitole 7.1.

Kapitola 7

Záver

Po oboznámení sa s teóriou, obecnou problematikou, hypotézami, postupmi využitými pri spracovaní a analýze dát, sa v poslednej kapitole pozrieme na zhrnutie výsledkov tejto záverečnej práce.

7.1 Výsledky

Výsledky sú rozdelené do troch skupín: výsledky na základe aktigrafických dát, výsledky na základe GPS dát a výsledky, ktoré boli počítané z datasetov obsahujúcich príznaky spočítané z ako aj z aktigrafických, tak aj z GPS dát.

Vysvetlivky ku skratkám v ľavých stĺpcoch tabuliek 7.1 – 7.9 sa nachádzajú v zozname skratiek B.2 na konci práce.

Zároveň sú v tabuľkách 7.1 – 7.9 hrubo vyznačené najlepšie hodnoty pre každý stĺpec.

7.1.1 Aktidáta

V Tabuľke 7.1 sú priemerné hodnoty presností všetkých algoritmov.

	DT	NB	SVC
G (%)	55.563	57.369	62.588
MCG (%)	59.993	61.179	**66.240**
H_KMG (%)	58.959	57.154	62.480
H_KM (%)	57.857	55.876	64.768
H_B (%)	57.435	58.345	64.236
H_E (%)	58.502	60.116	62.650
M_KMG (%)	56.874	61.907	62.482
M_KM (%)	56.186	61.449	64.708
M_B (%)	57.026	60.142	64.370
M_E (%)	57.516	60.237	62.520
M_E_60+ (%)	61.661	63.383	63.756
MC_M_E (%)	**63.031**	**64.994**	66.026
MC_H_E (%)	62.018	63.333	66.145

Tabuľka 7.1: Priemerné výsledky algoritmov nad rôznymi množinami aktigrafických príznakov.

Tieto hodnoty boli získané spriemerovaním priemerných presností na všetkých valídnych subjektoch.

V Tabuľke 7.1 dosahoval najlepšie výsledky algoritmus SVC.

TIež vidno, že lepšie výsledky dosahovali datasety, pri ktorých boli klasifikačné triedy zredukované len na dve.

Podobne boli vypočítané aj priemerné hodnoty najhorších a najlepších výsledkov. Tie sú v Tabuľke 7.2 a Tabuľke 7.3.

	DT	NB	SVC
G (%)	19.247	26.916	30.400
MCG (%)	23.653	30.284	**38.606**
H_KMG (%)	23.307	23.854	26.516
H_KM (%)	22.676	22.719	30.611
H_B (%)	21.735	26.119	29.444
H_E (%)	21.213	27.148	26.842
M_KMG (%)	20.320	29.593	26.516
M_KM (%)	17.908	30.681	30.169
M_B (%)	21.652	28.326	29.957
M_E (%)	21.868	27.079	26.516
M_E_60+ (%)	**31.579**	**32.758**	33.049
MC_M_E (%)	22.592	32.616	34.356
MC_H_E (%)	24.409	29.860	34.358

Tabuľka 7.2: Priemerné najhoršie výsledky algoritmov nad rôznymi množinami aktigrafických príznakov.

Z Tabuľky 7.2 vidno, že algoritmus SVC má stále navrch, čo sa týka presnosti klasifikácie aj v najhorších prípadoch.

	DT	NB	SVC
G (%)	86.296	84.837	88.663
MCG (%)	90.848	90.423	89.317
H_KMG (%)	88.902	86.305	89.359
H_KM (%)	89.651	86.031	92.828
H_B (%)	89.678	87.329	91.739
H_E (%)	90.150	90.369	89.949
M_KMG (%)	90.368	89.613	89.032
M_KM (%)	89.743	90.056	**93.212**
M_B (%)	87.583	89.115	92.018
M_E (%)	87.520	90.043	89.032
M_E_60+ (%)	89.546	91.405	87.696
MC_M_E (%)	92.446	**92.963**	91.196
MC_H_E (%)	**93.301**	91.348	91.863

Tabuľka 7.3: Priemerné najlepšie výsledky algoritmov nad rôznymi množinami aktigrafických a GPS príznakov.

V Tabuľke 7.3 sa priemerná stabilita a vyššia spoľahlivosť pre aktigrafické dáta potvrdzuje. Algoritmus DT síce SVC prekonal v najlepšom ideálnom výsledku, no rozdiel nie je významný.

7.1.2 GPS dáta

GPS dáta boli druhý typ dát, z ktorých boli vytvorené príznaky, ktoré sa následne vkladali do algoritmov.

Pre zrovnanie informačnej hodnoty dát s aktigrafickými príznakmi, bolo najprv vyskúšané klasifikovanie dní na základe čisto GPS dát.

	DT	NB	SVC
G (%)	**64.467**	**56.656**	**65.130**
MCG (%)	55.613	42.114	55.900

Tabuľka 7.4: Priemerné výsledky algoritmov nad GPS príznakmi.

	DT	NB	SVC
G (%)	**33.136**	**24.277**	**36.629**
MCG (%)	25.046	10.096	27.938

Tabuľka 7.5: Priemerné najhoršie výsledky algoritmov nad GPS príznakmi.

	DT	NB	SVC
G (%)	**89.111**	**85.403**	**88.148**
MCG (%)	81.735	72.978	79.809

Tabuľka 7.6: Priemerné najlepšie výsledky algoritmov nad GPS príznakmi.

Tabuľky 7.4 – 7.6 ukazujú, že najlepší algoritmus na klasifikáciu je opäť SVC, hoci DT už moc nezaostáva. Problém je tentokrát málo datasetov, resp. jeden dataset s dvomi rôznymi klasifikáciami.

V tomto prípade zredukovanie počtu tried zo štyroch na dve paradoxne výsledky zhoršilo pre všetky algoritmy.

7.1.3 Mix aktidát a GPS dát

Na koniec prišla snaha zvýšiť presnosť klasifikačných algoritmov kombináciou GPS a aktigrafických príznakov.

Jej výsledky sú vidieť v tabuľkách 7.7, 7.8 a 7.9.

Bohužiaľ k žiadnemu znateľnému nárastu presnosti nedošlo.

V Tabuľke 7.8 sa dokonca priemerná presnosť znižuje.

Opäť sa vo všetkých tabuľkách ukazuje algoritmus SVC ako najlepšia voľba pre klasifikáciu týchto dát.

	DT	NB	SVC
G (%)	59.126	54.546	61.929
MCG (%)	62.901	60.596	65.887
H_KMG (%)	60.275	57.656	62.060
H_KM (%)	60.089	57.443	62.398
H_B (%)	60.249	59.478	64.733
H_E (%)	60.664	61.134	62.195
M_KMG (%)	58.945	62.945	63.680
M_KM (%)	59.069	63.182	62.054
M_B (%)	60.044	60.690	64.936
M_E (%)	60.191	61.206	62.119
M_E_60+ (%)	62.553	63.371	64.812
MC_M_E (%)	**63.964**	**65.189**	66.027
MC_H_E (%)	63.629	63.840	**66.149**

Tabuľka 7.7: Priemerné výsledky algoritmov nad rôznymi množinami aktigrafických príznakov.

	DT	NB	SVC
G (%)	22.420	18.983	29.418
MCG (%)	25.491	26.063	35.704
H_KMG (%)	22.601	22.135	28.545
H_KM (%)	21.214	20.835	29.322
H_B (%)	23.675	23.051	30.188
H_E (%)	23.955	26.641	28.545
M_KMG (%)	20.727	28.469	29.146
M_KM (%)	21.699	28.980	28.545
M_B (%)	20.655	26.717	31.675
M_E (%)	20.857	26.979	28.545
M_E_60+ (%)	26.751	**34.666**	**34.701**
MC_M_E (%)	24.391	33.894	34.356
MC_H_E (%)	**27.463**	31.090	34.356

Tabuľka 7.8: Priemerné najhoršie výsledky algoritmov nad rôznymi množinami aktigrafických a GPS príznakov.

	DT	NB	SVC
G (%)	89.292	88.342	90.442
MCG (%)	**93.067**	91.393	90.539
H_KMG (%)	89.595	88.575	88.856
H_KM (%)	89.540	86.693	91.245
H_B (%)	91.983	87.250	92.781
H_E (%)	91.652	89.466	90.096
M_KMG (%)	88.640	91.568	92.187
M_KM (%)	89.508	90.924	88.856
M_B (%)	89.175	89.926	**93.496**
M_E (%)	88.840	89.656	90.116
M_E_60+ (%)	90.291	89.894	90.004
MC_M_E (%)	92.563	**92.291**	91.196
MC_H_E (%)	92.483	90.770	91.863

Tabuľka 7.9: Priemerné najlepšie výsledky algoritmov nad rôznymi množinami aktigrafických a GPS príznakov.

Obecne sa vo všetkých prípadoch priemerná presnosť odhadu drží pod 70%. Nie je to ideálny výsledok, avšak je pozitívny.

Pozitívny je tiež fakt, že priemerná presnosť neklesá pod 50%.

7.2 Diskusia

Indiferentnosť výsledkov po kombinácií GPS príznakov spolu s podmnožinami aktigrafických príznakov je sklamaním, no neznačí absolútny neúspech.

Ako bolo avizované hneď na úvode práce podmienky počas štúdie neboli ideálne.

Okrem malého počtu účastníkov, relatívne krátkeho obdobia na zber dát (ktorého neboli ani všetky subjekty účastné po celý čas) a technických problémov počas zberu sa nám do cesty postavila aj pandémia vírusu Covid-19. Vďaka nej a opatreniam prijatým proti jej nekontrolovanému šíreniu, sme už pred začiatkom semestrálneho projektu predpokladali, že zozbieraný dataset nebude dobrý. Primárne kvôli zníženému pohybu osôb a zavedeniu homeoffice pracovnej doby a distančnej výuke. Pre to sme očakávali, že denná aktivita bude voľné a pracovné dni jednoznačne rozlišovať len v hŕstke prípadov.

S veľkou pravdepodobnosťou sa tento predpoklad naplnil v obecnej rovine. Za predpokladu, že vybrané príznaky by sa dali v normálnych podmienkach využiť s vyššiou presnosťou, tak takmer nezaznamenateľná zmena v presnosti medzi aktidáta a GPS príznakmi je podozrivá.

Táto skoro až identická charakteristika dát, by mohla znamenať aj to, že väčšina subjektov má drvivú väčšinu fyzickej aktivity spojenú s presunom vonku.

Z dôvodu väčšinového pobytu v domácnosti, je vysoká pravdepodobnosť aj toho, že mobil ležal drvivú väčšinu času na jednom mieste a tým pádom nezaznamenal dôležité dáta z iných senzorov (hoci s tie neboli použité pri klasifikácií voľných a pracovných dní v tejto záverečnej práci).

Okrem toho značné problémy pri zpracovaní dát robili niektoré subjekty a ich datasety. Napríklad pri testovaní redukovaných klasifikačných tried, museli byť ignorovaný z dôvodu, že jediné ich oklasifikované dni by po redukcií vytvorili iba jednu triedu.

Vyššie bola zmienená veľkosť datasetu a dĺžka štúdie. Aj v prípade, že by klasifikácia bola značne presnejšia, než je, nemohli by sme prekázať, že sme nemali len šťastie. Maximálne by sme mohli poukázať na istý potenciál, ktorý by budúci výskum mohol mať.

Doposiaľ sa nám podarilo prikloniť k prvým dvom hypotézam tejto práce: Áno je možné rozlíšiť pracovné a voľné dni na základe aktidát2.3.2 aj mobilných dát2.3.2. Avšak tretia hypotéza: Rozlíšenie pracovných a voľných dní bude mať najlepšie výsledky pri kombinácií aktigrafických dát s dátami z mobilného zariadenia2.3.2, bola aktuálnou prácou vyvrátená.

Kvôli všetkým vyššie zmieneným faktorom sa dá tvrdiť, že práca bola zakončená úspechom. Hoci nepriniesla výsledky aké by sme si želali, priniesla lepšie výsledky, než sme mohli dúfať kvôli súčasnej situácií.

Dodatok A

Časti kódu

Highlighting častí kódu bol prevzatý zo stránky tex.stackexchange.com[45].

A.1 Konštanty

Typy výsledkov A.1.

Listing A.1: Result types

```
    RESULT_TYPES = {
    1: '15min_everything',
    # ... MORE RESULT TYPES ...
    25: 'merged_classes_1hr_everything',
}
```

A.2 Preprocessing dát

Zhrnutie najdôležitejších častí kódu týkajúcich sa preprocessingu dát.

A.2.1 Trieda DataContainer

DataContainer A.2 je jednoduchá abstrakcia nad jednotlivými druhmi dát. Podľa druhu dát (name) vie ako má s daným pandas DataFramom zaobchádzať. Zprehľadňuje a zjednodušuje kód.

Listing A.2: Trieda DataContainer

```
    class DataContainer:
    def __init__(self, name: PDName):
        self.name = name
        self.data = pd.DataFrame()

    def save_to_csv(self, folder, prefix=''):
        #save function

    def load_from_csv(self, folder, prefix=''):
```

```
        #load function

    def order_data_by_date(self):
        #order function

    def cluster_data(
        self,
        minutes,
        device_type) -> pd.DataFrame:
        #cluster function
```

Jedna z najdôležitejších metód je metóda cluster_data A.3. Táto metóda clusteruje dáta podľa času do jednoduchšie spracovateľných jednotiek. Subset datasetu odošle následne do komplexnejšej metódy, ktorá už rozhodne akým spôsobom daný druh dát zkomprimovať.

Listing A.3: Metóda cluster_data

```
    def cluster_data(
        self,
        minutes,
        device_type
    ) -> pd.DataFrame:
        clustered = pd.DataFrame()
        time = pd.to_datetime(STUDY_START_DATE)
        end_time = pd.to_datetime(STUDY_END_DATE)
        period = dt.timedelta(minutes=minutes)

        if is_clusterable(self.name):
            while time < end_time:
                cluster = self.data[
                    (self.data[
                        get_time_key(self.name)
                    ] >= f'{time}') &
                    (self.data[
                        get_time_key(self.name)
                    ] < f'{time + period}')
                ]

                row = cluster_data(
                    cluster,
                    self.name,
                    device_type
                )

                if bool(row):
                    clustered = clustered.append(
                        row,
```

```
                    ignore_index=True
                )

            time += period
    return clustered
```

A.2.2 Trieda PatientData

Trieda PatientData A.4 obsahuje funkcie na load všetkých alebo jednotlyvých dát z platformy M0 a Beiwe2. Obsahuje funkcie na ich spracovanie, ukladanie a správne čítanie (napríklad druh dňa z dotazníkov denných).

Listing A.4: Trieda PatientData

```
class PatientData:
    '''Simple class for manipulation with
    all essential patient data.'''
    readable_id:    str # name of the instance

    '''Simple dataframes'''
    actidata =      DataContainer(PDName.ACTIDATA)
    # ... OTHER DATA CONTAINERS ...
    q_week =        DataContainer(PDName.Q_WEEK)

    '''Dictionaries'''
    q_daily =   {
                    "V1": pd.DataFrame,
                    "V2": pd.DataFrame,
                    "VF": pd.DataFrame
                }

    valid_days = {}
    valid_beiwe_days = {}
    day_type = {}

    def __init__(self, readable_id):
        self.readable_id = readable_id
```

A.3 Rôzne

Ukážka inštalácieA.5 nadštandardných knižníc použitých v implementácií[1].

Listing A.5: Inštalácia nadštandardných knižníc

```
pip3 install --user scikit-learn pandas numpy
```

[1]Rovnakým spôsobom sa dá nainštalovať aj Jupyter-lab.

Dodatok B

Zoznam použitých skratiek

V tomto dodatku sa nachádza tabuľka B.1 vysvetliviek všetkých skratiek. Skratky sú zoradené podľa abecedy.

Skratka	Význam
ADHD	Attention deficit hyperactivity disorder
ASD	Autism spectrum disorder
AWS	Amazon Web Services
BIDMC	Beth Israel Deaconess Medical Centre
BP	Bipolárna porucha
BT	Bluetooth
CV	Cross validation
DF	Digitálne fenotypovanie (Digital phenotyping)
DT	Decision trees
GB	Gradient boosting
GDPR	General Data Protection Regulation
GNB	Gaussian Naive Bayes
HMM	Hidden Markov's model
HO	Homeoffice
ICC	Interclass correlation
kNN	k-Nearest Neighbors
NB	Naive Bayes
OS	Operačný systém
PC	Personal computer (Stolný počítač, môže byť ale aj notebook)
PD	Pracovný deň
PR	Public Relations
SVM	Support Vector Machine
SVC	Support Vector Classification
VD	Voľný deň
VR	Virtuálna realita
QoL	Quality of Life

Tabuľka B.1: Zoznam použitých skratiek s vysvetlivkami.

Skratka	Význam
G	General - Iba príznaky hlavné denný úsekov
H_B	Hour Barebones - Príznaky z aktidát sú len z hodinových úsekov
H_E	H. Everything - Všetky príznaky z aktidát
H_KM	H. Kurtosis Mean - Iba kurtóza a priemer
H_KMG	H. K. M. General - Kurtóza, priemer a príznaky hlavných úsekov
M_E_60+	Všetky príznaky, iba subjekty s viac ako 60 valídnymi dňami
MC_H_E	Merged Classes H. E. - klasifikácia zredukovaná na 2 triedy
MC_M_E	M. C. Minutes E. - 15 minútové useky
MCG	Merged Classes General
M_B	H_B ale z 15 minútových úsekov
M_E	H_E ale z 15 minútových úsekov
M_KM	H_KM ale z 15 minútových úsekov
M_KMG	H_KMG ale z 15 minútových úsekov

Tabuľka B.2: Zoznam použitých skratiek použitých v tabuľkách s výsledkami.

Dodatok C

Literatúra

[1] Yunji Liang, Xiaolong Zheng, and Daniel D. Zeng. A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, 52:290–307, 2019. https://doi.org/10.1016/j.inffus.2019.04.001.

[2] Michael I.C. Kingsley, Rashmika Nawaratne, Paul D. O'Halloran, Alexander H.K. Montoye, Damminda Alahakoon, Daswin De Silva, Kiera Staley, and Matthew Nicholson. Wrist-specific accelerometry methods for estimating free-living physical activity. *Journal of Science and Medicine in Sport*, 22(6):677–683, 2019. https://doi.org/10.1016/j.jsams.2018.12.003.

[3] Inc Fitbit. fitbit, 2021. https://www.fitbit.com/global/eu/home Hlavná stránka produktov značky fitbit.

[4] Xiaomi. Xiaomi Mi Band, 2021. https://www.mi.com/global/miband/ Hlavná stránka smart hodiniek značky Xiaomi.

[5] Google. Google adsense, 2021. https://www.google.com/adsense/start/ Stránka Google AdSense služby.

[6] Facebook. Facebook ads, 2021. https://www.facebook.com/business/ads Stránka Facebook ads.

[7] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A. Landay, and Jeremy N. Bailenson. Personal identifiability of user tracking data during observation of 360-degree vr video. *Scientific Reports*, 10:17404, 10 2020. https://doi.org/10.1038/s41598-020-74486-y.

[8] LLC. Facebook Technologies. Quest 2, 2020. https://www.oculus.com/quest-2/ Oficiálna stránka produktu Oculus Quest 2 od firmy Facebook.

[9] Patricio Robles. Will digital phenotyping ever be applied to pharma marketing? Centaur Media plc, March 2018. https://econsultancy.com/will-digital-phenotyping-ever-be-applied-to-pharma-marketing/ Článok zamýšľajúci sa nad možnosťou využitia digitálneho fenotypingu v oblasti medicínskeho marketingu.

[10] N Martinez-Martin, H.T. Greely, and M.K. Cho. Ethical Development of Digital Phenotyping Tools for Mental Health Applications: Delphi Study. *JMIR mHealth and uHealth*, 2021. https://pubmed.ncbi.nlm.nih.gov/34319252/.

[11] Mindpax. Mindpax, 2021. https://www.mindpax.me/ Hlavná stránka Mindpax projektu.

[12] Maria Hildebrand, Vincent T VAN Hees, Bjorge Hermann Hansen, and Ulf Ekelund. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. *Medicine and science in sports and exercise*, 46(9):1816—1824, September 2014. https://doi.org/10.1249/MSS.0000000000000289.

[13] Ellis K, Kerr J, Godbole S, Staudenmayer J, and Lanckriet G. Hip and wrist accelerometer algorithms for free-living behavior classification. *Medicine and science in sports and exercise*, pages 933–40, May 2016. https://doi.org/10.1249/MSS.0000000000000840.

[14] Toby G. Pavey, Nicholas D. Gilson, Sjaan R. Gomersall, Bronwyn Clark, and Stewart G. Trost. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *Journal of Science and Medicine in Sport*, 20(1):75–80, 2017. https://doi.org/10.1016/j.jsams.2016.06.003.

[15] Toby G. Pavey, Sjaan R. Gomersall, Bronwyn K. Clark, and Wendy J. Brown. The validity of the geneactiv wrist-worn accelerometer for measuring adult sedentary time in free living. *Journal of Science and Medicine in Sport*, 19(5):395–399, 2016. https://doi.org/10.1016/j.jsams.2015.04.007.

[16] Kangjae Lee and Mei-Po Kwan. Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results. *Computers, Environment and Urban Systems*, 67:124–131, 2018. https://doi.org/10.1016/j.compenvurbsys.2017.09.012.

[17] Ivan Miguel Pires, Gonçalo Marques, Nuno M. Garcia, and Eftim Zdravevski. Identification of activities of daily living through artificial intelligence: an accelerometry-based approach. *Procedia Computer Science*, 175:308–314, 2020. https://doi.org/10.1016/j.procs.2020.07.044 The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC),The 15th International Conference on Future Networks and Communications (FNC),The 10th International Conference on Sustainable Energy Information Technology.

[18] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen. Activity classification using realistic data from wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):119–128, 2006. https://doi.org/10.1109/TITB.2005.856863.

[19] Madelon van Hooff, Sabine Geurts, Michiel Kompier, and Toon Taris. Workdays, in-between workdays and the weekend: A diary study on effort and recovery. *International archives of occupational and environmental health*, 80:599–613, 08 2007. https://doi.org/10.1007/s00420-007-0172-5.

[20] Julieta G. Rodríguez-Ruiz, Carlos E. Galván-Tejada, Laura A. Zanella-Calzada, José M. Celaya-Padilla, Jorge I. Galván-Tejada, Hamurabi Gamboa-Rosales, Huizilopoztli Luna-García, Rafael Magallanes-Quintanar, and Manuel A. Soto-Murillo. Comparison of night, day and 24 h motor activity data for the classification of depressive episodes. *Diagnostics*, 10(3), 2020. https://doi.org/10.3390/diagnostics10030162.

[21] Onnela Lab. Beiwe research platform, 2021. https://www.beiwe.org/ Hlavná stránka Beiwe platformy.

[22] Harvard T.H.Chan. Onnela Lab, 2021. https://www.hsph.harvard.edu/onnela-lab/ Hlavná stránka Onnela Labtýmu.

[23] Division of Digital Psychiatry BIDMC. Lamp - digital psych, 2021. https://www.digitalpsych.org/lamp.html Hlavná stránka LAMP projektu.

[24] Onnela Lab. Wiki - Active Data, 2019. https://github.com/onnela-lab/beiwe/wiki/Active-Data Wiki stránka Beiwe platformy so zhrnutím aktívnych dát.

[25] Onnela Lab. Wiki - Passive Data, 2018. https://github.com/onnela-lab/beiwe/wiki/Passive-Data Wiki stránka Beiwe platformy so zhrnutím pasívnych dát.

[26] Laura Lovett. In-depth: Beth israel's digital psychiatry division looks to integrate passive, active phone data into patient care, January 2020. https://www.mobihealthnews.com/news/depth-beth-israels-digital-psychiatry-division-looks-integrate-passive-active-phone-data-00.

[27] Division of Digital Psychiatry BIDMC. mindlamp-activities repository, 2021. https://github.com/BIDMCDigitalPsychiatry/LAMP-activities Zdrojové kódy aktivít platformy mindLAMP.

[28] Division of Digital Psychiatry BIDMC. LAMP Platform User Guide, 2021. https://docs.lamp.digital/ Hlavná stránka dokumentácie LAMP platformy.

[29] Division of Digital Psychiatry BIDMC. Download mindLAMP, 2021. https://docs.lamp.digital/app Inštalácia mindLAMP 2 aplikácie na iOS a Android OS.

[30] Mindpax. Studies, 2021. https://mindpax.me/studies.html Zoznam všetkých štúdií z firmy Mindpax. Štúdia o kvalite dotazníku ASERT je

v dobe písania tejto práce na prvom mieste a nemá svoj vlastný odkaz na publikáciu.

[31] Y. Lecrubier, E. Weiller, T. Hergueta, P. Amorim, L. I. Bonora, J. P. Lépi, D. Sheehan, J. Janavs, R. Baker, K. Harnett-Sheehan, E. Knapp, and M. Sheehan. M.I.N.I. - MINI INTERNATIONAL NEUROPSYCHIATRIC INTERVIEW, 2006. https://huibee.com/wordpress/wp-content/uploads/2013/11/Mini-International-Neuropsychiatric-Interview-MINI.pdf Dotazník M.I.N.I.

[32] Amazon Web Services Inc. or its affiliates. Amazon Web Services, 2021. https://aws.amazon.com/ Hlavná stránka Amazon Web Services.

[33] Division of Digital Psychiatry BIDMC. Costs of Deploying the LAMP Platform, 2021. https://docs.lamp.digital/deploy/costs/ Hlavná stránka deployment návodu LAMP platformy „On-Premises“.Neobsahuje všetky prerekvizity pre úspešný deployment.

[34] Division of Digital Psychiatry BIDMC. Prerequisites for Deploying the LAMP Platform, 2021. https://docs.lamp.digital/deploy/prereqs Stránka deployment prerekvizít LAMP platformy „On-Premises“.

[35] Division of Digital Psychiatry BIDMC. Deploying the LAMP Platform, 2021. https://docs.lamp.digital/deploy/deploying Návod na deployment LAMP platformy „On-Premises“.

[36] mindLAMP. mindLAMP Dashboard, 2021. https://dashboard.lamp.digital/ Dashboard platformy mindLAMP s možnosťou pripojenia na ľubovoľný správne nastavený server hostujúci LAMP server.

[37] Division of Digital Psychiatry BIDMC. Cortex & API, 2021. https://docs.lamp.digital/data_science/intro Úvodná stránka dokumentácie Cortexu a API projektu mindLAMP.

[38] Division of Digital Psychiatry BIDMC. Data Types, 2021. https://docs.lamp.digital/data_science/data_types Dokumentácia datových typov zbieraných platformou LAMP.

[39] Division of Digital Psychiatry BIDMC. Preparing to Analyze Your Data in Python, 2021. https://docs.lamp.digital/data_science/python Úvod do analýzy dát získaných z platformy LAMP s pomocou programovacieho jazyka Python. Obsahuje odkazy na konkrétne riešenia a príklady získavania dát z databázy.

[40] Division of Digital Psychiatry BIDMC. Preparing to Analyze Your Data in R, 2021. https://docs.lamp.digital/data_science/r Úvod do analýzy dát získaných z platformy LAMP s pomocou programovacieho jazyka R. Obsahuje odkazy na konkrétne riešenia a príklady získavania dát z databázy. Oproti úvodu k Pythonu je rozsiahlejší.

[41] Division of Digital Psychiatry BIDMC. LAMP consortium, 2021. https://mindlamp.discourse.group Oficiálne fórum platformy LAMP.

[42] scikit learn. scikit-learn Machine Learning in Python, 2021. https://scikit-learn.org/stable/ Hlavná stránka projektu scikit-learn.

[43] ScienceDirect. Kurtosis, 2021. https://www.sciencedirect.com/topics/neuroscience/kurtosis Definície kurtózy z článkov a kníh.

[44] scikit learn. API Reference, 2021. https://scikit-learn.org/stable/modules/classes.html Dokumentácia API balíčku sklearn. Obsahuje (mimo iné) odkazy pre Decision tree algoritmy, SVM a Naive Bayes.

[45] Stack Exchange Inc. How to highlight python syntax in latex listings \lstinputlistings command, 2012. https://tex.stackexchange.com/questions/83882/how-to-highlight-python-syntax-in-latex-listings-lstinputlistings-command Prevzatá a upravená odpoveď užívateľa redmode. Odpoveď užívateľa CodingYourLife bola zvážená ale nesplňovala autorove požiadavky pre prácu s codeblockmi.

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Sakači** Jméno: **Ondrej** Osobní číslo: **487028**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Softwarové inženýrství a technologie**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Sběr a analýza dat získaných pomocí mobilních platforem pro sběr behaviorálních dat

Název bakalářské práce anglicky:

Data collection and analysis from behavioral mobile platforms

Pokyny pro vypracování:

Cílem této práce je seznámit se s digitálním fenotypizací a dostupnými mobilními platformam pro sběr behaviorálních dat
1. Seznamte se s aktuálním stavem a dostupnými mobilními platformami.
2. Vyberte a implementujte vybranou platformu.
3. Proveďte pilotní testování na vzorku 10 dobrovolníků.
4. Proveďte analýzu a vizualizaci behaviorálních dat

Seznam doporučené literatury:

1. MOOD STATE PREDICTION FROM SPEECH OF VARYING ACOUSTIC QUALITY FOR INDIVIDUALS WITH BIPOLAR DISORDER, Gideon, 2016
2. Recognition of Depression in Bipolar Disorder: Leveraging Cohort and Person-Specific Knowledge, Khorram, 2016
3. ECOLOGICALLY VALID LONG-TERM MOOD MONITORING OF INDIVIDUALS WITH BIPOLAR DISORDER USING SPEECH, Proc IEEE Int Conf Acoust Speech Signal Process. 2014 May; 2014: 4858–4862.
4. Realizing the Potential of Mobile Mental Health: New Methods for New Data in Psychiatry, Torous, P Staples, JP Onnela. Current psychiatry reports 17 (8), 61, 2015

Jméno a pracoviště vedoucí(ho) bakalářské práce:

doc. Ing. Daniel Novák, Ph.D., Analýza a interpretace biomedicínských dat FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **12.02.2021** Termín odevzdání bakalářské práce: **21.05.2021**

Platnost zadání bakalářské práce: **30.09.2022**

doc. Ing. Daniel Novák, Ph.D. podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. Datum převzetí zadání

Podpis studenta