In dit document leest u enkele email van alle communicatie die heeft plaatsgevonden aan mijn kant met CBS. Natuurlijk zijn er nog veel meer emails, maar dit geeft een indicatie over de communicatie waar ik voor gezorgd heb.

Ik heb het even gecontroleerd, BMR097, BMR044 hebben beide negatieve g-force bij zitten op z-as.

Van: Luiten, J.M.M. (Annemieke) <a.luiten@cbs.nl>
Verzonden: dinsdag 10 november 2020 16:03
Aan: 'matthew turkenburg' <matthew.turkenburg1@hotmail.com>
Onderwerp: RE: Meeting Questions

34 was die ene in het midden, die waarschijnlijk nog aan het staan was. Maar wie waren die drie met de negatieve g-force?

Van: matthew turkenburg <matthew.turkenburg1@hotmail.com>
Verzonden: dinsdag 10 november 2020 14:19
Aan: Luiten, J.M.M. (Annemieke) <a.luiten@cbs.nl>
CC: M.J.Boon@student.hhs.nl; J.F.B.Bolte@hhs.nl; A.Akbas@student.hhs.nl; A.A.Safdari@student.hhs.nl; C.R.Werkhoven@student.hhs.nl
Onderwerp: Re: Meeting Questions

Hi Annemieke,

Sorry for the late reply, I have been very busy. The respondent number was ['BMR043']. I also like to mention that one of our group members have quit this minor (D.Sekijevskis@student.hhs.nl), and from this point doesn't need to be included into the CC.

This also means that all the communication from this point, can be done in Dutch.

Kind regards,
Matthew Turkenburg

Misschien is het handiger als jullie mij uitnodigen vanuit jullie Teams omgeving. Vandaaruit is het misschien makkelijker om dingen te laten zien. Als ik jullie uitnodig dan wordt het Zoom.
Ik kan bijvoorbeeld van 10 tot 11.30 uur (reserveer maar wat ruimer voor de zekerheid) en Sigrid ook.

Annemieke

Van: matthew turkenburg <matthew.turkenburg1@hotmail.com>
Verzonden: woensdag 2 december 2020 15:19
Aan: Luiten, J.M.M. (Annemieke) <a.luiten@cbs.nl>; Adnan Akbas <A.Akbas@student.hhs.nl>; Ali Safdari <A.A.Safdari@student.hhs.nl>; Colin Werkhoven <C.R.Werkhoven@student.hhs.nl>; Mark Boon <M.J.Boon@student.hhs.nl>; Mattew Turkenburg <M.C.Turkenburg@student.hhs.nl>
Onderwerp: Re: Afsluiting fase - ActivePal

Hi Annemieke,

Op dit moment zitten we in week 13 van het project. Dit betekent dat de komende werken (week 13 tot en met 18) voornamelijk in het teken staan van het schrijven van de paper. Natuurlijk zijn de modellen en de processen nog niet helemaal af, en daarom blijven wij als groep hier nog steeds aan werken.

Als groep zijn we echt ontzettend blij dat we op zo een kort termijn nog een meeting kunnen hebben, kunt u de meeting aanmaken op een tijd dat u uitkomt? Ik heb op dit moment John nog niet uitgenodigd om de meeting bij te wonen, de meeting staat meer in het taken hoe we uiteindelijk de overdracht van kennis kunnen overbrengen en wat onze laatste stappen zullen zijn in het project.

Met vriendelijke groet,
Matthew Turkenburg

Van: Luiten, J.M.M. (Annemieke) <a.luiten@cbs.nl>
Verzonden: woensdag 2 december 2020 14:04
Aan: 'matthew turkenburg' <matthew.turkenburg1@hotmail.com>; Adnan Akbas <A.Akbas@student.hhs.nl>; Ali Safdari <A.A.Safdari@student.hhs.nl>; Colin Werkhoven <C.R.Werkhoven@student.hhs.nl>; Mark Boon <M.J.Boon@student.hhs.nl>; Mattew Turkenburg <M.C.Turkenburg@student.hhs.nl>
Onderwerp: RE: Afsluiting fase - ActivePal

Ha Matthew,

Nu al afsluiten? Ik dacht dat het project tot eind januari liep?

Maar een meeting kan altijd, en je hebt geluk, ik moet vrijdag werken en ik heb nog ruimte in de ochtend. En Sigrid ook volgens haar agenda.
Doet John ook mee?

Groeten,
Annemieke

Van: matthew turkenburg <matthew.turkenburg1@hotmail.com>
Verzonden: woensdag 2 december 2020 14:00
Aan: Luiten, J.M.M. (Annemieke) <a.luiten@cbs.nl>; Adnan Akbas <A.Akbas@student.hhs.nl>; Ali Safdari <A.A.Safdari@student.hhs.nl>; Colin Werkhoven <C.R.Werkhoven@student.hhs.nl>; Mark Boon <M.J.Boon@student.hhs.nl>; Mattew Turkenburg <M.C.Turkenburg@student.hhs.nl>
Onderwerp: Re: Afsluiting fase - ActivePal

Hi Annemieke,

Ter voorbereiding op het afsluiten van het project zouden we graag nog een meeting willen inplannen. In deze meeting zullen we al onze resultaten bespreken, de discussie, zwaktes en sterktes van het onderzoek.

De meeting zouden we graag zo snel mogelijk willen inplannen. Onze voorkeur gaat uit naar aankomende vrijdag, maar gezien het tijdstip van deze email, denken wij dat dit mogelijk niet meer kan. Indien vrijdag niet al mogelijk is, zouden we graag een meeting willen inplannen voor begin van volgende week (maandag, dinsdag)? Indien ook deze data niet mogelijk is, dan moeten we even bekijken wat de andere opties zijn!

Met vriendelijke groet,
Matthew Turkenburg

Van: Luiten, J.M.M. (Annemieke) <a.luiten@cbs.nl>
Verzonden: maandag 30 november 2020 09:27
Aan: Adnan Akbas <A.Akbas@student.hhs.nl>; Ali Safdari <A.A.Safdari@student.hhs.nl>; Colin Werkhoven <C.R.Werkhoven@student.hhs.nl>; Mark Boon <M.J.Boon@student.hhs.nl>; Mattew Turkenburg <M.C.Turkenburg@student.hhs.nl>
CC: 'Bolte, J.F.B.' <J.F.B.Bolte@hhs.nl>; Hoek, S.T. van (Sigrid) <st.vanhoek@cbs.nl>; 'Vuurens, J.B.P.' <j.b.p.vuurens@hhs.nl>
Onderwerp: FW: Presentatie studenten HHS

Ha jongens,

Afgelopen vrijdag kon ik zelf niet bij jullie presentatie zijn, maar Sigrid was er wel. Zij had een aantal vragen aan jullie. Zouden jullie die willen beantwoorden?

Classificeren van activiteiten:

- Op welke data (features/ respondenten/ frequentie) is de Random Forest gebouwd?
- Hoe hebben ze de train (en validatie) en test set gesplitst?
- Welke activiteiten zijn er geclassificeerd? Later zag ik in de geclassificeerde weekdata een kolom ‘unknown activity’, hoe is deze bepaald?

- Welke modellen naast Random Forest zijn geprobeerd en wat was ongeveer hun performance?

Voorspelling MET waardes

- Bij de MET predictie nemen ze de variabelen weight, length, age, mag value, meets balance guidelines, mean speed mee als features. Welke train-test split is er gebruikt en wat is de frequentie van de data?
- Wat houden de variabelen mag value en meets balance guidelines in?
- Length was bijvoorbeeld gemeten in cm. Omdat we maar max. 40 proefpersonen hebben, is het waarschijnlijk dat er voor elke lengte maar één proefpersoon is in de data. Hoe goed zou dit model dan extrapoleren naar andere proefpersonen? Stel dat je een andere test set neemt, wordt de performance van het model dan veel lager?
- In de grafieken die ze lieten zien, wat betekent de x-axis label 'Prediction number'?

Applicatie:

- Waarom zijn er andere categorieën gekozen dan in de labsessies? Fietsen licht en zwaar zijn samengenomen en springen zag ik ook niet terug.

Met vriendelijke groet,
Annemieke

Hi Annemieke,

Leuk om te horen dat Sigrid aanwezig was bij de presentaties! Voor de leesbaarheid heb ik de vragen opgedeeld in twee categorieën namelijk Activity Recognition en voorspellingen MET waardes.

Activity Recognition

Op welke data (features/ respondenten/ frequentie) is de Random Forest gebouwd?

Training/validatie respondenten

BMR002	BMR008	BMR011	BMR012
BMR014	BMR018	BMR030	BMR030
BMR032	BMR033	BMR036	BMR040
BMR041	BMR042	BMR043	BMR044
BMR052	BMR053	BMR055	BMR058
BMR064	BMR098		

80% van training/validatie dataset wordt gebruikt als training dataset en 20% wordt gebruikt als validatie dataset.

Test respondenten

BMR004	BMR034	BMR097

Features

Kolom	**Feature 1**	**Feature 2**
X	Standard deviatie	Gemiddelde
Y	Standard deviatie	Gemiddelde
Z	Standard deviatie	Gemiddelde

De features wordt berekend de frequentie 9.4S.

Hoe hebben ze de train (en validatie) en test set gesplitst?
De test dataset is bestaat uit 3 respondenten. Deze respondenten zijn gekozen met behulp van een script dat kijkt of test respondenten representatief zijn in vergelijking met de training dataset. De overige respondenten gaan in de training/validation dataset. De dataset wordt daarna met behulp sklearn train_test_split function gesplits in training (80%) en validatie (20%) dataset. In de functie geef ik ook aan dat stratificatie gedaan moet worden op basis van de labels.
De random_state parameters is staat op 0.
Welke activiteiten zijn er geclassificeerd? Later zag ik in de geclassificeerde weekdata een kolom 'unknown activity', hoe is deze bepaald?
De activiteiten die zijn geclassificeerd

Fietsen	Lopen	Rennen
Staan	Zitten	

'Unknown activity' wordt alleen gebruikt in de applicatie. De label 'unknown activity' wordt alleen toegepast wanneer de activity recognition model niet kan bepalen welke activiteit label geplakt moet worden

Welke modellen naast Random Forest zijn geprobeerd en wat was ongeveer hun performance?
Src: all_steps_activity recognition

validatie dataset resultaten			
Model	**Accuracy**	**Precision**	**Recall**
Decision Tree	0.96	0.96	0.96
Random forest	0.97	0.98	0.97

Waarom zijn er andere categorieën gekozen dan in de labsessies? Fietsen licht en zwaar zijn samengenomen en springen zag ik ook niet terug.
Fietsen licht en zwaar zijn gecombineerd tot 1 activiteit omdat we dachten door gebruik te maken van de feature snelheid het te kunnen onderscheiden en veel accurater MET waardes te berekenen. Helaas kwamen we achter dat niet het geval is en daarom zijn we nu bezig om MET predicition modellen te creëren voor fietsen licht en zwaar. De activity recognition model is na de presentatie al hiervoor aangepast.

	Accuracy	Precision	Recall
Validation dataset	0.96	0.96	0.96
Test dataset	0.85	0.85	0.85
K-fold cross validation	0.81(+/- 0.05)	0.81 (+/- 0.05)	0.82 (+/- 0.04)

Src: all_steps_activity recognition_final_version_split_cycling
Springen en traplopen zijn niet meegenomen omdat we hiervoor weinig of
geen VO2 data van Vyntus hebben. Hierdoor konden we geen MET-waardes berekenen dit heeft als gevolg dat we modellen geen ground truth waardes hebben om hierop te trainen.

Voorspelling MET waardes

Bij de MET predictie nemen ze de variabelen weight, length, age, mag value, meets balance guidelines, mean speed mee als features. Welke train-test split is er gebruikt en wat is de frequentie van de data?

In de dataset die we gebruiken voor de MET models maken we gebruik van 25 gebruikers, 22 worden gebruikt voor de train/validation datasets en 3 voor de test set. Op deze 22 gebruikers voeren we `train_test_split` method uit met een `test_size` van `0.2`. Elke gebruiker heeft 5 rows aan data omdat de MET waardes per minuut worden berekend en de meeste activiteiten in het lab 5 minuten zijn uitgevoerd. Dit houdt in `22x5 rows` = 110 train/validation rows en `3x5 = 15` test rows.
Na wat feedback van Brian Keijzer over onze train_test_split methode hebben we een andere manier toegepast. Voor deze methode hebben we de 22 gebruikers gesplitst op hele gebruikers i.p.v. random 20% pakken van alle 110 rows. Bij deze methode werden er 17 hele gebruikers voor de training gebruikt en 5 hele gebruikers voor de validation. In de alinea hierboven werden er random rows gepakt. Het kon dus voorkomen dat er data van bepaalde gebruikers niet in de validation set kwamen.
We hebben de hierboven genoemde methode toegepast met representatieve gebruikers voor de train, validation en test set, maar dit zorgde voor veel mindere resultaten voor onze modellen. We hebben besloten om deze methode niet te gebruiken.

Wat houden de variabelen mag value en meets balance guidelines in?

`sum_mag_acc`
Deze variabel houdt in sum of magnitude of acceleration. Omdat de MET waardes per minuut berekend worden is de acceleration bij elkaar opgeteld. Deze acceleration bestaat uit de X, Y en Z waarde tot de macht 2 in het kwadraat. De formule die we hiervoor hebben gebruikt:

```
def convert_to_acceleration(row):
return math.sqrt(row['pal_accX'] ** 2 + row['pal_accY'] ** 2 + row['pal_accZ'] ** 2)
```

Meets_balance_guidelines
In het `respondenten.csv` bestand staat een column `voldoet aan richtlijn balansoefeningen`. Deze column bestaat uit een `ja/nee` waarde. We hebben deze naar een numerieke `1/0` geconverteerd zodat we deze in ons model konden gebruiken.

Length was bijvoorbeeld gemeten in cm. Omdat we maar max. 40 proefpersonen hebben, is het waarschijnlijk dat er voor elke lengte maar één proefpersoon is in de data. Hoe goed zou dit model dan extrapoleren naar andere proefpersonen? Stel dat je een andere test set neemt, wordt de performance van het model dan veel lager?

De lengtes van de respondenten zijn inderdaad verschillend van elkaar. Omdat onze modellen nog niet helemaal compleet zijn hebben wij de test set nog niet toegepast op ons model. We gaan er van uit dat

de resultaten hiervan wat minder zullen zijn omdat er weinig data is waarmee we kunnen trainen, valideren en testen.
We hebben ook gekeken naar Body Mass Index(BMI). Deze waarde wordt middels de `weight` en `length` berekend, maar we hebben ervoor gekozen om BMI weg te laten omdat deze waarde een sterkte correlatie heeft met de `weight` en `length`. Is dit aangeraden of kunnen we beter `BMI` gebruiken en `weight` en `length` weghalen?
In de grafieken die ze lieten zien, wat betekent de x-axis label 'Prediction number'?
Uit onze train_test_split komt een training en validation set. Als voorbeeld; de validation set bevat 20 ground truth MET waardes die voorspeld moeten worden. Prediction number is in dit geval de MET waarde die voorspeld wordt. De 'prediction number' axis zelf is een naam die we hebben gebruikt voor regels binnen in validatie set. We konden hier geen passende naam voor bedenken.