

Ten eerste wil ik namens de projectgroep u en de betrokken collega's enorm bedanken voor alle tijd en energie die tijdens het ActivePal project in ons zijn gestoken. Als groep hebben we hierdoor enorm veel kunnen leren en dit komt mede door jullie. Ter afsluiten van het project willen we naast de scientific paper resultaten, ook de resultaten delen via deze email. Naast deze email ontvangen jullie ook twee andere e-mails, waarin de gehele documentatie van ons project is opgenomen, in de bijlage van de e-mails vind u ook onze Python code en de applicatie code die nodig is om ons werk te reproduceren.

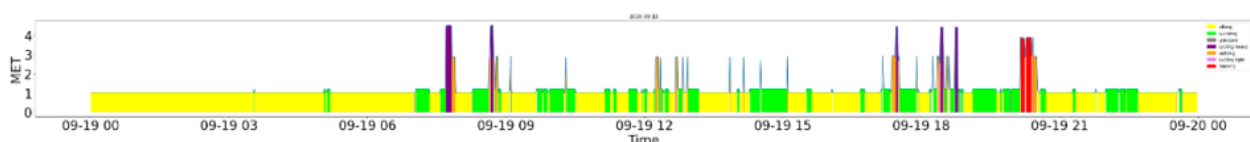
Aan het begin van het project is ons gevraagd om onderzoek naar de volgende hoofdvraag: Can Machine Learning be used to determine whether people did their 150 minutes of moderate physical activity in ActivPal accelerometer data of an entire week? Vanaf het begin van het project zijn wij als groep al enorm bezig geweest om deze vraag te beantwoorden. Doormiddel van het onderzoek kunnen wij de volgende conclusies geven.

Conclusie

Machine learning kan worden ingezet om te voorspellen of een persoon zijn/haar 150 minuten of matige lichamelijke activiteit heeft uitgevoerd. Dit kan worden voorspelt doormiddel van een combinatie van twee verschillende modellen. Ten eerste zal een activiteiten herkenning model moeten worden uitgevoerd op de weekdata. Dit model voorspelt doormiddel van verschillende features welke activiteit door de persoon werd uitgevoerd. Binnen het project hebben wij ervoor gekozen om een random forest model te gebruiken om de activiteiten te voorspellen. Bij deze voorspelling is ervoor gekozen om de activiteiten springen en traplopen niet mee te nemen omdat hiervoor geen lab data aanwezig was (vyntus data) en om deze reden ook geen ground truth voor de MET values.

Nadat de activiteit is voorspelt, komt er een tweede model in het gebruik. Doormiddel van een model dat per activiteit een MET waarde kan voorspelt, is het mogelijk om de intensiviteit van een activiteit te bereken. Voor het voorspellen van de MET waarde, is er gekozen voor twee verschillende modellen waar de resultaten vrij dicht bij elkaar liggen. Namelijk een random forest model en een XGBoost model. Voor het berekenen van de MET waarde, is ervoor gekozen om deze te bereken per minuut zoals aangegeven in de literatuur.

Een combinatie van de activiteiten voorspelling en de MET waarde voorspelling maakt het mogelijk om te voorspellen of een persoon zijn/haar 150 minuten of matige lichamelijke activiteiten heeft uitgevoerd binnen de weekdata. Om dit zo goed mogelijk te illustreren hebben wij een applicatie gebouwd waarin de modellen daadwerkelijk worden toegepast op de weekdata. Een illustratie van de applicatie ziet u hieronder. In dit voorbeeld is voor één van de gebruikers binnen de dataset een overzicht gemaakt van één dag d.m.v onze modellen.



Met de modellen en de applicatie tonen wij aan op welke wijze machine learning kan worden ingezet om te voorspellen of personen hun 150 minuten van matige lichamelijke activiteiten heeft uitgevoerd.

Advies

Tijdens ons onderzoek zijn wij tegen een aantal problemen aangelopen die in het vervolg onderzoek het liefst voorkomen kunnen worden. Ten eerste raden wij aan om in het vervolg een grotere groep correspondenten te verzamelen. Door een grotere dataset te creëren kunnen de modellen beter worden getrained en zullen ze ook beter kunnen voorspellen. Ook adviseren wij dat hier rekening wordt gehouden met leeftijden binnen de groep. Probeer deze evenredig te verspreiden over de gehele doelgroep. Op deze manier kunnen de modellen worden getrained op de volledige doelgroep. Hierbij moet ook rekening worden gehouden met het profiel van de doelgroep. Het is belangrijk dat van de doelgroep ook verschillende karakteristieken aanwezig zijn. Denk hierbij aan mensen die sportief zijn, een hoge BMI hebben of juist heel weinig sporten.

Ook moet er in de toekomst beter worden gecontroleerd bij het opstellen van het dataset. Tijdens ons onderzoek is naar voren gekomen dat het voor kan komen dat er tijdens de afnamen van de lab data niet altijd dezelfde werkwijze is toegepast. Dit zorgt ervoor dat het model getrained moeilijker getrained kan worden op het herkennen van activiteiten.

Als laatste raden wij aan dat er een vervolg onderzoek plaatsvindt waarbij er meer data beschikbaar is, de data op een vaste werkwijze wordt verzameld en dat de dataset representatief is voor de gehele doelgroep. Wanneer dit advies wordt opgevolgd, kunnen de bestaande modellen nogmaals worden uitgevoerd op deze dataset, om te onderzoeken of dit betere resultaten biedt. Hierbij willen wij graag de wel mee geven dat er opnieuw moet worden gekeken naar de sample rate van de MET waarde. Op dit moment hadden wij ervoor gekozen om deze zoals aangegeven in de literatuur naar één minuut te resamplen. Alleen zorgt dit ervoor dat er enorm moeilijk wordt om op grote schaal data te verzamelen. In de toekomst raden wij daarom aan om dit niet op te volgen, maar in plaats daarvan de MET waarde op te bouwen gebaseerd op seconden en de MET waarde op dezelfde schaal te plaatsen. Op deze manier ontstaan er meer data punten waarop de data getrained kan worden.