Description of Thesis

ECTS Points: 32.5
Start date of project: August 2 2021
End date of project: January 14 2022
English Title: "Fairness-oriented interpretability of predictive algorithms"
Danish Title: “Fairness-orienteret fortolkning af prædiktive algoritmer”

English description:
This project will focus on fairness-oriented interpretability of predictive algorithms. The aim is to investigate existing methods to evaluate fairness of predictive algorithms and assess the advantages and disadvantages of the methods. We want to work towards developing a method or tool to evaluate algorithms by taking several fairness and bias criteria into account in an appropriate way. The tool should deliver insight to algorithm owners about the fairness of their prediction algorithm, such that they can take action to address potential issues. We will develop and assess the tool by using it on models trained on a variety of datasets such as ADNI and COMPAS.

Problem statement:
The aim of the project is to combine fairness metrics in order to perform nuanced analyses of fairness of predictive algorithms and develop a toolkit which enables data scientists to include fairness assessment in their modelling workflow. To reach this goal we will

- Research and understand current fairness criteria along with their advantages and disadvantages
- Develop a toolkit in Python that combines several fairness measures in order to gain a nuanced and comprehensive analysis of the fairness of a predictive algorithm
- Showcase the possibilities and limitations of the toolkit on constructed examples consisting of predictive algorithms modelling simple real-world datasets and generated synthetic data
- Set up a predictive algorithm on a more comprehensive medical dataset aiming for high performance without taking fairness of the model into account
- Use the toolkit to assess the above predictive model with respect to fairness and take steps to mitigate potential issues of unfairness