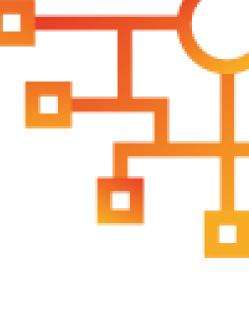


Interactive biobank feature selection using Pearson's correlation



Tim Swarts^{1,3}

Dr Sipko van Dam^{2,3}

- ¹ Student Bioinformatics, Institute for Life Science & Technology, Hanze University of Applied Sciences, Groningen, The Netherlands
- ² Head of Bioinformatics, Ancora Health B.V., Groningen, The Netherlands
- ³ Department of Endocrinology, University of Groningen, University Medical Center Groningen, P.O. Box 30001, 9700 RB, Groningen, The Netherlands

February 1st, 2024



Background

Despite the continuous rise in global health expectancy, prevalence of preventable diseases is increasing [1-3]. To allow effective prevention, Ancora uses machine learning algorithms to identify individuals at high risk [4]. They data from the UK Biobank. [5] This database contains biomarkers, measurements, questionnaire answers, and genetic data from 500,000 UK participants [6]. Due to strong collinearity between features effective feature selection is necessary. This can increase efficiency and performance of machine learning models trained on such large datasets [7, 8].



Project Goal

Creating graphical user interface that:

- Allows interactive feature selection for machine learning
- Helps prevent use of multicollinear features
- Handles both numeric and categorical data
- Is easy to use for doctors and researches without a background in informatics

Selecting a Feature

Decided by highest absolute value of Pearson's correlation with the target

Correcting for Multicollinearity

Effect of selected feature is removed with linear regression

regression <- lm(remaining ~ selected)</pre> remaining <- regression\$residuals</pre>

Calculate Corrected Correlations

For the remaining features, their correlation to the target is revaluated, now with the effect of the selected feature removed



Results

The recursive feature addition method shown in the visualisation was used to reduce a subset from the UK Biobank down from 313 features to just 15, optimised for diabetes classification. This resulted in the following improvements the training of a logistic regression model:

- Training speed increased by a factor of 81.
- AUC improved from 0.89 to 0.91.
- Comparison using Akaike's Information Criterion favoured the simpler model with a difference of 133.



Discussion

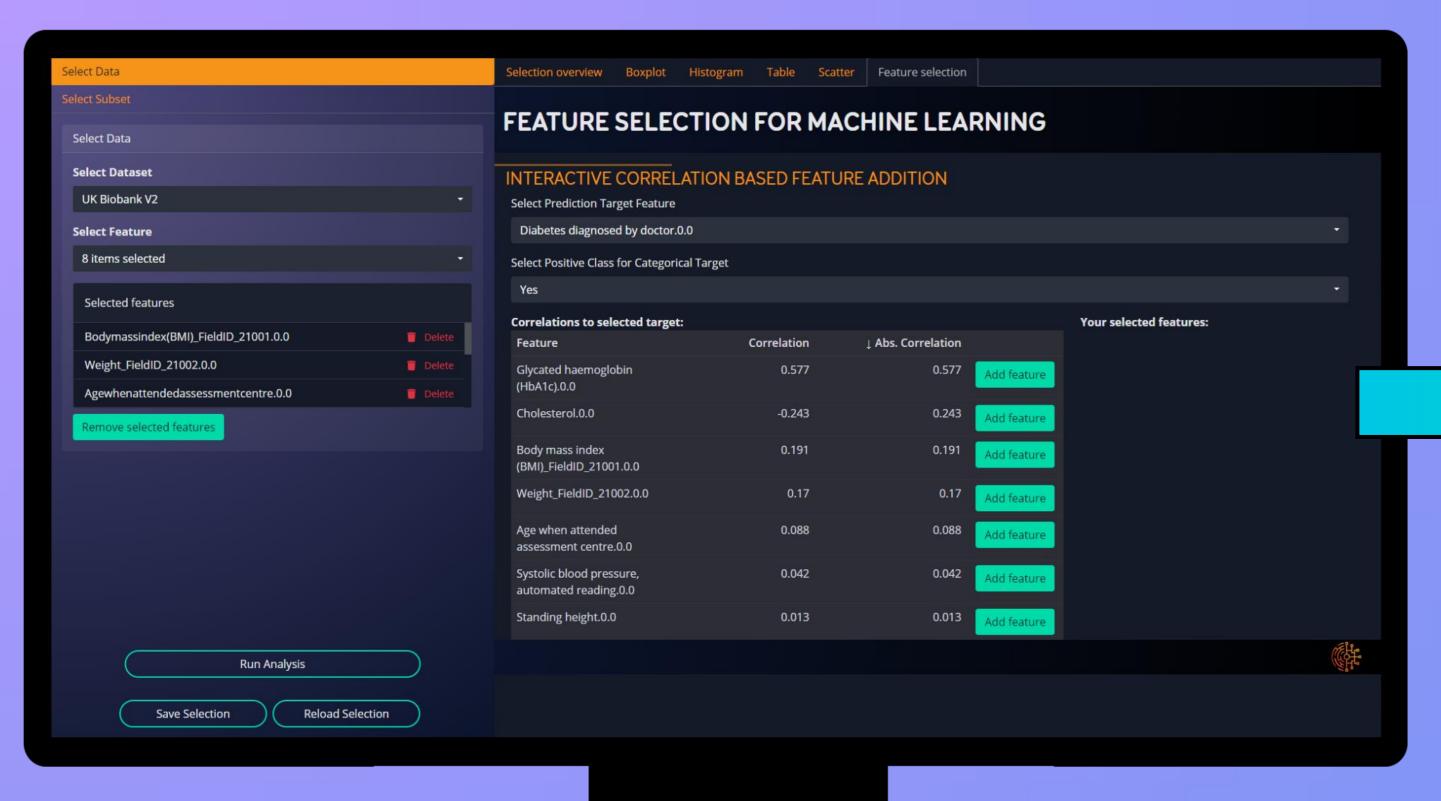
The interface is user friendly and will allow noninformaticians to do complex analyses with the click of a button (figure below).

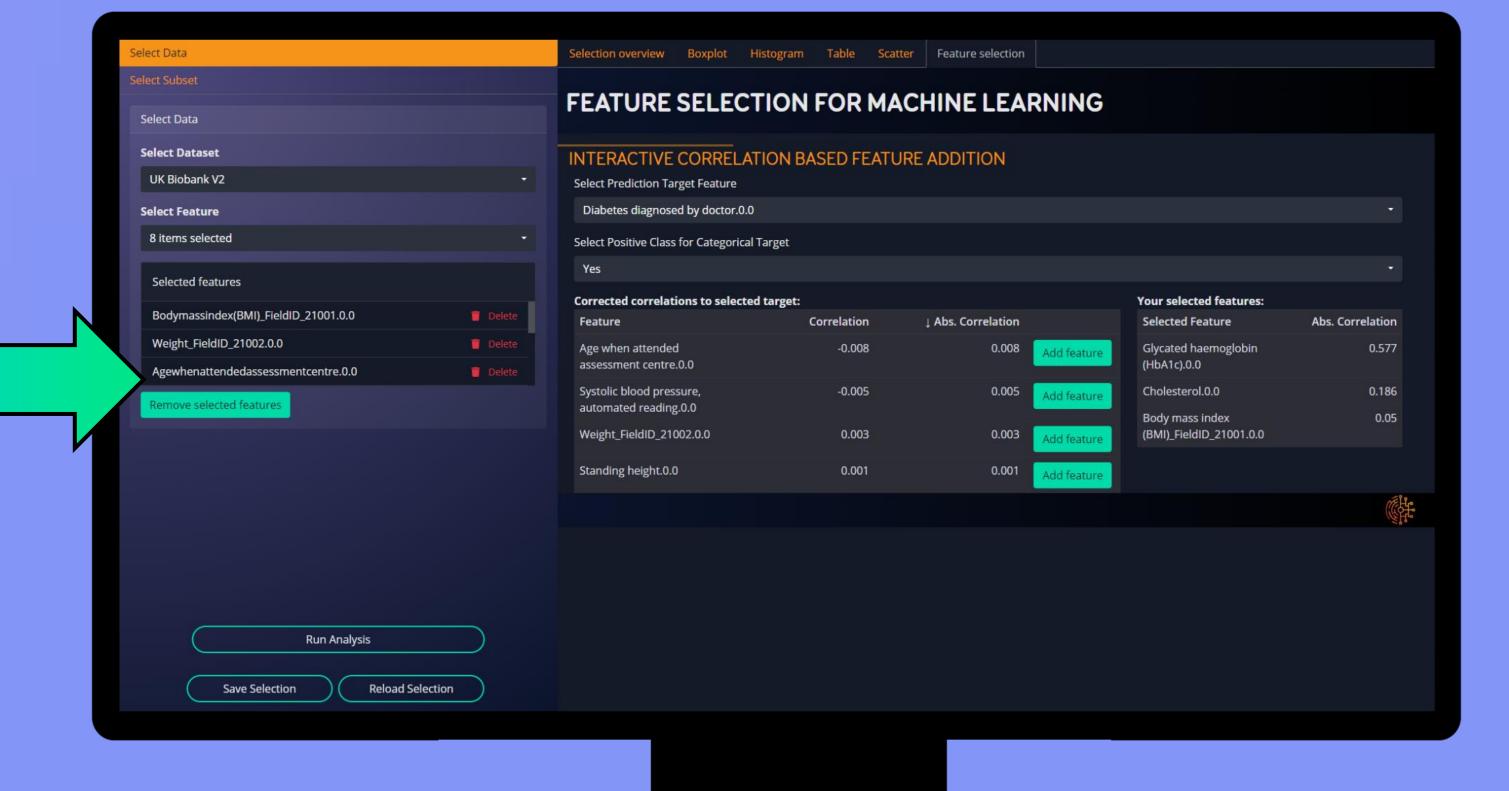
Possible future improvements:

- Integrating the feature selection tool with a machine learning interface for model training and evaluation.
- Optimisation of the correlation table computation to enhance tool responsiveness.
- Further validating the feature selection method by training multiple models and assessing the average performance difference.



Example





References:

- [1] AbdulRaheem, Y. (2023). Unveiling the significance and challenges of integrating prevention levels in healthcare practice. J. Prim. Care
- Community Health, 14, 21501319231186500. Retrieved from https://journals.sagepub.com/doi/10.1177/21501319231186500 • [2] Dattani, S., Rodés-Guirao, L., Ritchie, H., Ortiz-Ospina, E., & Roser, M. (2023). Life Expectancy. Our World in Data. Retrieved from
- [3] Dattani, S., Rodés-Guirao, L., Ritchie, H., Ortiz-Ospina, E., & Roser, M. (2023). Life Expectancy. Our World in Data. Retrieved from
- https://ourworldindata.org/life-expectancy • [4] Ancora Health. (2023, September). Ancora Health. Retrieved from https://ancora.health/

https://ourworldindata.org/life-expectancy

- [5] van Dam, S., Folkertsma, P., Castela Forte, J., de Vries, D. H., Herrera Cunillera, C., Gannamani, R., & Wolffenbuttel, B. H. R. (2023). The necessity of incorporating non-genetic risk factors into polygenic risk score models. Sci. Rep., 13(1), 1351. Retrieved from https://www.nature.com/articles/s41598-023-27637-w
- [6] UK Biobank UK Biobank 2023. (2023, May). UK Biobank UK Biobank. Retrieved from https://www.ukbiobank.ac.uk/ [7] Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (01 2010). Advancing feature selection research. ASU Feature Selection
- Repository Arizona State University, 1–28. • [8] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79.
- **Student:**

Tim Swarts - 372975 t.swarts@st.hanze.nl **Supervisor:** Dr Sipko van Dam sipko@ancora.health

Supervising Teacher: Lübeckweg 2 Ronald Wedema 9723 HE Groningen r.wedema@pl.hanze.nl





Zernikeplein 7 9747 AS Groningen