Assignment for *Bioinformatics aspects of aging and rejuvenation* lecture

Zoltán Szarvas[1]

lecturer: Csaba Kerepesi[2]

[1] Faculty of Computer Science, Eötvös Loránd University
[2] HUN-REN Institute for Computer Science and Control

May 2024

Assignment 1

Train epigenetic clock for microarray-based methylation dataset[3]

```
CCCGGGTCCGGGCGGGGAAGAGCCGCCTCAACGGCAGGGCCCATCCGCGA
GAGGCCAGCGCCCCCGGCCGGTCCAGCCCAGGCCCGCCGCCTCCGCCCTG
GGCTGCTCCCTCCGGGCCCTGCACCGCCCTCCTGCTACTTGGACCGCTTC
CTCACGCCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCAGCGTGGGG
ATCTCGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCCTCCTCCAGCCGCCAAGGCCCCGGCCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACTCGCACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGCGTGG
AATGCGAGCGCGTCTTAAAATCGATGGCGCCTAGGAGTCCATGAAATACG
GTACAGGCTTCCGGCGACGGATGCCCCGCCCCTCACCCACGCTCCGCCCT
CCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCGTCCCCGCGCAGGCG
CGCTCGGGCTGCCGCTGGCTCTTCGCACGCGGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCCA
GGCGTGGTATTCAGGTGCACGCACAGGCCGCCCTCGTGGCGCCCCGACCT
GCGGGCCTACGGATGGGAGCGCGTGGCCCGCGACCTCCGGGCGGGCGGGG
CGGGAACCCTCGTCTTTCGCCCCCGGGGCCCTGCCCTCCTTCGGCCCCGG
CGTCACCAGGCCTGTCCTTGGGTCCAGGGACATCTCGCCCGTCCTGAAGG
ACCCCGCCTCCTTCCGCGCCGCCATCGGCCTCCTGGCGCGACACCTGAAG
GCGACCCACGGGGGCCGCATCGACTACATCGCAGGCGAGTGCCCAGTGGC
CGCATCTAGGGCGCTTCCGCCTCTGCGCGCGCCGAGGGCAGCACGTGGGC
TCTGCGCGTCTGCTTGGGGGAGGGCCTTTGGGGTGCTTCAGGGGGCGCCG
GGACGGGCGCCGTGCTTGGGTCGCCCGGGAAGGGTTGTGAGATTGAGCCC
CCGAGGCCGCCGCGCTGTGCAGGCGTCCTTCCCGCAGGTTCCGGGTCCCC
AGCCCAGGACAGGCGTGACCGAGTTGCCGGGTCAGTTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCACAGGGCCCAGCTGCCTTGCTTCTTGTTCCTTCT
GCGAGCTGGTATTGAGCGCCTGCCACGAGCCAGGCCTTCCCTGGTGAAGA
TCACGGAATGCCCACCCAGGGAAGGGAGGCCTGGAGGCCTCCGGGAGAGC
CCAAGAGGTGGCCCAGGGAGAACAGAGTGTTCCTGGCCGTCTTGCCTCTC
CTAGGGTGTGACAGCCCACTCCCTGGACACTGCCCTGAGGAAAGCGCCAG
CTCTTGCTGGAGCCACAACACTGCCAGAGCTCCCTTCTCACCTCCTGCAG
GAAGCCCTCCCTGACCTCCTGCCAGGCCGGGGCAGGGTTTCCCTGAGCGT
CCCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCCTTTGAGAGTAAGCTGAGGCCTGTCAGGT
TTCTACCAGCCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCCTAGGAACACACAGCTAAGAAGTGGTCCCTTAAAAGACAGAC
CCAGGTCTGCACTCTGACCTGGAAGCAGCTCCGGGTAGGTGATGGGTAAC
ATTCCTTAAATGGTGCATGTCACTGGCCTTTCAGCTGGGAGCCAACCAGG
TACCCCTTGCCACCGGCCAACCCTGGCCCCTGGGGATTCCCATGCTGCCG
AGTCACTCCTGTCACTTACCCTGACAGGCCTAGACTCCCGAGGCTTCCTC
TTTGGCCCCTCCCTGGCCCAGGAGCTTGGACTGGGCTGCGTGCTCATCCG
AAAGCGGGGGAAGCTGCCAGGCCCCACTCTGTGGGCCTCCTATTCCCTGG
AGTACGGGAAGGTAAGAGGGCTGGGGTGGCCAGAGGAAGGGCAGGGCCAG
GCCACCGTGGCCACTCTCCCCCAGTTCTAAAAGGCCTTCCCAGGCGTGTC
AAGTGGAGCTGCTGTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAACAGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTTCTGCG
GGAGGCTGAGGACCACACACCACTTCCCACTCCAGGCTGAGCTGGAGATT
CAGAAAGACGCCCTGGAGCCAGGACAGAGGGTGGTCGTCGTGGATGATCT
GCTGGCCACTGGTGGTAAGGGTCTCCCCGCAGCCAACTGCTGTGGCTCCA
AGGGCCTGGTGGGAGTGGGACAGGACCTCGCTGTGTGACATGGGATGCAG
CTTACTGTTGTCCAGAGGGTGCCTGGTGGCCAGGCCGACACCTTCCTCTC
CCCATGCCTTCCCCTCCCCAACCCAGGGGCTGGCCTGGAGCACCTGCTCT
CTGCAGCCCAGGCCAACTGGGGACCTCACCCTCCCATCCCCAGGAACCAT
GAACGCTGCCTGTGAGCTGCTGGGCCGCCTGCAGGCTGAGGTCCTGGAGT
GCGTGAGCCTGGTGGAGCTGACCTCGCTTAAGGGCAGGGAGAAGCTGGCA
CCTGTACCCTTCTTCTCTCTCCTGCAGTATGAGTGACCACAGGGCCTCCC
AGCCCAACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGGCAACT
GCAGTGACCAGGGGCACCGGCTGCCCACAGGGAACACATTCCTTTGCTGG
GGTTCAGCGCCTCTCCTGGGGCTGGAAGTGCCAAAGCCTGGGGCAAAGCT
GTGTTTCAGCCACACTGAACCCAATTACACACAGCGGGAGAACGCAGTAA
ACAGCTTTCCCAC
```

Dataset

Source

- Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates (Gene Expression Omnibus) [5] [6]
- approximately 450k CpGs from **human whole blood**
- Illumina Infinium 450k (microarray)

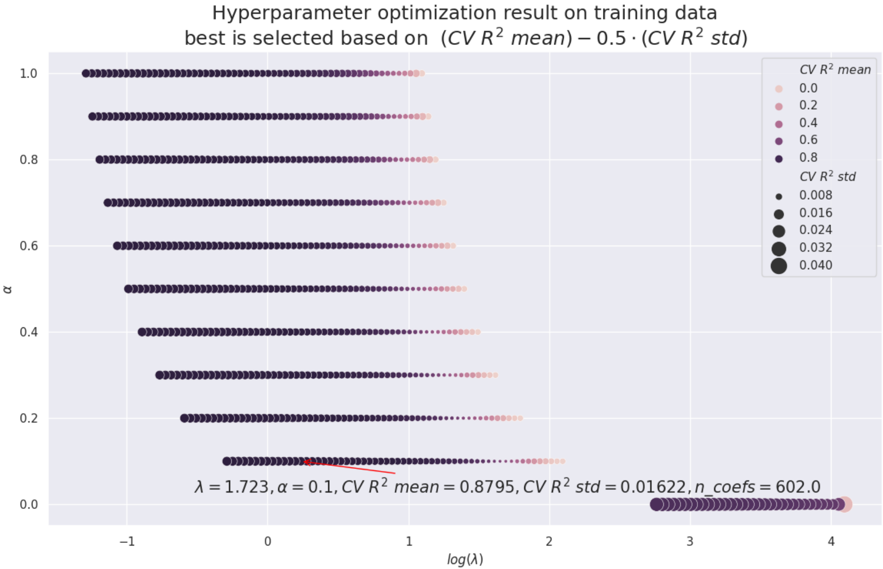
Details

- sample N=656
- features=473 034

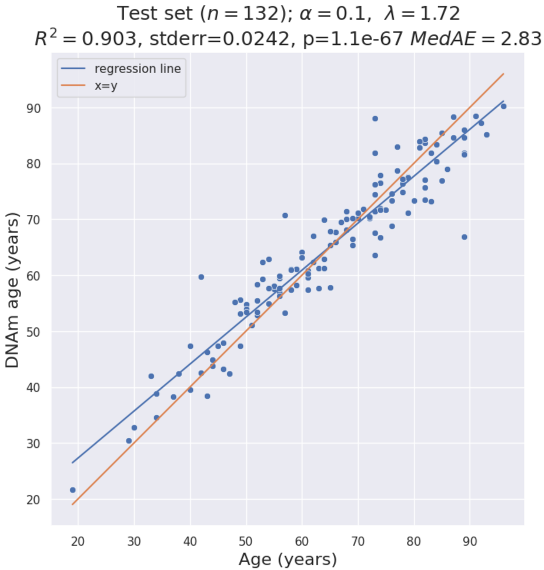
Model

- Elasticnet using Glmnet [1]
- 80-20 train-test split
- 10-fold CV

Hyperparameter Optimization (ElasticNet)

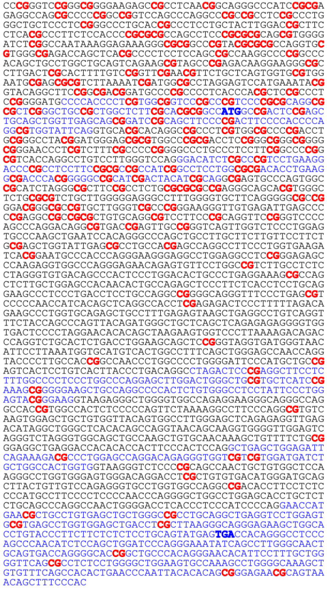


Results on Test Set



Assignment 2

Train epigenetic clock for RRBS-based methylation dataset (with missing data)[4]



```
CCCGGGTCCGGGCGGGGAAGAGCCGCCTCAACGGCAGGGCCCATCCGCGA
GAGGCCAGCGCCCCCGGCCGGTCCAGCCCAGGCCCGCCGCCTCCGCCCTG
GGCTGCTCCCTCCGGGCCCTGCACCGCCCTCCTGCTACTTGGACCGCTTC
CTCACGCCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCAGCGTGGGG
ATCTCGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCCTCCTCCAGCCGCCAAGGCCCCGGCCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACTCGCACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGCGTGG
AATGCGAGCGCGTCTTAAAATCGATGGCGCCTAGGAGTCCATGAAATACG
GTACAGGCTTCCGGCGACGGATGCCCCGCCCCTCACCCACGCTCCGCCCT
CCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCGTCCCCGCGCAGGCG
CGCTCGGGCTGCCGCTGGCTCTTCGCACGCGGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCCA
GGCGTGGTATTCAGGTGCACGCACAGGCCGCCCTCGTGGCGCCCCGACCT
GCGGGCCTACGGATGGGAGCGCGTGGCCCGCGACCTCCGGGCGGGCGGGG
CGGGAACCCTCGTCTTTCGCCCCCGGGGCCCTGCCCTCCTTCGGCCCCGG
CGTCACCAGGCCTGTCCTTGGGTCCAGGGACATCTCGCCCGTCCTGAAGG
ACCCCGCCTCCTTCCGCGCCGCCATCGGCCTCCTGGCGCGACACCTGAAG
GCGACCCACGGGGGCCGCATCGACTACATCGCAGGCGAGTGCCCAGTGGC
CGCATCTAGGGCGCTTCCGCCTCTGCGCGCGCCGAGGGCAGCACGTGGGC
TCTGCGCGTCTGCTTGGGGGAGGGCCTTTGGGGTGCTTCAGGGGGCGCCG
GGACGGGCGCCGTGCTTGGGTCGCCCGGGAAGGGTTGTGAGATTGAGCCC
CCGAGGCCGCCGCGCTGTGCAGGCGTCCTTCCCGCAGGTTCCGGGTCCCC
AGCCCAGGACAGGCGTGACCGAGTTGCCGGGTCAGTTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCACAGGGCCCAGCTGCCTTGCTTCTTGTTCCTTCT
GCGAGCTGGTATTGAGCGCCTGCCACGAGCCAGGCCTTCCCTGGTGAAGA
TCACGGAATGCCCACCCAGGGAAGGGAGGCCTGGAGGCCTCCGGGAGAGC
CCAAGAGGTGGCCCAGGGAGAACAGAGTGTTCCTGGCCGTCTTGCCTCTC
CTAGGGTGTGACAGCCCACTCCCTGGACACTGCCCTGAGGAAAGCGCCAG
CTCTTGCTGGAGCCACAACACTGCCAGAGCTCCCTTCTCACCTCCTGCAG
GAAGCCCTCCCTGACCTCCTGCCAGGCCGGGGCAGGGTTTCCCTGAGCGT
CCCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCCTTTGAGAGTAAGCTGAGGCCTGTCAGGT
TTCTACCAGCCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCCTAGGAACACACAGCTAAGAAGTGGTCCCTTAAAAGACAGAC
CCAGGTCTGCACTCTGACCTGGAAGCAGCTCCGGGTAGGTGATGGGTAAC
ATTCCTTAAATGGTGCATGTCACTGGCCTTTCAGCTGGGAGCCAACCAGG
TACCCCTTGCCACCGGCCAACCCTGGCCCCTGGGGATTCCCATGCTGCCG
AGTCACTCCTGTCACTTACCCTGACAGGCCTAGACTCCCGAGGCTTCCTC
TTTGGCCCCTCCCTGGCCCAGGAGCTTGGACTGGGCTGCGTGCTCATCCG
AAAGCGGGGGAAGCTGCCAGGCCCCACTCTGTGGGCCTCCTATTCCCTGG
AGTACGGGAAGGTAAGAGGGCTGGGGTGGCCAGAGGAAGGGCAGGGCCAG
GCCACCGTGGCCACTCTCCCCCAGTTCTAAAAGGCCTTCCCAGGCGTGTC
AAGTGGAGCTGCTGTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAACAGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTTCTGCG
GGAGGCTGAGGACCACACACCACTTCCCACTCCAGGCTGAGCTGGAGATT
CAGAAAGACGCCCTGGAGCCAGGACAGAGGGTGGTCGTCGTGGATGATCT
GCTGGCCACTGGTGGTAAGGGTCTCCCCGCAGCCAACTGCTGTGGCTCCA
AGGGCCTGGTGGGAGTGGGACAGGACCTCGCTGTGTGACATGGGATGCAG
CTTACTGTTGTCCAGAGGGTGCCTGGTGGCCAGGCCGACACCTTCCTCTC
CCCATGCCTTCCCCTCCCCAACCCAGGGGCTGGCCTGGAGCACCTGCTCT
CTGCAGCCCAGGCCAACTGGGGACCTCACCCTCCCATCCCCAGGAACCAT
GAACGCTGCCTGTGAGCTGCTGGGCCGCCTGCAGGCTGAGGTCCTGGAGT
GCGTGAGCCTGGTGGAGCTGACCTCGCTTAAGGGCAGGGAGAAGCTGGCA
CCTGTACCCTTCTTCTCTCTCCTGCAGTATGAGTGACCACAGGGCCTCCC
AGCCCAACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGGCAACT
GCAGTGACCAGGGGCACCGGCTGCCCACAGGGAACACATTCCTTTGCTGG
GGTTCAGCGCCTCTCCTGGGGCTGGAAGTGCCAAAGCCTGGGGCAAAGCT
GTGTTTCAGCCACACTGAACCCAATTACACACAGCGGAGAACGCAGTAA
ACAGCTTTCCCAC
```

Dataset

Source

- Data from ”Depression and suicide risk prediction models using blood-derived multi-omics data” [2]
- Reduced Representation Bisulfite Sequencing (RRBS)

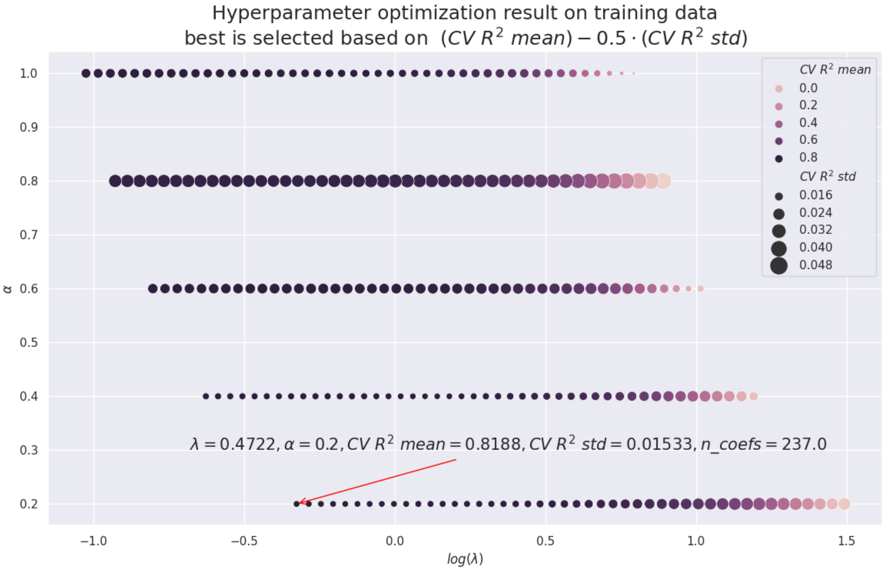
Details

- sample N=182 (Healthy/Control=87, Major Depressive Disorder (MDD)=39, Suicide Attempters (SA)=56)
- features=8 722 096

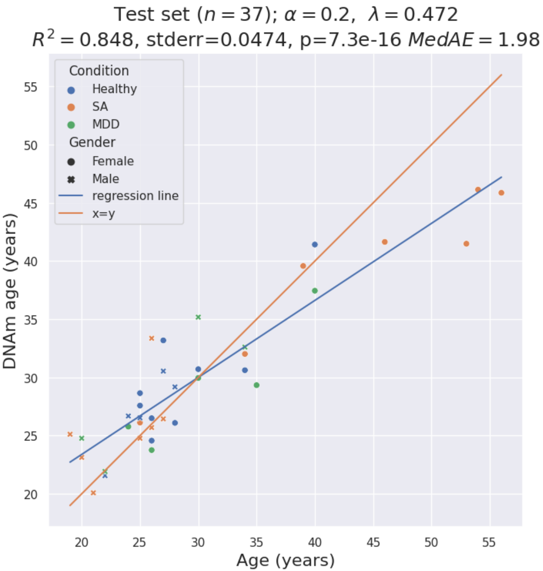
Model

- Elasticnet using Glmnet [1]
- 80-20 train-test split, 10-fold CV
- data for missing CpG sites was filled with **mean**

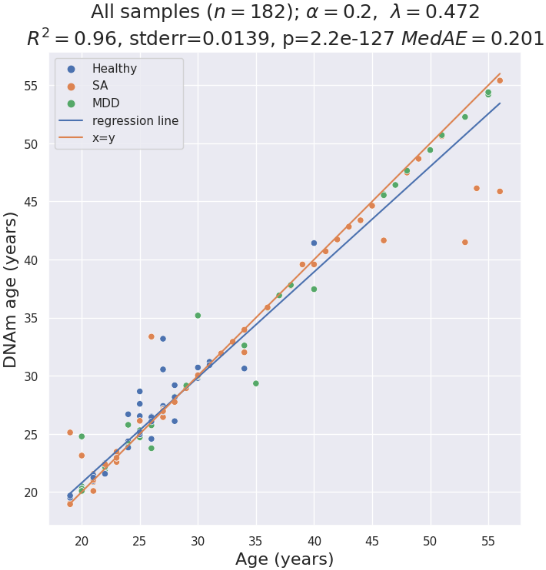
Hyperparameter Optimization (ElasticNet)



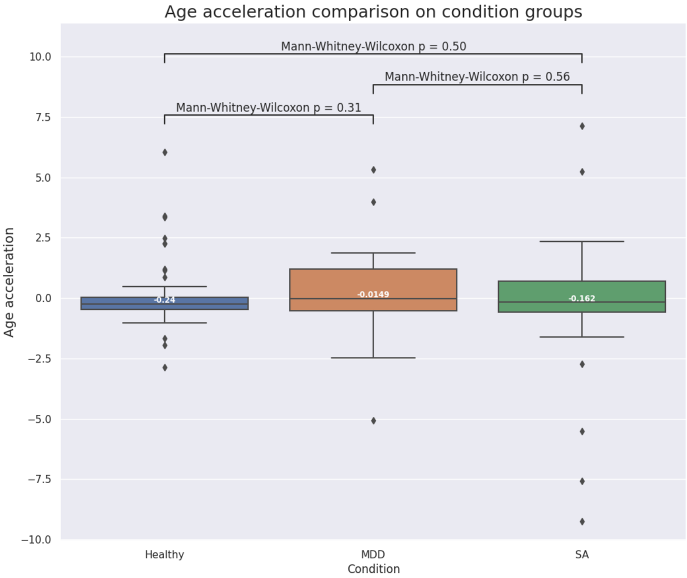
Results on Test Set



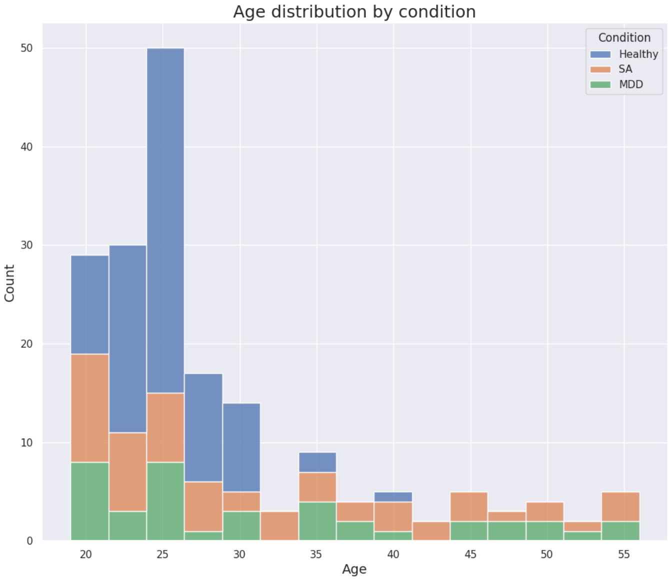
Results for All Instances



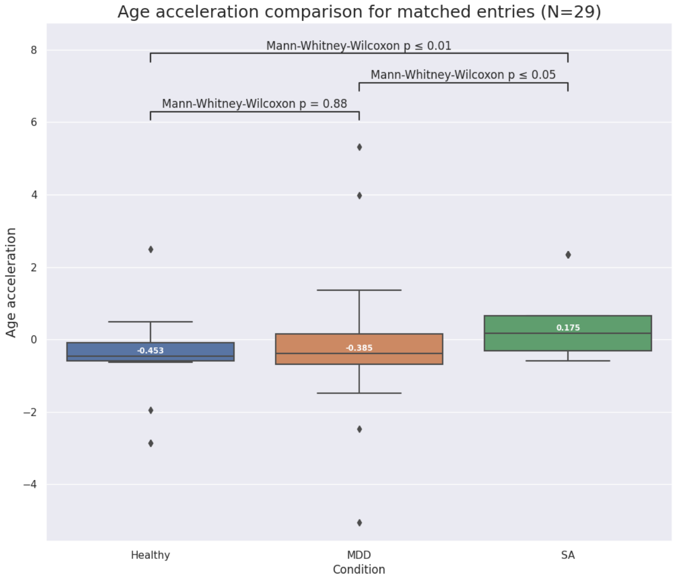
Age acceleration comparison by condition



Age distribution by condition



Age acceleration comparison by condition (controlled)



References I

[1] *An Introduction to ‘glmnet‘ — glmnet.stanford.edu*. https://glmnet.stanford.edu/articles/glmnet.html. [Accessed 16-05-2024].

[2] Youngjune Bhak et al. “Depression and suicide risk prediction models using blood-derived multi-omics data”. en. In: *Transl. Psychiatry* 9.1 (Oct. 2019), p. 262.

[3] *Epigenetic clock assignment 1*. https://github.com/szazo/epigenetic-clock/blob/main/assignment1_microarray.ipynb. 2024.

[4] *Epigenetic clock assignment 2*. https://github.com/szazo/epigenetic-clock/blob/main/assignment2_rrbs.ipynb. 2024.

References II

[5] *Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates*. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279. [Accessed 13-05-2024]. 2012.

[6] Gregory Hannum et al. “Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates”. In: *Molecular Cell* 49.2 (Jan. 2013), pp. 359–367. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.10.016. URL: http://dx.doi.org/10.1016/j.molcel.2012.10.016.