National Research University Higher School of Economics

Data analysis and mining

13 December 2025

Safe aging challenge. Predicting healthy longevity from lifestyle and health data

by Sovetkina Ludmila

Challenge: https://www.kaggle.com/datasets/zoya77/aging-risk-assessment-dataset

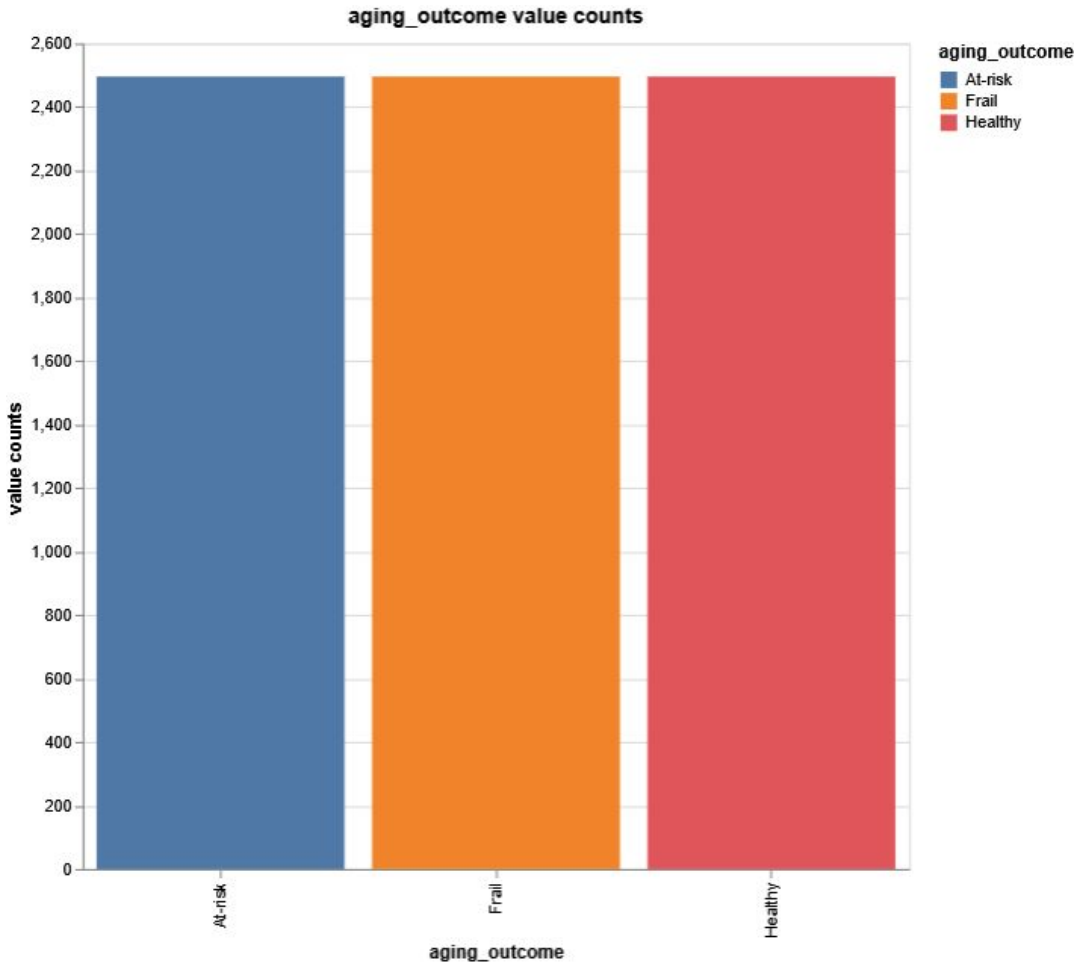
Research objective

The goal of the study is to analyse the **most probable causes** and **metrics** of **aging outcomes**. Additionally we have provided a classification model for reliable prediction of outcomes.

Objectives:

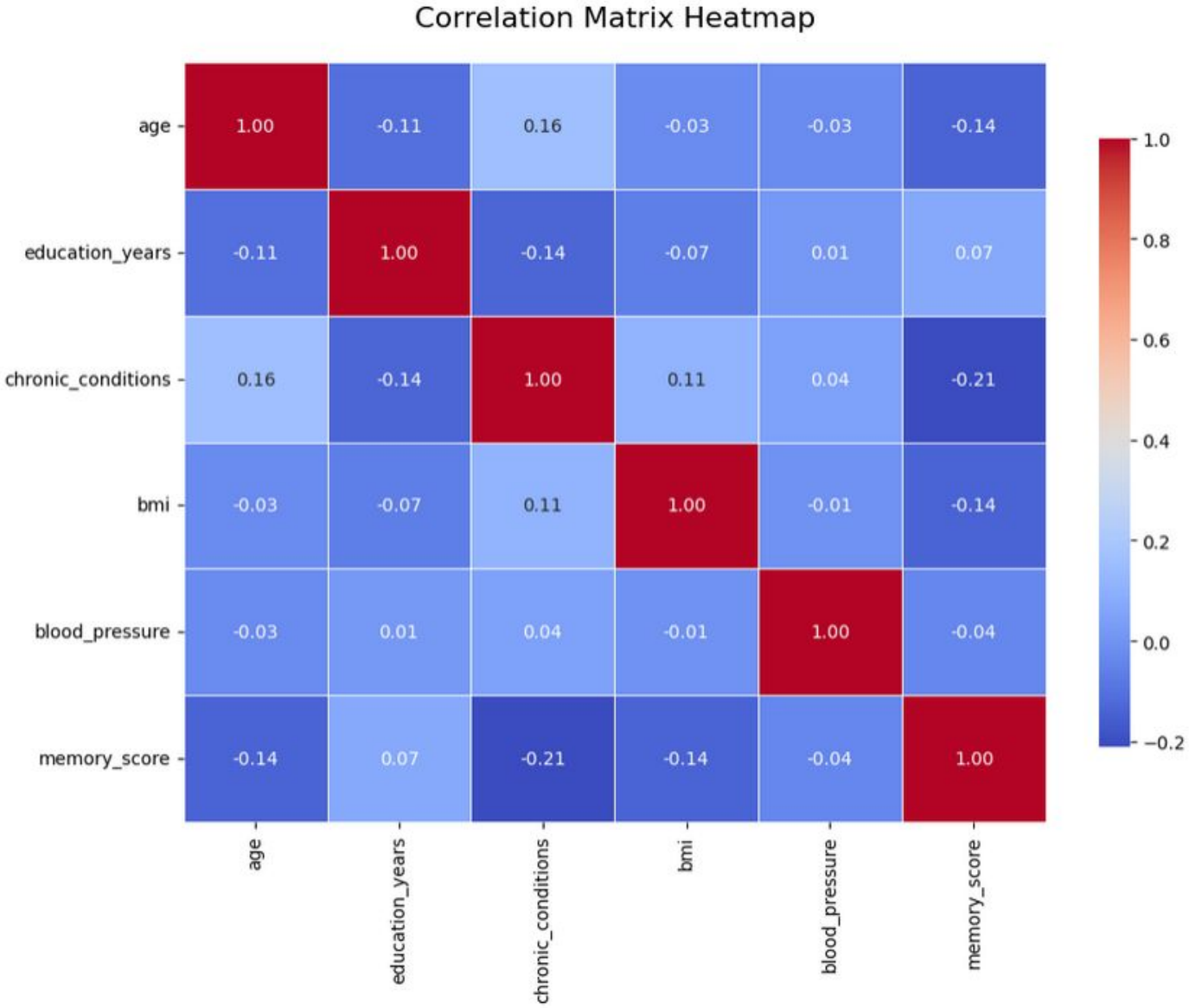
- analyze **most reliable metrics** to predict undesirable outcome of aging;
- analyse **most definite causes** of undesirable outcomes of aging;
- make a **classification model** to predict undesirable outcomes.

About the dataset

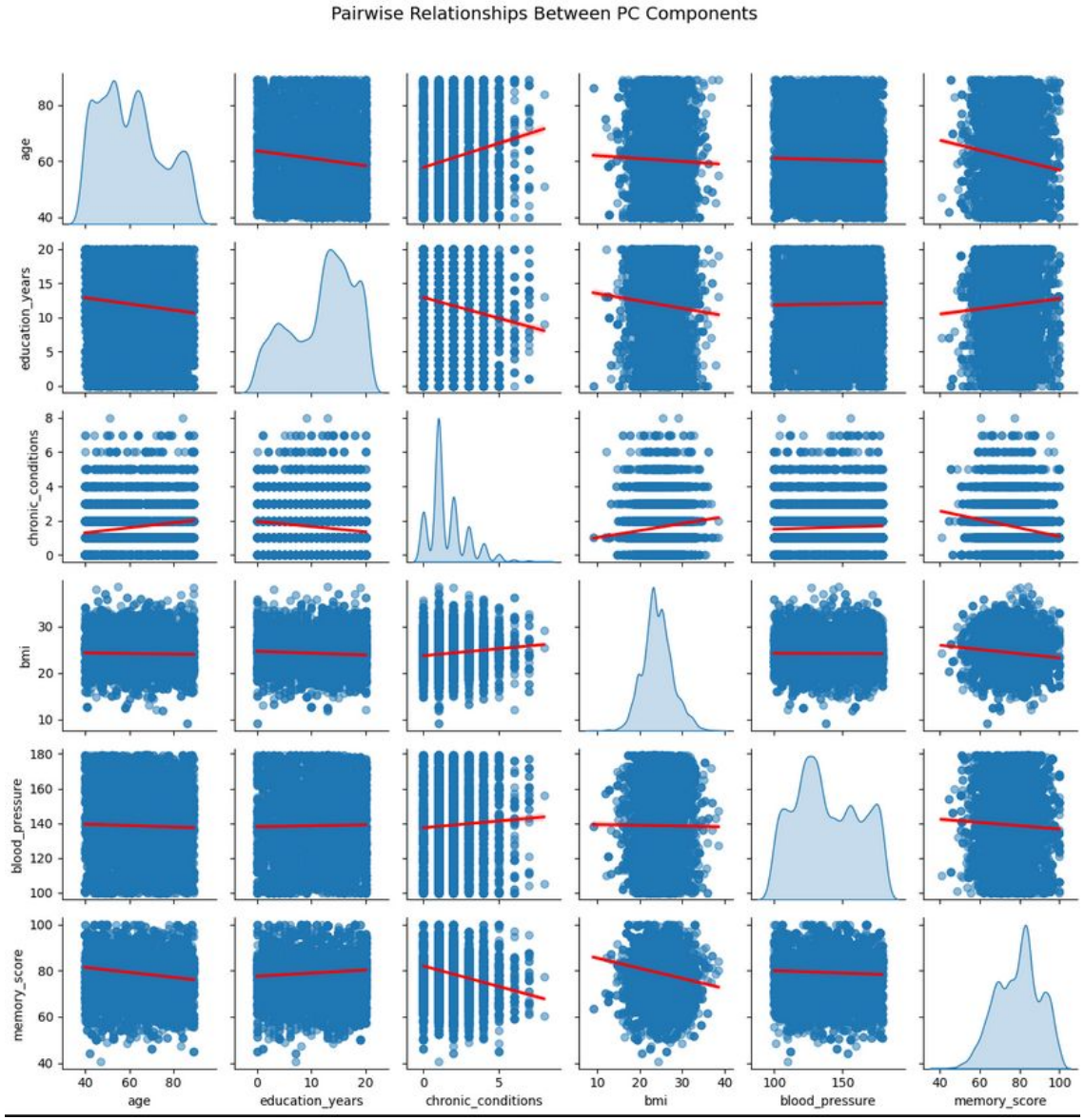


This dataset contains 3456 records and is intended for machine learning tasks, particularly classification and regression. It includes synthetic data on aging prediction, where the outcome variable is derived from several features related to an individual's health and lifestyle.

The dataset aims to assist in understanding aging patterns and predicting the likelihood of an individual falling into one of the three aging categories: Healthy, At-risk, or Frail.



Overall the dataset contained 5 numerical, 8 ordinal, and 2 nominal variables with target being an ordinal variable which allows us to explore methods of strict classification and ordinal regressions. We can apply for example Random Forest methods or choice regressions. Because of absence of linear relationships between numeric variables ordinal linear methods would not be considered



The dataset mainly consists of categorical and nominal variables. Additional inspection has showed that dataset is not intendent much for ecomoterics or tasks outside prediction. Most of numeric data is not normally distributed which inflicts certain limitations on interpretations..

Predictive modeling

In order to examine additional relationships we directly used **catboosting models**.

The catboost is used here because of its **fluid categorical** columns handling and **ani-overfit measures** built in.

Compared to other approaches it is produces much less overfitting like **RandomForests**, or **XGBoosting** or **AdaBoosting**. Unlike **XGBoosting** the performance won’t be affected as much by dataset size

Grid searching results

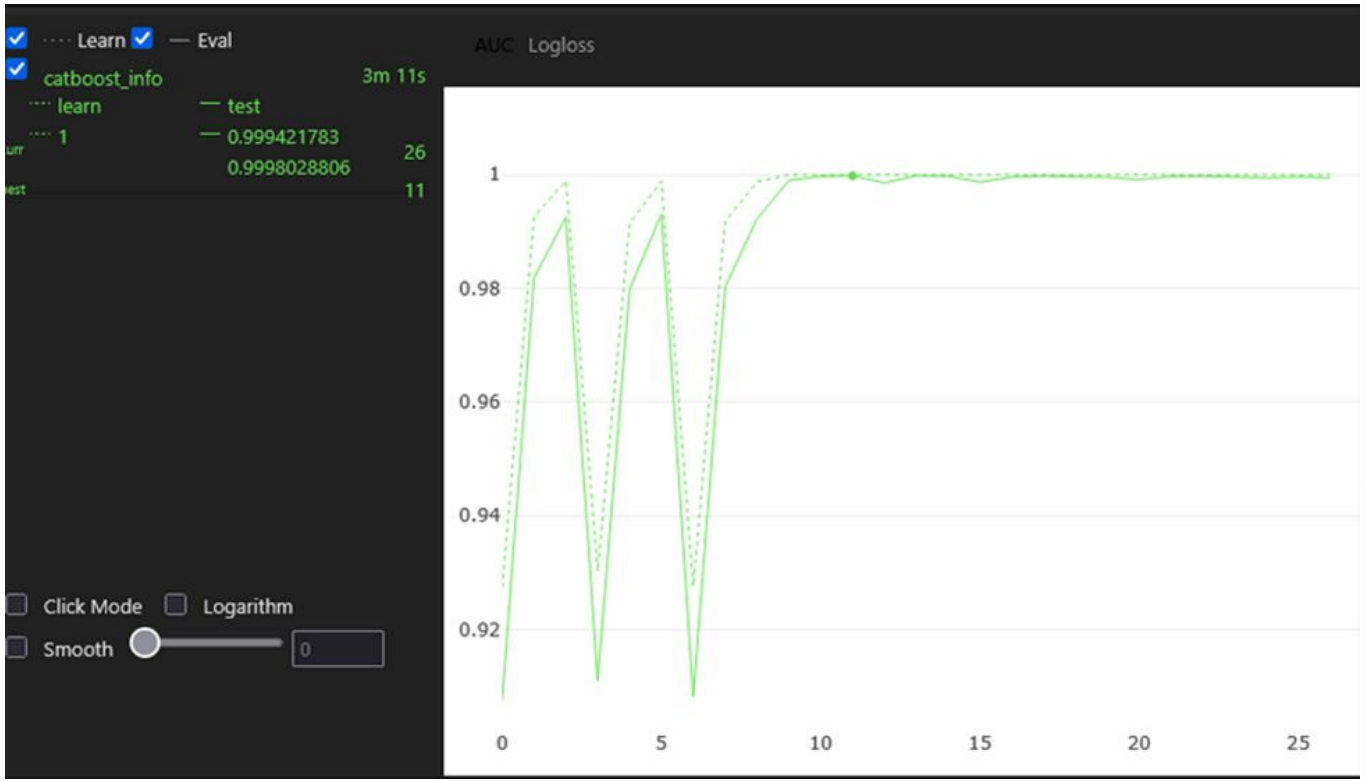
The most appropriate metric in for grid searching usually **AUC**. This metric is very reliable when assessing multiclass like this one.

The best group of parameters is:

{'depth': 5,

'learning_rate': 0.1,

'l2_leaf_reg': 1}

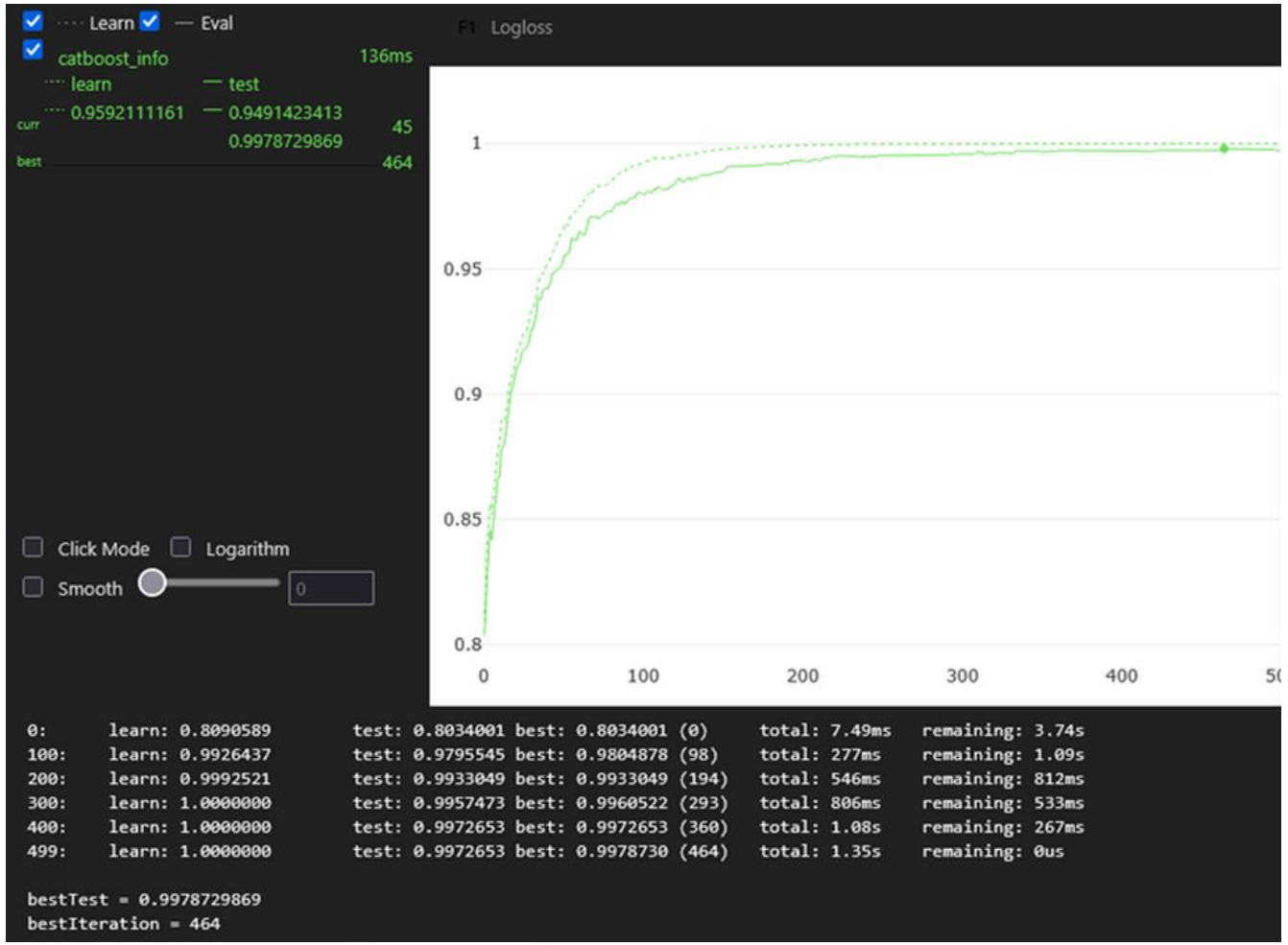


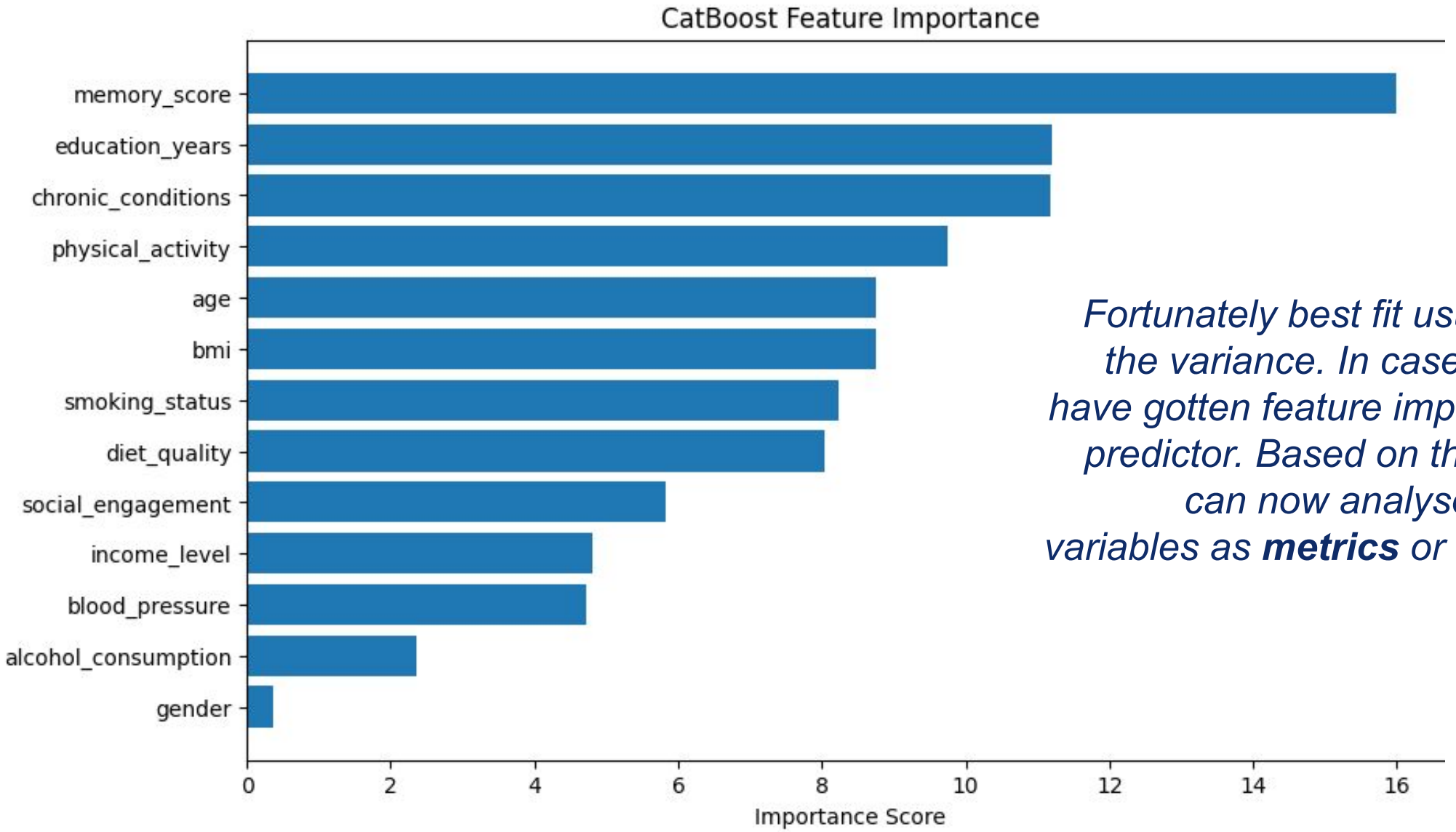
Prediction model design

Based on grid search results we trained best model with **F1** evaluation metric to ensure absence of **false positive**

(better safe than sorry).

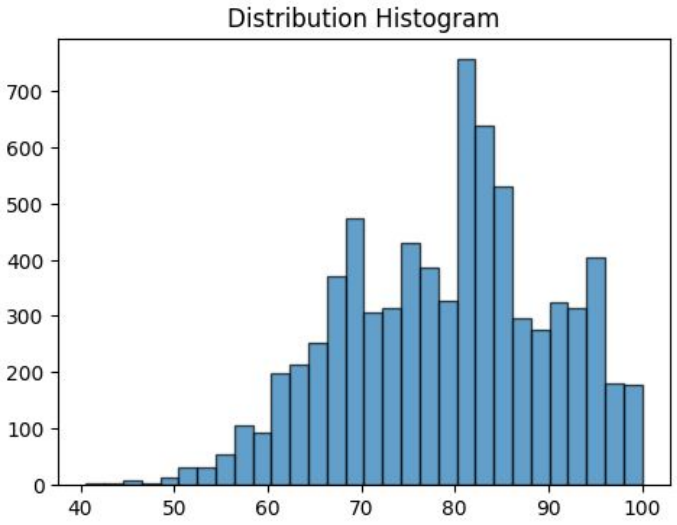
Evidentiality in all trials we ended up with what seems **overfitting models**. Unfortunately, there's no room to add artificial variance to combat overfitting.

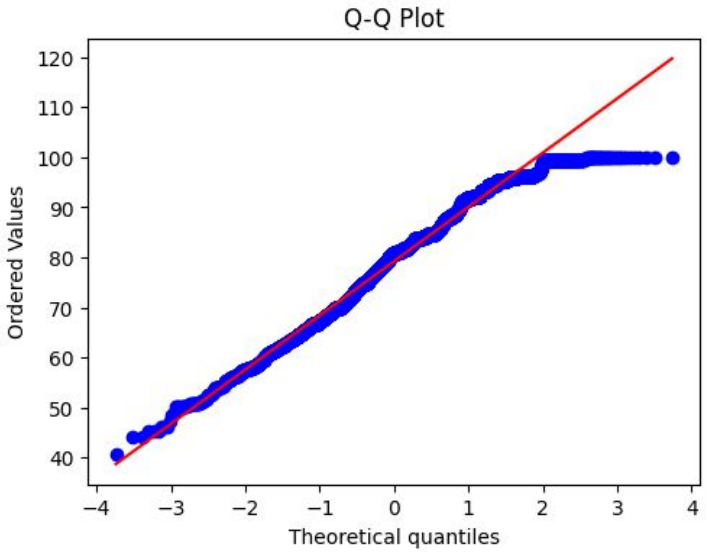




*Fortunately best fit usually explains all the variance. In case of CatBoost we have gotten feature importance for each predictor. Based on these features we can now analyse **importance** of variables as **metrics** or **causes** of aging outcomes*

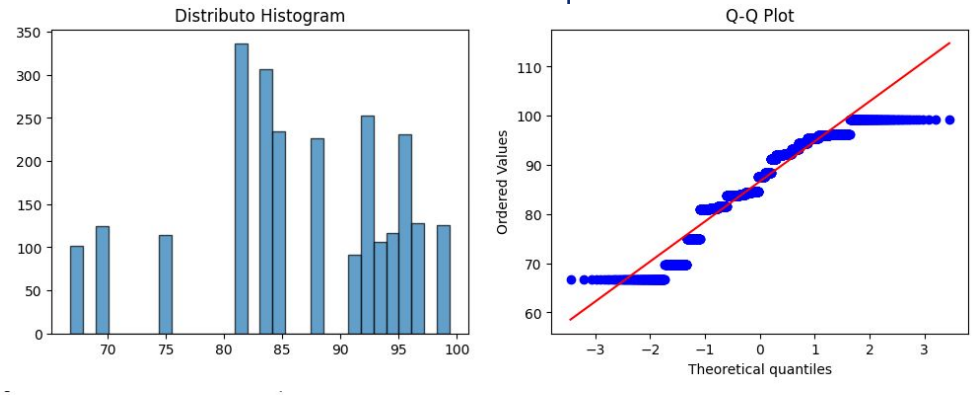
Memory score: the most reliable metric

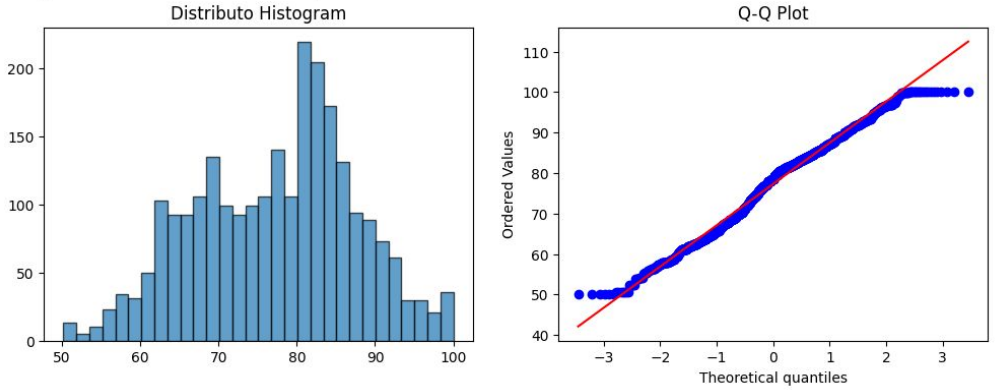


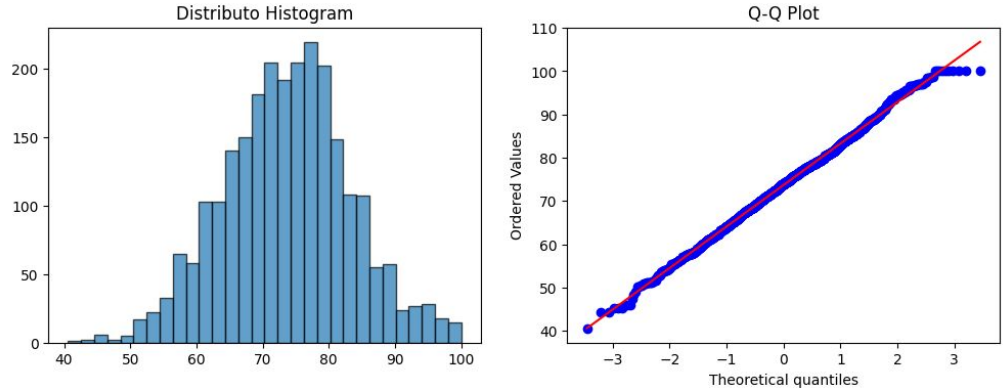


In order to examine differences between groups **ANOVA** was implemented.

aging_outcome	memory score mean
At-risk	77.28552017676857
Frail	73.68908996510517
Healthy	86.65273814173862







Memory score: not much of a “healthy”

We continue with ANOVA because there’s other evidence of high differences in MS for groups. Also we assumed that was ‘**Healthy**’ downsampled with regards to its distribution and its normality

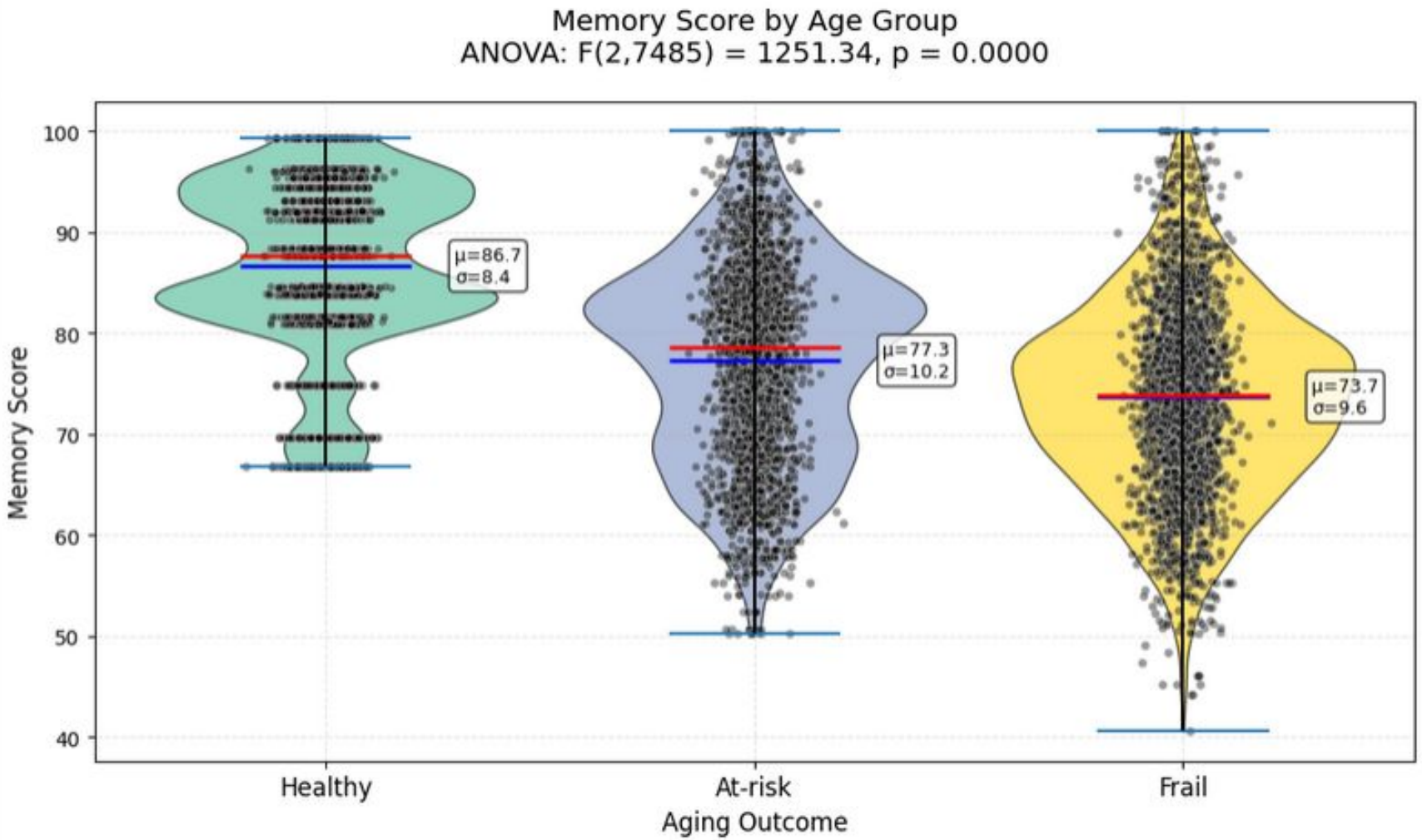
Results: Ordinary least squares

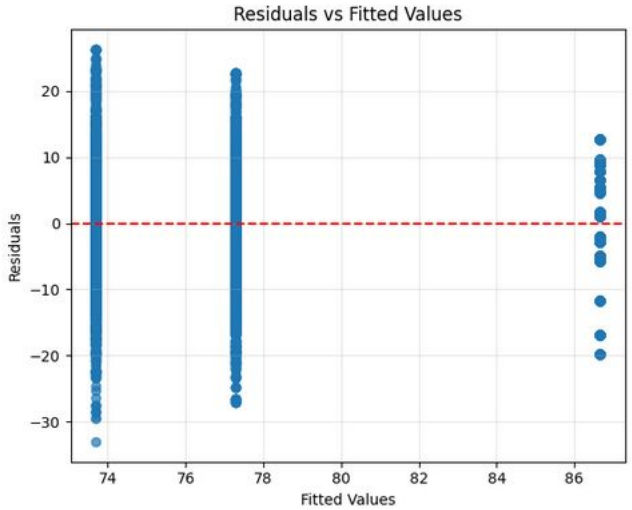
Model:	OLS	Adj. R-squared:	0.250
Dependent Variable:	memory_score	AIC:	54892.4545
Date:	2025-12-09 14:39	BIC:	54913.2177
No. Observations:	7488	Log-Likelihood:	-27443.
Df Model:	2	F-statistic:	1251.
Df Residuals:	7485	Prob (F-statistic):	0.00
R-squared:	0.251	Scale:	89.340

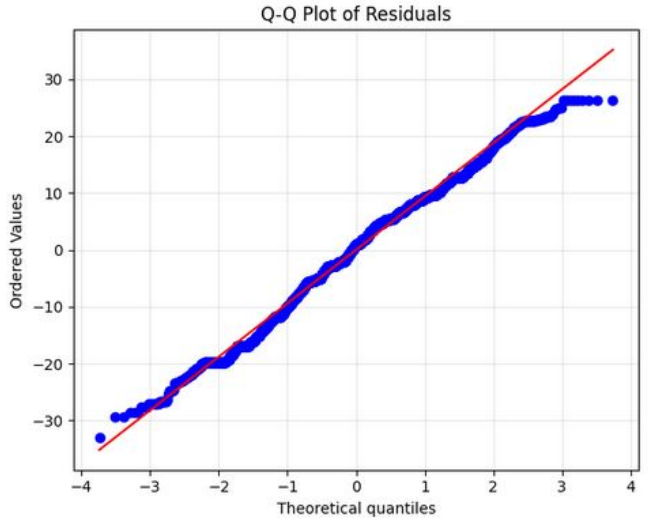
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	86.6527	0.1892	458.0179	0.0000	86.2819	87.0236
C(aging_outcome_cat)[T.Frail]	-12.9636	0.2676	-48.4521	0.0000	-13.4881	-12.4392
C(aging_outcome_cat)[T.At-risk]	-9.3672	0.2676	-35.0103	0.0000	-9.8917	-8.8427

Omnibus:	83.141	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	80.793
Skew:	-0.229	Prob(JB):	0.000
Kurtosis:	2.778	Condition No.:	4

Memory score: differences in distributions







Sources of risk: education in years

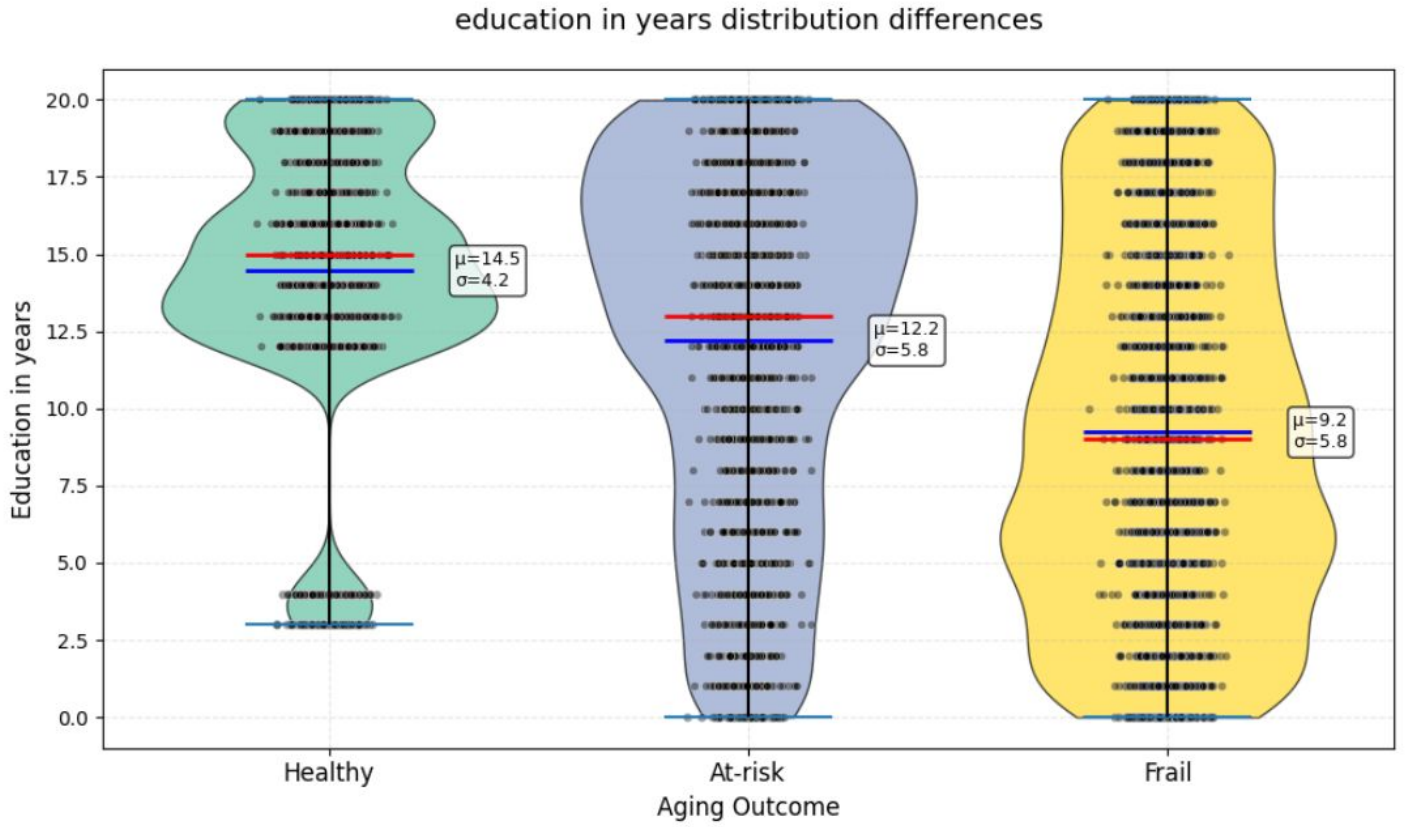
Parametric test Kruskal-Wallis:

```
Kruskal-Wallis Test:
  H-statistic = 953.738
  p-value = 0.0000
  Significant at α = 0.05

Effect Size:
  Epsilon-squared (ε²) = 0.127
  Interpretation: medium effect

Pairwise Comparisons (α = 0.05):
  At-risk vs Healthy: p = 0.0000 ***
  At-risk vs Frail: p = 0.0000 ***
  Healthy vs Frail: p = 0.0000 ***

Cliff's Delta Effect Sizes:
  At-risk < Healthy: δ = -0.201 (small)
  At-risk > Frail: δ = 0.286 (small)
  Healthy > Frail: δ = 0.509 (large)
```



Sources of risk: number of chronic conditions

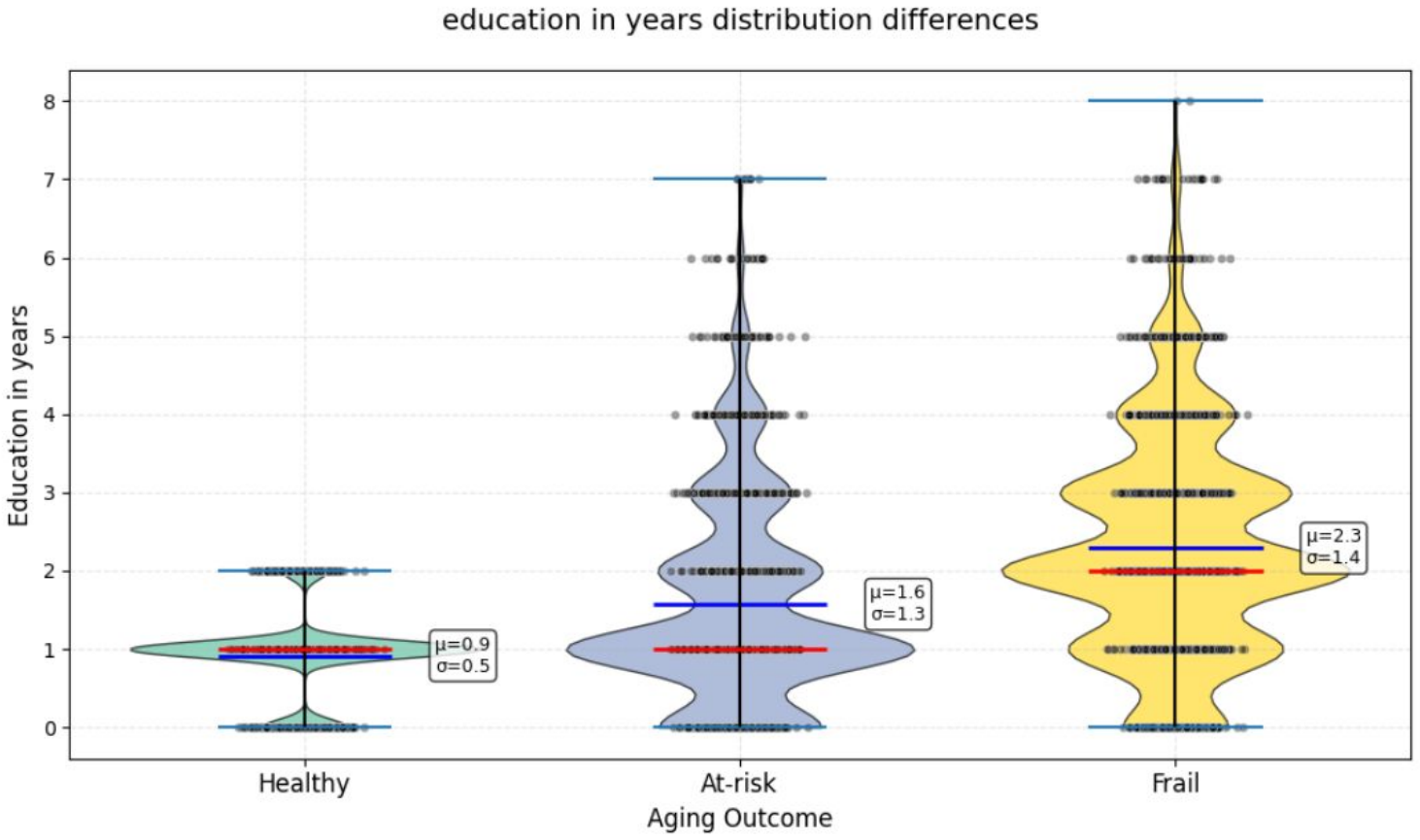
Parametric test Kruskal-Wallis:

```
Kruskal-Wallis Test:
  H-statistic = 1526.740
  p-value = 0.0000
  Significant at α = 0.05

Effect Size:
  Epsilon-squared (ε²) = 0.204
  Interpretation: large effect

Pairwise Comparisons (α = 0.05):
  At-risk vs Healthy: p = 0.0000 ***
  At-risk vs Frail: p = 0.0000 ***
  Healthy vs Frail: p = 0.0000 ***

Cliff's Delta Effect Sizes:
  At-risk > Healthy: δ = 0.268 (small)
  At-risk < Frail: δ = -0.312 (small)
  Healthy < Frail: δ = -0.618 (large)
```



Limitations

- Undersampling the Healthy class: Artificially pruning the records compressed the distributions (short boxes, waves in memory score), causing bias in ANOVA
- Model overfitting: CatBoost explains almost all the variance in the training data, but does not transfer well without data augmentation

What to do in the future?

Anti-overfitting: K-fold CV, early stopping in CatBoost, dropout; Add real data (NHANES).

Validation: External test set from real sources; SHAP for causal insights.

Conclusion

Most Important Predictive Factors (according to CatBoost):

· Top Predictors: memory_score (cognitive function), education_years (education), chronic_conditions (chronic diseases), and physical_activity.

· Least Important Factors: Gender (gender) and income level (income_level) had minimal predictive value.

Strong Statistical Relationships

· Memory Score: Showed a significant and graded decline from the Healthy group (highest) to the At-risk and then the Frail group (lowest). All pairwise differences were statistically significant.

· Education Years: Exhibited a clear downward trend from the Healthy to the Frail group.

· Chronic Conditions: Displayed a significant upward trend, with the Frail group reporting the highest disease burden.


HSE UNIVERSITY
EST 1992