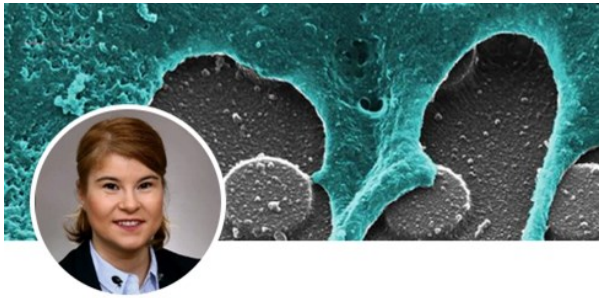
Diagnostic Prediction: Let’s make age just a number



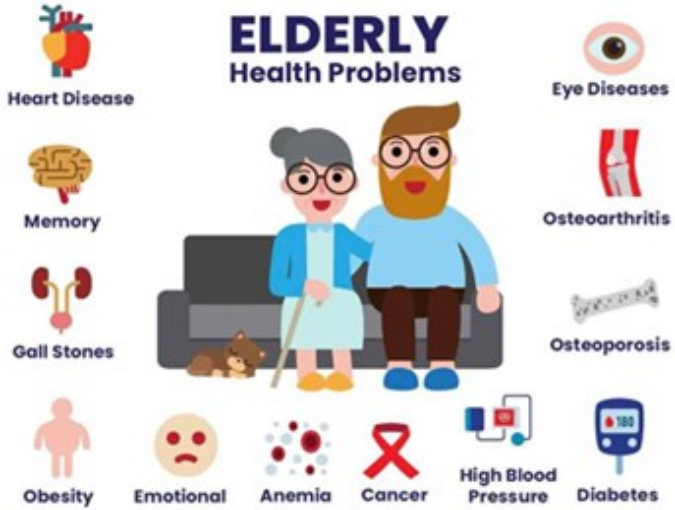
Andreea Belu
🎓 PhD in Natural Sciences | 📊 Data Scientist | 🧪 Chemistry & Biochemistry Specialist | 🚀 Product Development | 💻 Python, R

Final Project Data Science Bootcamp
03.08.2023

Aging

Physical and mental capacity

Risk of disease



- There are common health conditions associated with ageing.
- <u>The diversity seen in older age is not random.</u>

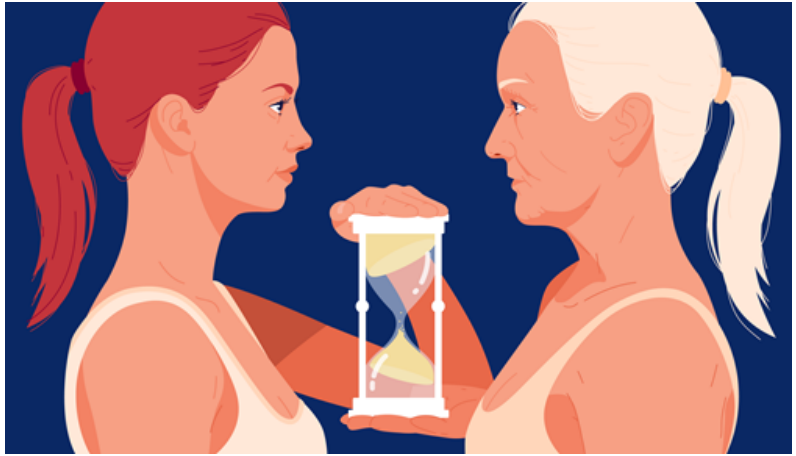
https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

Data science as method to predict aging-related diseases

Research using data science focused on:

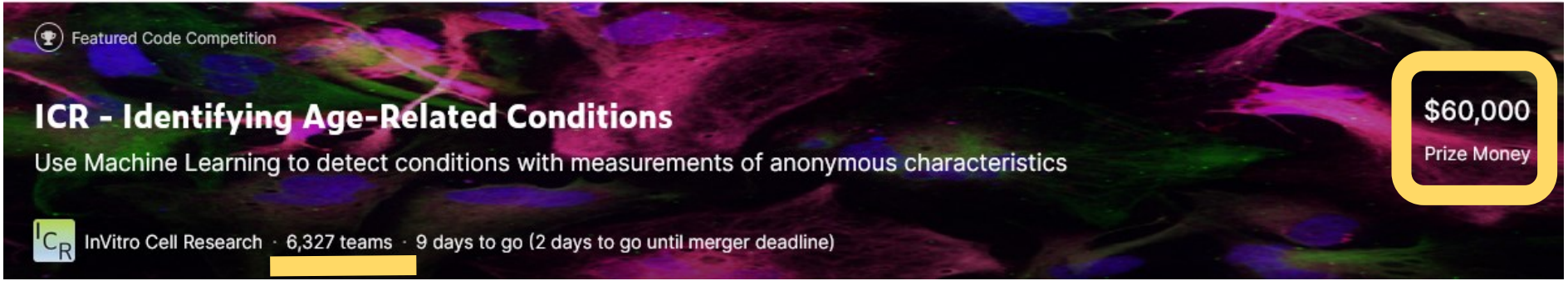
- slow down
- reverse
- prevent major age-related diseases

➢ improve healthcare outcomes



Wassan, J.T. et al, Role of Deep Learning in Predicting Aging-Related Diseases: A Scoping Review. Cells 2021, 10, 2924.

Kaggle competition



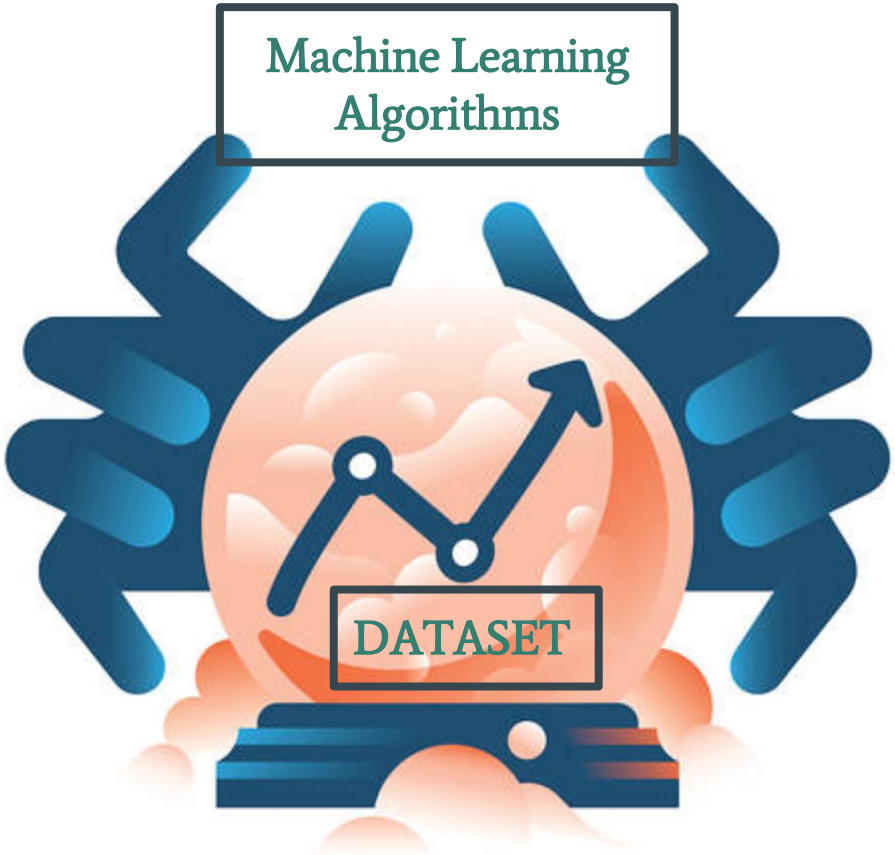
Goal: predict if a person has any of three medical conditions based on health characteristics.

Dataset:

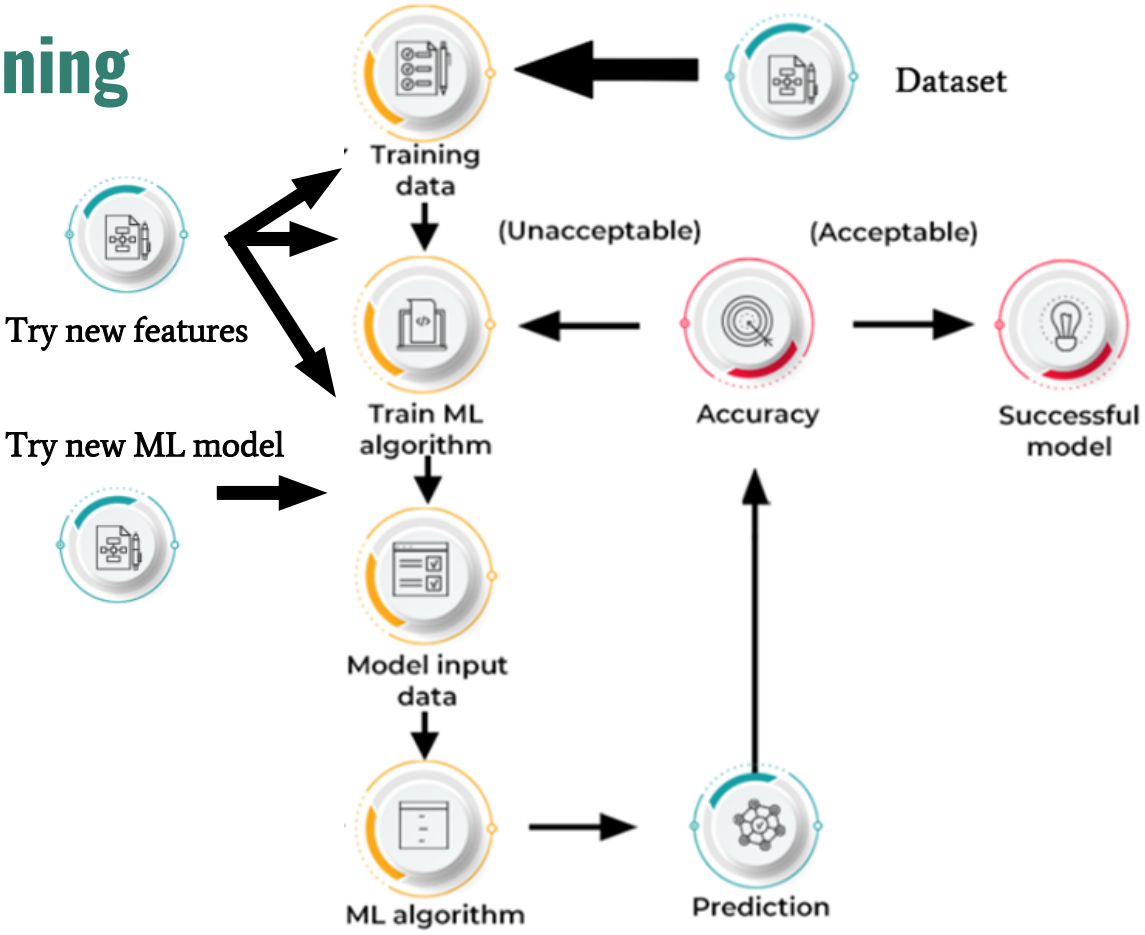
- Id — Unique identifier for each observation.
- AB-GL: — 56 anonymized health characteristics.
- Class: — - 1 subject has been diagnosed with one of the three conditions
- — - 0 no medical conditions

https://www.kaggle.com/competitions/icr-identify-age-related-conditions/overview

Prediction



Machine Learning



Results: Kaggle submissions

Submission and Description	Public Score
20230801_Kaggle_tabPFN_XGBoost_th_no-EJ - Version 1 Succeeded · 1d ago · Notebook 20230801_Kaggle_tabPFN_XGBoost_th_no-EJ | Version 1	0.06
20230731_Kaggle_tabPFN_XGBoost_th - Version 1 Succeeded · 2d ago · Notebook 20230731_Kaggle_tabPFN_XGBoost_th | Version 1	0.06
20230728_Kaggle_tensorflow_ - Version 3 Succeeded · 5d ago · Notebook 20230728_Kaggle_tensorflow_ | Version 2	17.26
20230727_Kaggle_tensorflow - Version 1 Notebook Threw Exception · 6d ago · Notebook 20230727_Kaggle_tensorflow | Version 1	
20230726_Kaggle_tabPFN_XGB_stack - Version 1 Succeeded · 7d ago · Notebook 20230726_Kaggle_tabPFN_XGB_stack | Version 1	0.52
20230725_Kaggle_tabPFN_2 - Version 1 Succeeded · 8d ago · Notebook 20230725_Kaggle_tabPFN_2 | Version 2	0.21
20230724_Kaggle_tabPFN - Version 1 Succeeded · 9d ago · Notebook 20230724_Kaggle_tabPFN | Version 1	0.07
20230721_Kaggle_1Model_feature-selection - Version 2 Succeeded · 10d ago · Notebook 20230721_Kaggle_1Model_feature-selection | Version 2	0.34
20230721_Kaggle_1Model_feature-selection - Version 1 Succeeded · 12d ago · Notebook 20230721_Kaggle_1Model_feature-selection | Version 1	0.48
20230720_Kaggle_XGBoost - Version 1 Succeeded · 13d ago · Notebook 20230720_Kaggle_XGBoost | Version 1	0.26
Notebook_starter20230718 - Version 3 Succeeded · 14d ago · Notebook Notebook_starter20230718 | Version 3	0.31

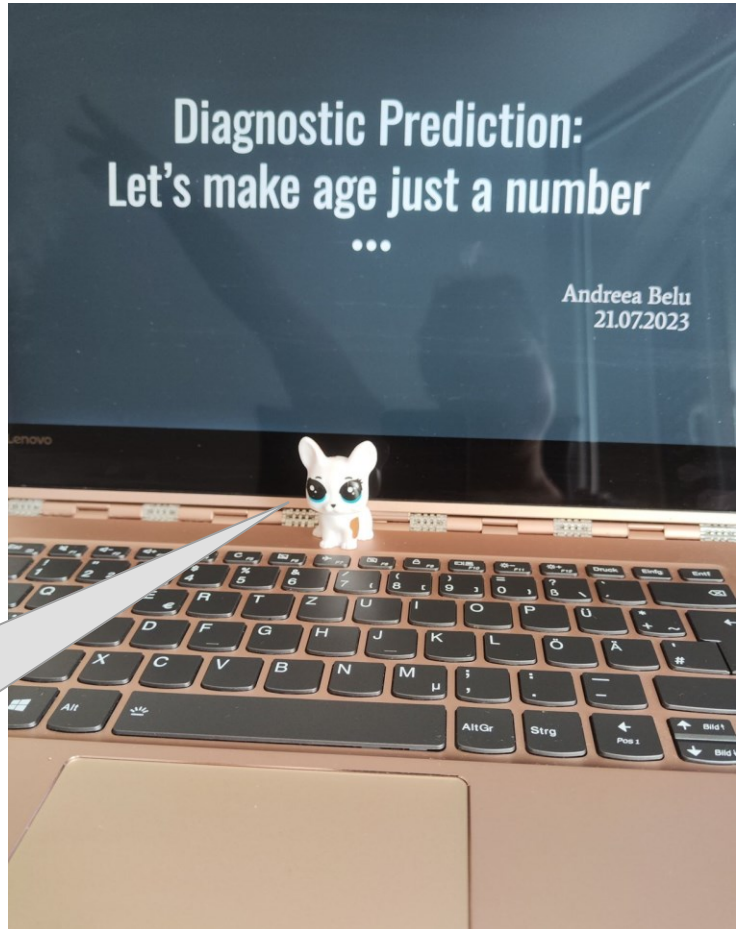
Code Competition:
- Log loss lower as possible
- Score = 0.00 was achieved
- 58% of test set is hidden
- Final standings may be different

Present Results:
- Best score = 0.06
- Place ~1200 out of 6300 teams
- 8 days to go

Amazing time!

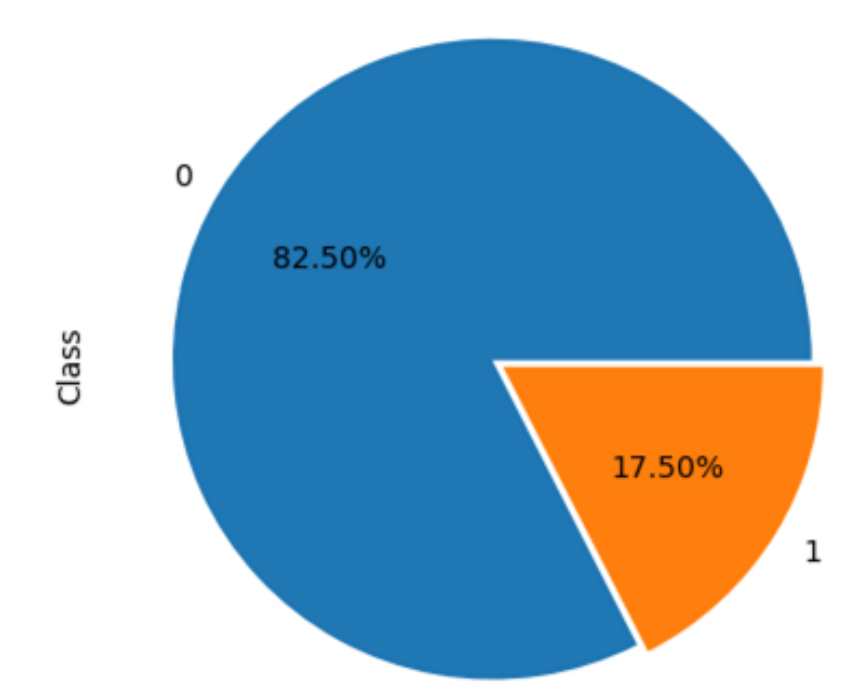
- 😊DATA 17 Team
- 😊Henrik
- 😊Vasil
- 😊WBS Community

Thank you!



Extra slides

Dataset is heavily imbalanced



Total: 617

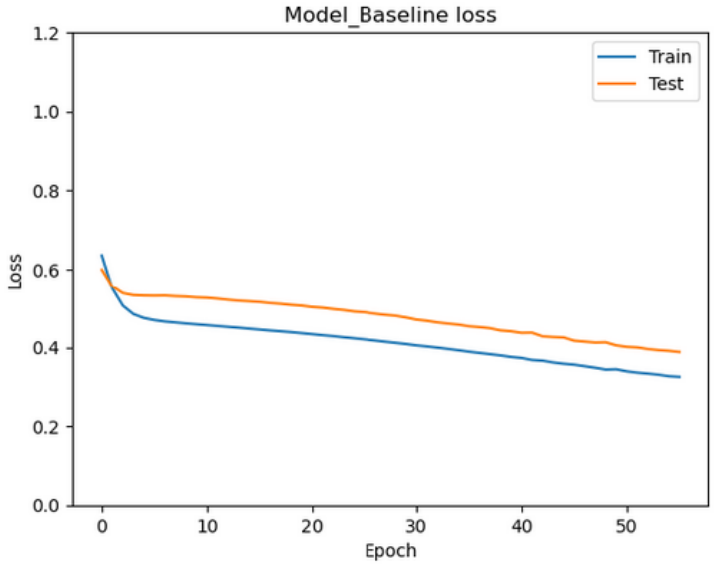
Best ML model

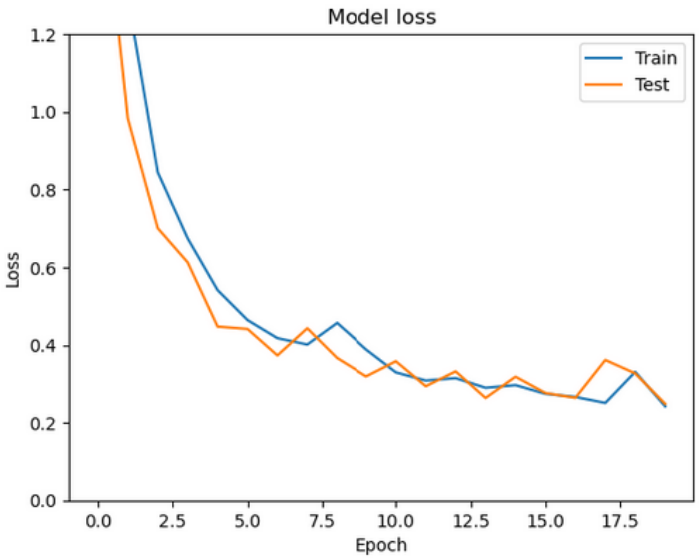
Best_Result_tabPFN-XGBoost-stack_score=0.06

- Random undersampling for the balance the "Class" distribution (train dataframe).
- Two cross-validation strategies using K-Fold in order to get a more reliable estimate of the model's performance on unseen data and avoid overfitting.
- Ensemble model that combines the predictions of two classifiers: XGBClassifier from XGBoost and TabPFNClassifier from the TabNet framework.
- Training function was performed by training and evaluation of a given model using nested cross-validation.
- Random oversampling with respect to the greeks dataframe.
- Final training on the balanced dataset that assess the model's performance and returns the best trained model.
- Best trained model was used to make predictions on test data.

Dive into Neural Network - not really beneficial

score: 17.26





Sequential with 2 layers, 1000 neurons/layer, epochs = 56

Train accuracy: 86%
Test accuracy: 79%

Sequential with 3 layers, 1000 neurons/layer, epochs = 20, regularizer, dropout

Train accuracy: 96%, loss: 0.2153
Test accuracy: 85%, loss: 0.3008