Machine Learning in Computational Biology – 1st Assignment

Rafail Adam

7115152400009

Abstract

In recent years, the microbiome has become a very important indicator of health and disease, so much so that it has been associated with diseases not only correlated with the gut, but also including autoimmune diseases, neurodegenerative diseases and even cancer treatment[1–3]. In this study, we are examining the predictive power of the gut microbiome composition from several individuals for calculating their Body-Mass index (BMI) via three different machine learning algorithms for regression.

Github Link:
https://github.com/RafailTn/Assignment-1

Introduction

Historically, the first to introduce BMI (initially called the Quetelet index) was Adolphe Jacque Quetelet in 1835, who noted that "*the body mass relationship to height in normal young adults was least affected by height when the ratio of weight to height squared was used rather than merely using the ratio of the weight to height or weight to height raised to the third power"*[4]. In 1972 Keys et.al[4] popularized the Quetelet index and the formula for its calculation was finalized to ***body weight (kilograms) divided by height squared (meters)***. Since then, the term Body Mass Index is used interchangeably.

Body Mass Index (BMI) is a widely known metric that is mostly used as a preliminary screening tool. Taking into account an individual's height and weight, BMI places them into one of seven (according to WHO) distinct categories:

1. Severely Underweight: <16 kg/m^2
2. Underweight: 16.0 to 18.4 kg/m^2
3. Normal weight: 18.5 to 24.9 kg/m^2
4. Overweight: 25.0 to 29.9 kg/m^2
5. Moderately Obese: 30.0 to 34.9 kg/m^2
6. Severely Obese: 35.0 to 39.9 kg/m^2
7. Morbidly Obese: ≥40.0 kg/m^2

Individuals classified as underweight and morbid and severely obese have increased risk of adverse health outcomes. Specific BMI ranges have been associated with higher mortality[5] as well as poor responses to several diseases and/or medication[6–8].

Gut microbiome has become an increasingly important factor for matters of health and disease. In the early 1900s Eli Metchnikoff associated the longevity of rural Bulgarians to the consumption of fermented milk, proposing that bacteria in lactic acid provide anti-ageing effects. He later named the bacterium *Lactobacillus bulgaricus*[9].

One of the most important efforts to understand the effect of the gut microbiome in health and disease is the Human Microbiome Project[10,11]. The HMP used multiomics data to characterize the effect of the microbiome in human health. Thus far it has reported the structure and function of the microbiome of more than 300 healthy individuals at 18 body sites from a single point, while the gut microbiome project has contributed significantly to the knowledge of how various types of microorganisms in the gastrointestinal tract affect human health. Finally, the microbiome is hardly static and is influenced by many factors including but not limited to age, diet and disease[12,13].

Recently, the contribution of the microbiome to many diseases has been uncovered, ranging from neurodegenerative and autoimmune diseases to cancer[1–3]. Due to this ever growing importance of the microbiome and the dependence of BMI on diet as well as its wide use, we decided to examine the potential use of the gut microbiome in predicting the BMI of a person as a regression problem. This will be attempted via the use of three different machine learning algorithms for regression:

1. ElasticNet
2. Bayesian Ridge
3. Support Vector Regression

As these are simple models, relative to deep neural networks, they will give us the opportunity to examine if there are simple correlations between specific microbial species and BMI that allow for its approximation or if more complex relationships are at play.

The first and most simple machine learning algorithm that we are using is ElasticNet[14]. ElasticNet Regression is a regularization technique that utilizes both Lasso (L1) and Ridge Regularization (L2). Lasso regularization adds a penalty equivalent to the absolute value of the coefficients which reduces some coefficients to zero, introducing sparsity. Ridge/L2 regularization adds a penalty equivalent to the square of the coefficients which reduces large coefficients and helps reduce collinearity.

Bayesian Ridge Regression is an extention of classical Ridge Regression that utilizes the Expectation Maximization algorithm. By modelling the weights, their precision and the precision of the noise as random variables, we end up forming a Gaussian posterior distribution that is dependent on the precision of weights and the precision of the noise. These are then estimated via using the EM algorithm while utilizing a gamma prior and using type 2 maximum likelihood to update the two uknown variables by maximizing the evidence[15].

Support Vector Regression (SVR) is an extension of the classical use of Support Vector Machines, which were initially developed for classification[16], for regression tasks. SVR estimates a function by introducing a margin (epsilon) that defines the maximum deviation within which the actual values fall. Furthermore, SVR utilizes the kernel trick which maps data to

a higher dimensional space according to specific functions (eg Linear, Polynomial, Radial Basis Function).

Results

Data Exploration

The dataset that was used is comprised of 700 samples in total and contains information about the Sex, Age and measurments for more than 130 different microorganisms in the gut coming from metagenomics studies. The goal of this study is to examine the correlation of these microorganisms with BMI by using the metagenomics data as features for accurate BMI prediction.

Before trying any of our models, we examined the dataset to evaluate its quality and perform data cleaning. Among the first characteristics that we can observe is that our dataset does not contain the same number of male and female samples (302 and 186 respectively), which does make it less comparable with the real world distribution, while at the same time, the age distribution is significantly different among the male and female samples. These characteristics are important as age and sex affect the microbiome of a person[12,13,17].

From the following figures, differences in some microbial measurments can be observed between male and female samples (eg. Akkermansia muciniphila) which further indicates that the bias in the distribution of sexes affects also microbiome data. Since the goal of the study is to study how microbiome data correlates to BMI, all other features were dropped.

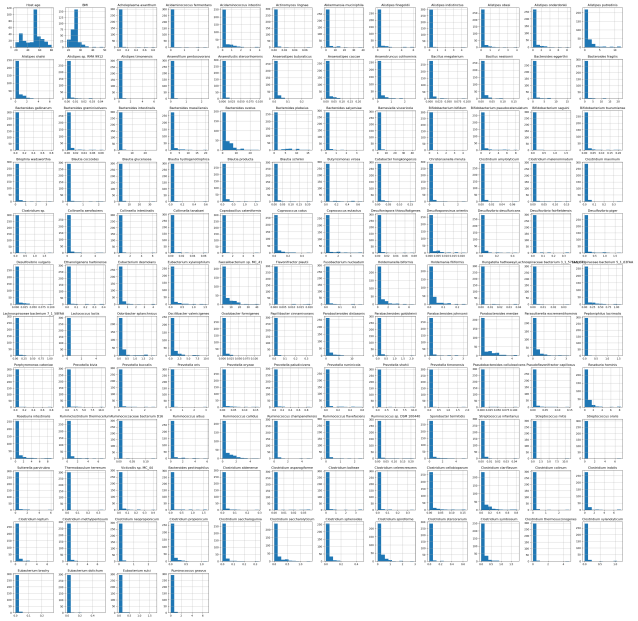

Figure 1. Age and microbiome data for male samples.

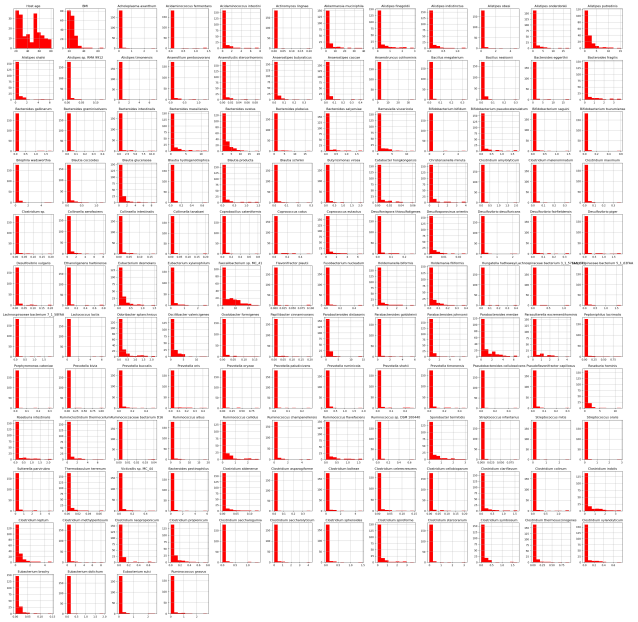

Figure 2. Age and microbiome data for female samples.

The same plots were also constructed for the evaluation set (please check the model_analysis notebook for this and additional plots). The evaluation set also has a bias for male samples being more than female samples, with a different

ratio, while the age distribution is also different. This should not pose a problem as ideally our models should be able to generalize to new datasets with varying distributions among their features.

Finally, for the data exploration step, a plot for BMI against age and sex was constructed to visually examine their relationship in the dataset. Based on it, sex and/or age do not seem to be able to separate BMI into groups.

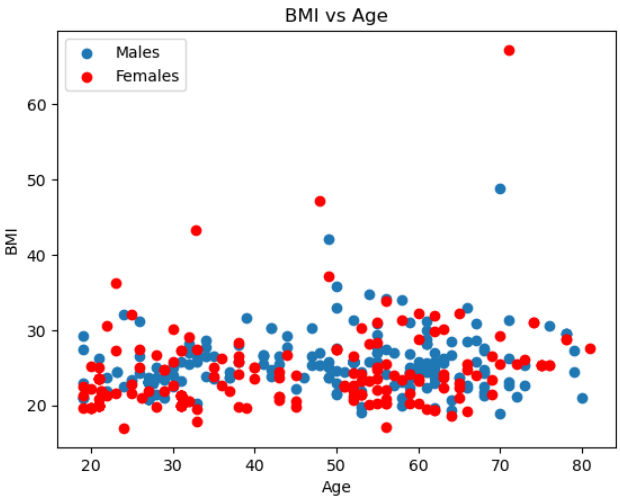

Figure 3. BMI plotted against Age and Sex.

Despite sex and age being excluded in the final steps of preprocessing, sex proved to be an important factor at least in some initial, experimental trials made for feature selection. Thus, a future goal would be to include sex in the features as it is likely to reduce the number of necessary microbial features and reduce the error metrics. Since the decision of the most important microbial features will be discussed more thoroughly in the feature selection section, additional plots for relationships between microbiome data and BMI were not constructed.

Model Analysis

The baseline model results are shown below, the reasoning behind the choice of the metrics is explained in the Methods section. The three models examined are ElasticNet, Bayesian Ridge and Support Vector Regression.

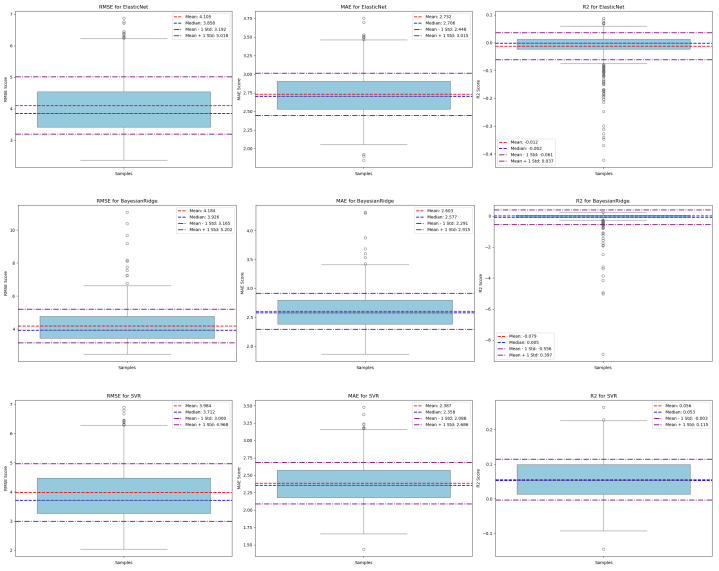

Figure 4. RMSE, MAE and R^2 boxplots for baseline models (200 kfold cross-validation).

Baseline models show an RMSE around 4, MAE scores around 2.5 and very low R^2 values. More specifically, ElasticNet seems to be the worst performing model as its RMSE and MAE are slightly higher than Bayesian Ridge, while their R^2 score is essentially 0 showing that the quality of predictions might be even worse than always predicting the mean. The SVR model is significantly better judging from the mean and median of the plots, indicating that this model might be promising. It is important to note however that the R^2 score has a higher standard deviation than ElasticNet and Bayesian Ridge, which shows that the predictions of this model are not always reliable. In order to mitigate this issue trying different kfold strategies and plotting

results per fold is needed, in order to determine where the problem arises. For now, we will proceed with the evaluation boxplots, in order to also examine potential issues arising regarding generalization.

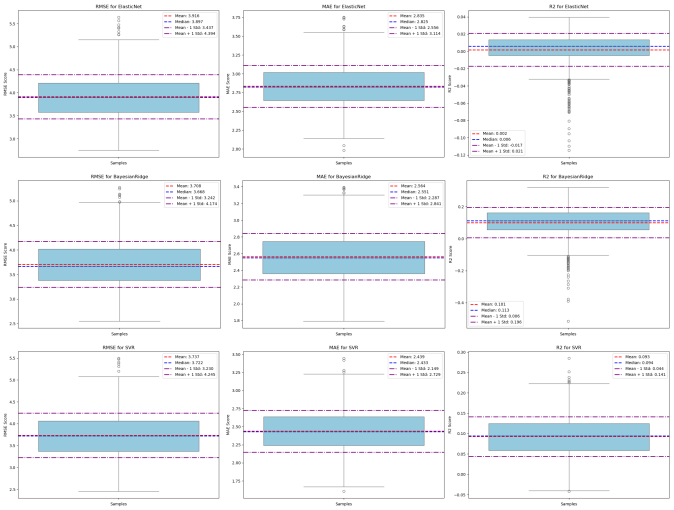

Figure 5. Evaluation boxplots by bootstrapping(1000fold).

From the evaluation boxplots above, we can see that the RMSE scores have lowered for all of the models, with the SVR retaining the lowest error score. MAE is also reduced for Bayesian Ridge and SVR but is higher in ElasticNet, indicating potential overfitting. R^2 is increased for all models and above 0, in fact for Bayesian Ridge and SVR it has improved by 0.1 which shows that these models might be able to generalize well to new data after feature selection and fine-tuning.

Feature Selection

Several feature selection strategies were attempted, including Recursive Feature Selection, PCA, KernelPCA, Mutual Information/Pearson based feature selection as well as a brute force method that checks all possible combinations of N features for each model. Initially PCA was performed in order to determine the number of features that explain 95% of the variance and help make decisions regarding the next steps (95 features). Despite the percentage of variance explained by each feature being miniscule, we attempted to use Optuna to search for the best feature selection method based on RMSE minimization. This can lead to overfitting but that will become apparent from the resulting kfold and evaluation boxplots. Based on kfold, PCA with 47 features was deemed to be the best feature selection method for ElasticNet and Bayesian Ridge while PCA with 90 features was the best for SVR. Choosing these methods yields slightly lower RMSE and MAE scores as well as higher R^2 scores in kfold cross-validation, with the exception of ElasticNet.

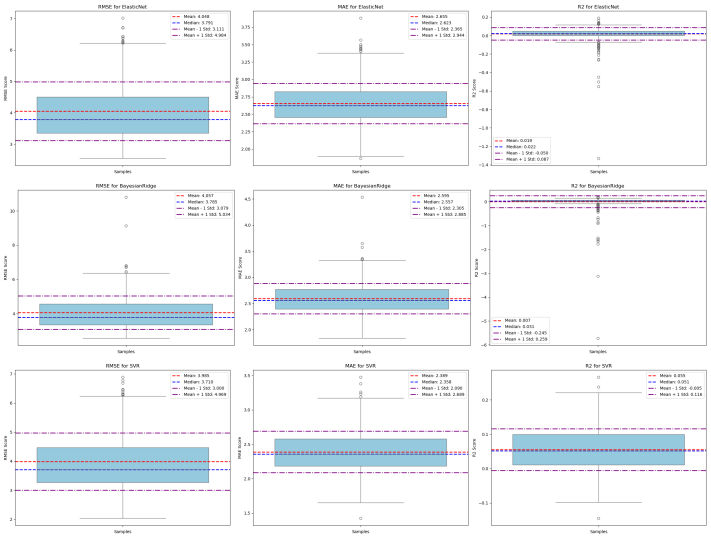

Figure 6. Kfold Boxplots after PCA.

After performing bootstrapping on the evaluation set, we can observe that for ElasticNet and Bayesian Ridge there is a slight reduction in RMSE and MAE however R^2 has not improved. SVR metrics seem to have remained stable after

feature selection indicating that it did not compromise the model.

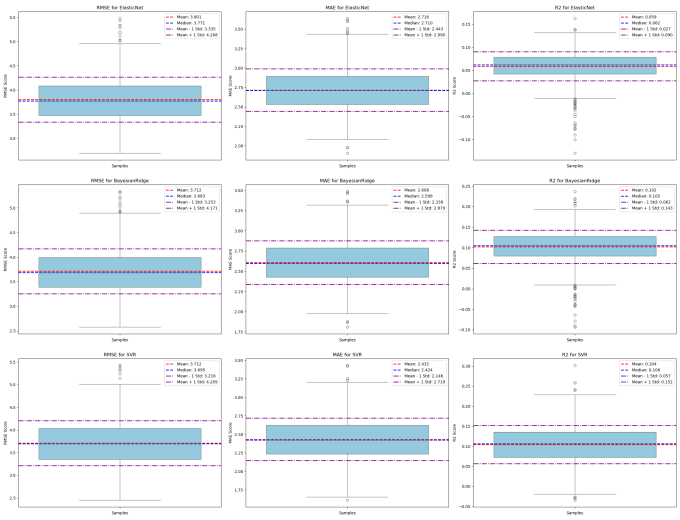

Figure 7. Evaluation boxplots by bootstrapping after PCA (1000 fold)

Fine-Tuning

In order to perform fine-tuning, Optuna was used to suggest the best hyperparameters for each model. Since the sampler it uses by default utilized a Bayesian approach, a different one will be assumed in the bonus question that does so differently. The different ranges for each hyperparameter of each model are shown below.

1. ElasticNet
    - Alpha : 0.1 -10.0
    - L1 Ratio: 0-1
2. Bayesian Ridge
    - Alpha1: 1e-8 – 1e-4
    - Alpha2: 1e-8 – 1e-4
    - Lamda1: 1e-8 – 1e-4
    - Lamda2: 1e-8 – 1e-4
3. Support Vector Regression
    - C: 0.1 - 10.0
    - Gamma: auto/scale
    - Epsilon: 0.01 - 1.0
    - Kernel: Linear/Poly/RBF
    - Degree: 2 – 5
    - Coeff0: 0 – 1

After retrieving the Optuna suggested hyperparameters, kfold and evaluation boxplots were constructed to judge the magnitude of improvement.

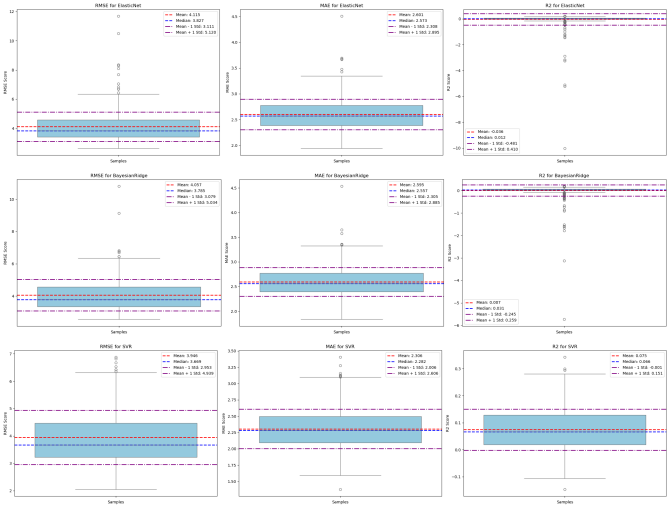

Figure 8. Kfold on fine-tuned models

The results of fine tuning based on kfold cross-validation regarding show that it did not manage to significantly improve the models as ElasticNet and Bayesian Ridge have remained largely the same whereas SVR has only slightly improved in terms of R^2. These results indicate that these models are likely overfitted, especially when only utilizing microbiome data.

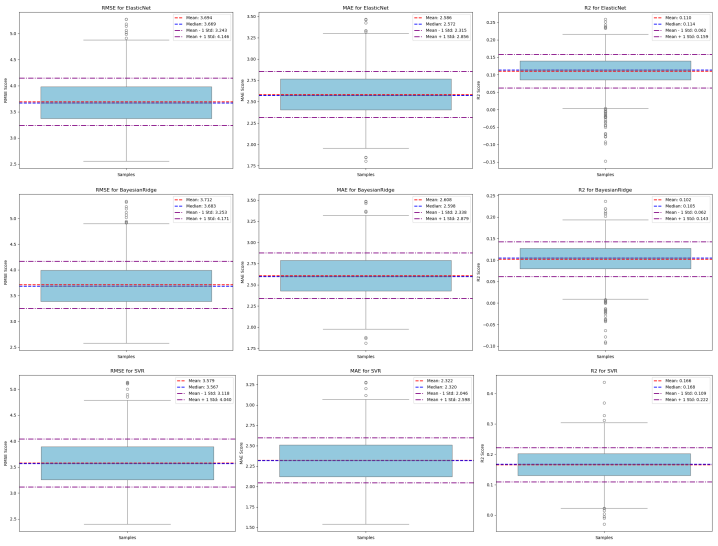

Figure 9. Evaluation boxplots after fine-tuning (1000 fold)

Evaluation boxplots show better scores for the models indicating that training on the entire dev set and predicting on the evaluation set leads to good results of relatively high reliability. Overall, the results are not substantially improved, even after fine-tuning. The best model seems to be SVR as its metrics remained relatively stable throughout, with the feature selection method being PCA with 90 features. Since SVR is the most complex model out of the three, the interpretability is reduced compared to the other two while at the same time PCA being chosen leads to a compromised capability to choose specific features as more than one of them contribute to the construction of Principal Components. It is even more challenging if one considers the fact that we need 95 Principal Components to explain 95% of the variance all of which explain a low percent of the variance.

Methods

All of the methods used are found in pre-existing python libraries. Scikit-Learn [18] was used extensively as it provides a vast number of machine learning tools and methods including all the algorithms that were used by the models.

ElasticNet is the simplest one of the three models used in this exercise. It solves ordinary least squares while utilizing 2 regularization techniques L1 (lasso), removes less important features (performs feature selection) and L2 (Ridge) shrinks coefficients to prevent overfitting[14]. In scikit learn, a coefficient is specified that modifies the solution to be different than ordinary least squares (alpha).

Bayesian Ridge performs ridge regression while applying a bayesian framework to estimate regression coefficients. Briefly, it assumes Gaussian priors for model coefficients, updates the prior using Bayes' Theorem to obtain a posterior distribution, uses an empirical Bayes approach to estimate α\alpha and λ\lambda and predicts the output by using the mean of the posterior distribution[15].

Support Vector Regression is an extension of Support Vector Machines (SVM) for regression tasks. It aims to find a hyperplane that best fits the data while allowing for some tolerance of error[16].

Metagenomic features were scaled using Standard scaler from scikit learn which applies the same formula as z-scaling.

PCA, KernelPCA, SelectKBest, RFECV and SFS were used also directly from scikit-learn and comparisons were made between them based on an initial limited number of kfold cross validations (10) to find the best method, utilizing Optuna. Recursive feature selectors like RFECV and SFS were not included as they need to be set up separately, so the boxplots were compared with the best result of the non-recursive options.

By default, Optuna uses a sampler called Tree Parzen Estimator. TPE is a Bayesian optimization method that builds a probabilistic model of the objective function to guide the search process. TPE models the search space

using L(x) which models the distribution of hyperparameters that led to the best results (lower loss values) and g(x), which models the distribution of all other hyperparameters. Then Optuna splits previous trials into trials where the objective function (e.g., RMSE) was low and trials where the objective function was high. The split is determined using a threshold γ\gamma (default: 10% best trials). If a hyperparameter value good trials, it has a high l(x) and it is more likely to be selected [19].

The metrics that were used to evaluate the models are the following:

Root Mean Squared Error (RMSE) is a measure that quantifies the error between the predicted and actual values. It is calculated by taking the square root of the mean of the squared differences between the predicted and actual values. The lower the RMSE, the better the model's performance. Moreover, RMSE is sensitive to outliers, which can skew the results.

Mean Absolute Error (MAE) is a measure that also quantifies the error between the predicted and actual values. It is calculated by taking the mean of the absolute differences between the predicted and actual values. The lower the MAE, the better the model's performance. MAE is less sensitive to outliers than RMSE. Ideally, we aim to optimize both RMSE and MAE to achieve the best model performance.

R2 score measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It quantifies the goodness of fit, indicating how well the model explains the variance in the target variable. The maximization of this score is desirable. When this score is negative, it indicates that the model's fit is worse than a random guess.

RMSE was chosen as via its minimization it lets us minimize the number of errors. However due to it not being robust to outliers MAE is also used. Ideally both of these metrics should be reduced for the models to be optimized. It is expected however that after a threshold value they will not converge together as minimizing the divergence introduced by outliers, which have the greatest magnitude of error, leads to an increase in MAE. R^2 is used to monitor the reliability of the predictions of the models.

Discussion

The gut hosts a vast variety of microbes whose concentration has been associated in many ways with health and disease. BMI is a characteristic used for early screening based primarily on height and weight, that classifies people on seven different classes. Increased and decreased BMI than that of the normal class shows increased risk for several health issues.

By using metagenomics data from over 100 different gut microbial species we attempted to test 3 different machine learning algorithms to predict BMI as a regression task. ElasticNet, Bayesian Ridge and SVR are popular algorithms for regression that are intriguing to test when

faced with an uknown regression task before proceeding with more complex models.

The performance on these models on the regression task were underwhelming, even after fine-tuning, which can be attributed either to the models themselves or to the dataset and task at hand. Given the correlation of Sex and Age with BMI it would be interesting and necessary to check if including these features will improve the prediction quality. Additionally, there was significant underrepresentation of samples for severely underweight and severely overweight people and an imbalance in male vs female samples. Improvements in the comparison method of the boxplots could help us get a better picture on the significance in model metric differences, for example one based in p-values rather than inspection of mean and median. Another interesting option would also be to use Optuna to optimize for more than one score at the same time. Additionally, plotting each fold might also looks like a promising potential improvement, that would give us insight into problematic folds and allow us to consider better strategies. Finally, it is possible that the task at hand is also difficult and that making it a classification task would be more suitable.

Bibliography

1. Borbolis, F., Mytilinaiou, E. & Palikaras, K. The Crosstalk between Microbiome and Mitochondrial Homeostasis in Neurodegeneration. *Cells* **12**, 429 (2023).
2. Villemin, C. *et al.* The heightened importance of the microbiome in cancer immunotherapy. *Trends Immunol.* **44**, 44–59 (2023).
3. De Luca, F. & Shoenfeld, Y. The microbiome in autoimmune diseases. *Clin. Exp. Immunol.* **195**, 74–85 (2019).
4. Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N. & Taylor, H. L. Indices of relative weight and obesity. *J. Chronic Dis.* **25**, 329–343 (1972).
5. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *Lancet Lond. Engl.* **388**, 776–786 (2016).
6. Yanbing, L., Zijun, L., Hongbo, Z. & Zhi, W. Relationship between BMI and chemotherapy-induced peripheral neuropathy in cancer patients: a dose-response meta-analysis. *World J. Surg. Oncol.* **23**, 77 (2025).
7. Dubovyk, V. *et al.* Obesity is a risk factor for poor response to treatment in early rheumatoid arthritis: a NORD-STAR study. *RMD Open* **10**, e004227 (2024).

8. Tong, M. Z. *et al.* Elevated BMI reduces the humoral response to SARS-CoV-2 infection. *Clin. Transl. Immunol.* **12**, e1476 (2023).
9. Cresci, G. A. & Bawden, E. Gut Microbiome: What We Do and Don't Know. *Nutr. Clin. Pract. Off. Publ. Am. Soc. Parenter. Enter. Nutr.* **30**, 734–746 (2015).
10. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
11. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
12. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4578–4585 (2011).
13. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4586–4591 (2011).
14. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
15. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer, New York, 2006).
16. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
17. Jašarević, E., Morrison, K. E. & Bale, T. L. Sex differences in the gut microbiome–brain axis across the lifespan. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150122 (2016).
18. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
19. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. *OptunaHub* https://hub.optuna.org/samplers/tpe_tutorial/.