BMI prediction using immune markers

Project Category: Life Sciences
Project Mentor: Honglin Yuan

Solveig Einarsdottir
SUNet ID: einarsd
Department of Computer Science
Stanford University
name@stanford.edu

Laura Bravo Merodio
SUNet ID: laubra77
Department of Computer Science
Stanford University
laubra77@stanford.edu

Rouven Spiess
SUNet ID: rouvens
Department of Computer Science
Stanford University
rouvens@stanford.edu

1 Introduction

Our main goal is to further advance the understanding of immune-metabolic relationships by modeling the association between Body Mass Index (BMI), adipokines and other immune markers. By doing so, we aim to generate novel biological insight to help target immune- metabolic dysfunctions such as obesity, diabetes or cancer. Our input is immunological and clinical data (i.e cell counts, age, blood marker concentrations..) of 902 patients and our output will be the predicted BMI.

2 Related Work

Adipose or fat tissue is considered a very active endocrine organ able to bridge metabolism and the immune system together through the secretion of hormones (i.e adipokines) and other factors. As such, body fat has been linked to systemic inflammation, insulin resistance, coagulation and other diseases [1]. BMI, defined as body weight divided by the square of body height (kg/m^2), is actively used in clinic as a proxy of body fat given its easier measurement. Previous research has attempted to predict BMI through blood markers, with work mainly looking at the relationship between categories of BMI (World Health Organization (WHO) establishes: underweight ($< 18.5kg/m^2$), normal weight ($18.5 - 24.9kg/m^2$), overweight ($25 - 29.9kg/m^2$), obese ($\geq 30kg/m^2$))[2]. In this particular paper [2], only univariate linear associations with BMI categories were acknowledged. In 2015, a study [3] aiming to link blood markers to overweight status utilized Extreme Learning Machine, an artificial neural network that contains a single hidden layer, to predict BMI. While they looked into feature selection like we did their data did not include immune markers like ours and so generating little biological insight. More recently, given the associations between obesity and immune dysregulation, BMI was studied in COVID patients [4]. In all, associations with immune markers was apparent, although not significant and no clear mechanism of action was described.

In this work, we are collaborating with Dr David Furman, Director of the Stanford 1000 Immunomes Project (Stanford 1KIP) and Dr Jordan Baechle, postdoctoral fellow in his team. Dr Furman has shared his project´s data, which consists of participants with no severe health conditions that were recruited at Stanford University between the years 2007 and 2016 for various studies of aging and vaccination (Sayed et al. 2021[5]). Our dataset is rich in immunological information, with features such as: cell frequencies, serum protein concentrations (cytokines, chemokines, adipokines and growth factors) and studies of responses to cytokine stimulation.

Such an extensive resource of information has been exploited before, with previous studies mainly focusing on immune state description (Furman et al. 2017[6]; Brodin et al. 2015[7]) and its changes

after external challenges such as vaccination (Furman et al. 2015[8]). In an attempt to start bridging links with metabolism, this is the first time BMI will be explored in this dataset.

3 Dataset and Features

The dataset consists of deep immune phenotyping (approx. 170 immune markers) of 902 participants (603 female and 299 male, ages 8-90 years). In the analysis we have split the data into a training and test set in a 75:25 split. The dataset had been previously curated and so had no missing values or outliers.

Given the high dimensionality of our dataset, we performed PCA analysis, assessing the variance of our data, possible patient clusters and performing dimension reduction.

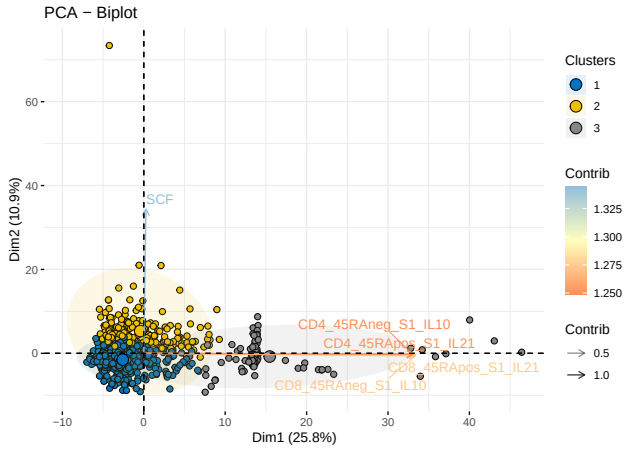

Figure 1: PCA (dimensions 1 and 2) with highest contributing features (>1) plotted and participants coloured according to clustering through hierarchical ascending classification (FactoMineR package)

4 Methods

4.1 Model building

4.1.1 Model choice

Running a preliminary comparison of regression algorithms with Pycaret [9] on our dataset gave us the ranking seen in Figure 2 where the algorithms are sorted by their root means square error, RMSE and R squared. This automated, data-driven results with default hyperparameter tuning using 10 fold cross validation directed our model selection towards decision trees given their higher performance.

4.1.2 Feature importance

Having narrowed down three of the decision tree algorithms (light gradient boosting machine, decision tree (rpart package) and decision tree (partykit package)), the whole dataset and PCA dimensions only, were both fitted, using BMI as outcome variable. Performance was assessed using R-squared and features evaluated through both model dependent variable importance algorithms and a moDel Agnostic Language for Exploration and eXplanation (DALEX) [10]

4.2 Extra Modeling

4.2.1 Feature Engineering with Linear Regression

Generating new features with feature interaction for a dataset of n existing features creates $\approx n^2$ new features. It is therefore important to restrain the feature space before the feature generation to keep runtime down and we naturally want to restrain it down to the features that contribute the most in

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	2.5381	12.0136	3.4404	0.5327	0.1219	0.0988	0.9660
rf	Random Forest Regressor	2.6654	12.9129	3.5640	0.4996	0.1268	0.1040	1.7710
gbr	Gradient Boosting Regressor	2.6936	13.0395	3.5893	0.4948	0.1278	0.1050	0.7380
lightgbm	Light Gradient Boosting Machine	2.7022	13.2069	3.6111	0.4908	0.1290	0.1055	0.2970
ada	AdaBoost Regressor	2.8345	14.1328	3.7372	0.4515	0.1345	0.1121	0.3080
omp	Orthogonal Matching Pursuit	2.9838	15.4416	3.9168	0.3962	0.1442	0.1186	0.0150
br	Bayesian Ridge	2.9968	16.2568	4.0189	0.3717	0.1467	0.1187	0.0220
lasso	Lasso Regression	3.0691	16.5692	4.0561	0.3642	0.1472	0.1217	0.0130
en	Elastic Net	3.1176	17.3719	4.1496	0.3373	0.1497	0.1233	0.0280
knn	K Neighbors Regressor	3.5470	22.4819	4.7060	0.1504	0.1679	0.1381	0.0420
dt	Decision Tree Regressor	3.7757	25.2361	4.9634	0.0418	0.1742	0.1449	0.0590
llar	Lasso Least Angle Regression	3.8193	26.8276	5.1479	-0.0138	0.1839	0.1497	0.4130
dummy	Dummy Regressor	3.8193	26.8276	5.1479	-0.0138	0.1839	0.1497	0.0180
huber	Huber Regressor	3.5612	45.4340	6.0123	-0.9831	0.1885	0.1413	0.1030
par	Passive Aggressive Regressor	4.5302	39.3229	6.1880	-0.5935	0.2315	0.1798	0.0370
ridge	Ridge Regression	3.6929	69.5105	6.8399	-2.0935	0.1933	0.1480	0.0200
lr	Linear Regression	4.4053	197.8321	10.7364	-7.8074	0.2229	0.1812	0.0230

Figure 2: PyCaret model ranking

predicting our target variable. We tried restricting the feature space to a few different numbers of features and added feature interactions, polynomial features of degree 3, or both.

4.2.2 BMI Prediction as a Classification Problem

Attempting to better understand the relationship between BMI and our features, and acknowledging the great variability in participants (children, old, men,women) we predicted BMI categories.Performance was assessed through the metric AUROC.

Work was performed both in python version 3.7 and R version 4.0 with all libraries mentioned above.

5 Experiments and Results

5.1 Model Building

Two datasets, the whole 177 features (AllData) and the 10 PCA dimensions (PCAData), were fitted using two boosted tree algorithms (rpart and partykit) and lightgradient boosting. The algorithms were then tuned using grid search (25 random parameters) in 10 fold cross validation. The parameters tuned were: cost complexity,tree depth and min n for the decision trees and mtry, tree depth, min n and learn rate for the boosted. Average performance results using best hyperparamters (as measured by R2) can be seen in 3

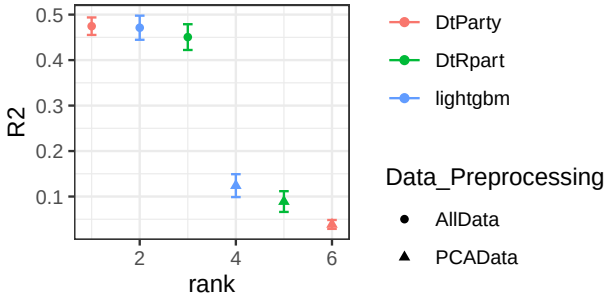

Figure 3: Model comparison using workflowsets

5.2 Feature Importance

Attending to the difficulty in interpreting the behaviour of more complex, non-linear models, we proceeded to assess the variable importance of each of the models. Two methodologies were used, method dependent variable importance (i.e betas for linear regression) and model agnostic methodologies such as DALEX.In the interest of space the mapping of the PCA dimensions to features is not shown but the most highly contributing features in each dimension were extracted too.

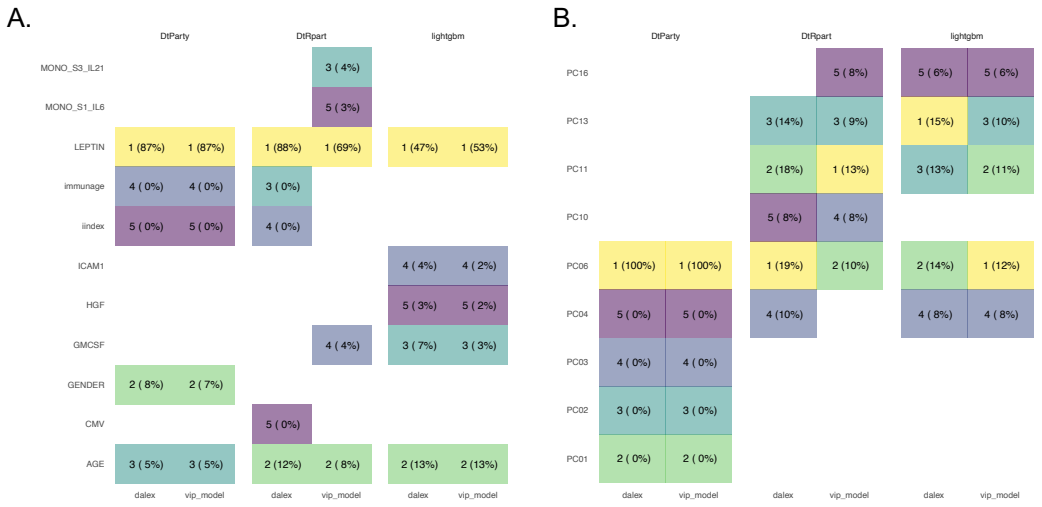

Figure 4: Feature importance heatmap.Most important features ranked according to two variable importance methodologies (vip) and DALEX for each of the three models trained. A) Whole dataset and B) PCA dimensions

5.3 Extra modeling

5.3.1 Feature Engineering with Linear Regression

The tables in Figure 5 shows the RMSE and R2 metrics for linear regression after feature selection and new feature generation.

Feature interaction

#features	RMSE	R2
20	3.9798	0.3768
9	3.8871	0.4143
4	3.7834	0.4371

Polynomial features

#features	RMSE	R2
20	3.7346	0.4529
9	3.8063	0.4309
4	3.8023	0.4348

Both

#features	RMSE	R2
20	3.9433	0.3784
9	3.8157	0.4292
4	3.8375	0.4246

Figure 5: Performance of Linear Regression with new generated features

The best performance we are able to get is an R2 of 0.4529, with decision trees performing better (0.489)

5.3.2 Classification

Achieved our best score with a light Gradient Boosting Machine, with accuracy of 84% using 3 classes (class 1: <18.5, class 2: 18.5-30, class 3: >30), Figure 6.

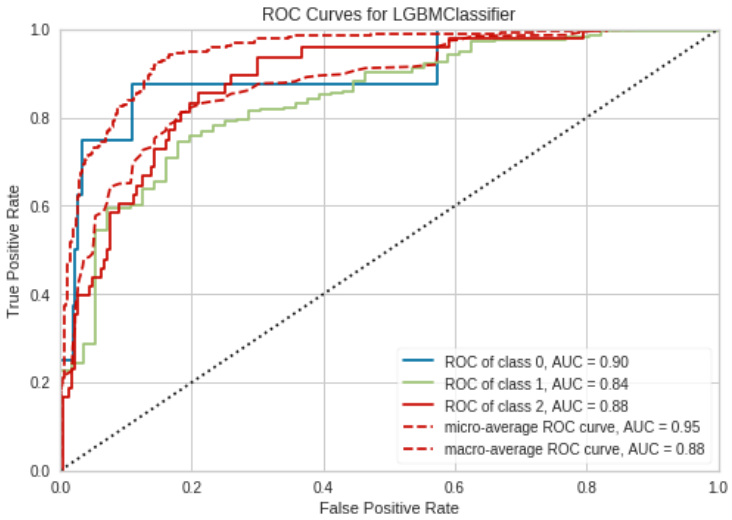

Figure 6: Best classification model performance

6 Conclusions and Future Work

In this work we have been able to build a model and generate biological insight in the process. The adipokine leptin as key driver of the model stems from it being a driving force in the association between immune parameters and adipose tissue (Naylor y Petri 2016[11]), with the other key features GMCSF and ICAM1 warranting further investigation. Gender and age related variables such as iindex, immuneage also appear as important as envisioned.

For the future, further analysis of possible interactions between features would be of interest. Also, given the good performance of our classification model and the widespread age range and variability in our data (from 8 to 90 years old), exploration of BMI prediction in different subgroups would be of interest. Also, genetic and disease information is available for this participants too, a study that would definitely enrich our predictions and research given the high heritability component in BMI as well as the high multimorbidity association.

7 Contributions

We all contributed equally to the project, with sharing of ideas and implementation. More specifically, Solveig ran feature generation on a reduced feature space with linear regression, Rouven ran pycaret models and expanded hyerparameter tuning information and Laura, performed model building and feature importance.

References

[1] Vera Francisco, Jesús Pino, Victor Campos-Cabaleiro, Clara Ruiz-Fernández, Antonio Mera, Miguel A Gonzalez-Gay, Rodolfo Gómez, and Oreste Gualillo. Obesity, Fat Mass and Immune System: Role for Leptin. *Frontiers in physiology*, 9:640–640, June 2018. Publisher: Frontiers Media S.A.

[2] Silvia Ilavská, Mira Horváthová, Michaela Szabová, Tomáš Nemessányi, Eva Jahnová, Jana Tulinská, Aurélia Líšková, Ladislava Wsolová, Marta Staruchová, and Katarína Volkovová. Association between the human immune response and body mass index. *Human Immunology*, 73(5):480–485, 2012.

[3] Liu D Liu W Liu Y Zhang X et al. Chen H, Yang B. Using Blood Indexes to Predict Overweight Statuses: An Extreme Learning Machine-Based Approach. *PLoS ONE 10(11): e0143003.*, 2015.

[4] Emma J. Kooistra, Aline H. de Nooijer, Wout J. Claassen, Inge Grondman, Nico A. F. Janssen, Mihai G. Netea, Frank L. van de Veerdonk, Johannes G. van der Hoeven, Matthijs Kox, Peter Pickkers, Pleun Hemelaar, Remi Beunders, Tim Frenzel, Jeroen Schouten, Sjef van der Velde, Hetty van der Eng, Noortje Roovers, Margreet Klop-Riehl, Jelle Gerretsen, Nicole Waalders, Hidde Heesakkers, Tirsa van Schaik, Leonie Buijsse, Leo Joosten, Quirijn de Mast, Martin Jaeger, Ilse Kouijzer, Helga Dijkstra, Heidi Lemmers, Reinout van Crevel, Josephine van de Maat, Gerine Nijman, Simone Moorlag, Esther Taks, Priya Debisarun, Heiman Wertheim, Joost Hopman, Janette Rahamat-Langendoen, Chantal Bleeker-Rovers, Hans Koenen, Esther Fasse, Esther van Rijssen, Manon Kolkman, Bram van Cranenbroek, Ruben Smeets, Irma Joosten, and on behalf of the RCI-COVID-19 study group. A higher BMI is not associated with a different immune response and disease course in critically ill COVID-19 patients. *International Journal of Obesity*, 45(3):687–694, March 2021.

[5] Yingxiang Huang Khiem Nguyen Zuzana Krejciova-Rajaniemi Anissa P. Grawe Tianxiang Gao Robert Tibshirani et al. Sayed, Nazish. An Inflammatory Aging Clock (IAge) Based on Deep Learning Tracks Multimorbidity, Immunosenescence, Frailty and Cardiovascular Aging. *Nature Aging 1 (7): 598–615.*, 2021.

[6] Junlei Chang Lydia Lartigue Christopher R. Bolen François Haddad Brice Gaudilliere Edward A. Ganio et al. Furman, David. Expression of Specific Inflammasome Gene Modules Stratifies Older Individuals into Two Extreme Clinical and Immunological States. *Nature Medicine 23 (2): 174–84. https://doi.org/10.1038/nm.4267.*, 2017.

[7] Vladimir Jojic Tianxiang Gao Sanchita Bhattacharya Cesar J. Lopez Angel David Furman Shai Shen-Orr et al. Brodin, Petter. Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell 160 (1): 37–47.*, 2015.

[8] Vladimir Jojic Shalini Sharma Shai S. Shen-Orr Cesar J. L. Angel Suna Onengut-Gumuscu Brian A. Kidd et al. Furman, David. Cytomegalovirus infection enhances the immune response to influenza. *Science Translational Medicine 7 (281): 281ra43-281ra43.*, 2015.

[9] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. PyCaret version 1.0.

[10] Przemyslaw Biecek. Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, 19(84):1–5, 2018.

[11] William A. Petri. Naylor, Caitlin. Leptin Regulation of Immune Responses. *Trends in Molecular Medicine 22 (2): 88–98.*, 2016.