SDS 322E Project: Aging Biomarkers

Team: James Groh, Erin Pennington, Bella Vincent, Jilliane Lagus, Janice Oh, Anthony Tang

Introduction

Our project goals were: (1) explore the relationships between biomarkers and age and (2) utilize biomarker data to create a model that predicts the biological age of an individual. This project was inspired by Hastings et al.'s 2019 study "Comparability of biological aging measures in the National Health and Nutrition Examination Study, 1999-2002", in which the authors used biomarkers to model biological age.

Chronological age is defined as the number of years that have passed since the individual's birth while biological age relates to an individual's health. An individual's biological age can be derived by the collection and analysis of biomarkers, e.g. systolic blood pressure. In short, biological age tells us how old the individual appears on a biochemical level. An individual who leads a healthy lifestyle may be biologically younger than his chronological age while an individual who leads an unhealthy lifestyle may be biologically older than his chronological age. We believed our model could serve as a useful tool. For example, it might show an individual that he is biologically older than his chronological age and may need to change his lifestyle.

In Part 1 of the project we determined how biomarker levels changed with age, what biomarkers should be incorporated into our model, and what type of model we should design. In Part 2 we designed and tested several models and experimented with subsampling to create the most accurate model possible.

Data

The National Health and Nutrition Examination Survey (NHANES) contains multiple studies assessing the health and nutrition of adults and children in the United States. It offers different data files containing the results of interviews, physical examinations, and lab tests. We downloaded and merged several data files from the NHANES 2017-2018 dataset for our exploratory analysis. Our choice of variables was based on the biomarkers used in the biological age clocks referenced by Hastings et al. The resultant dataset included the following variables: serum albumin, alkaline phosphatase, blood urea nitrogen, creatinine, C-reactive protein, glycated hemoglobin, total cholesterol, uric acid, white blood cell count, lymphocyte percent, mean corpuscular volume, systolic blood pressure, glucose, red blood cell distribution, age and gender.

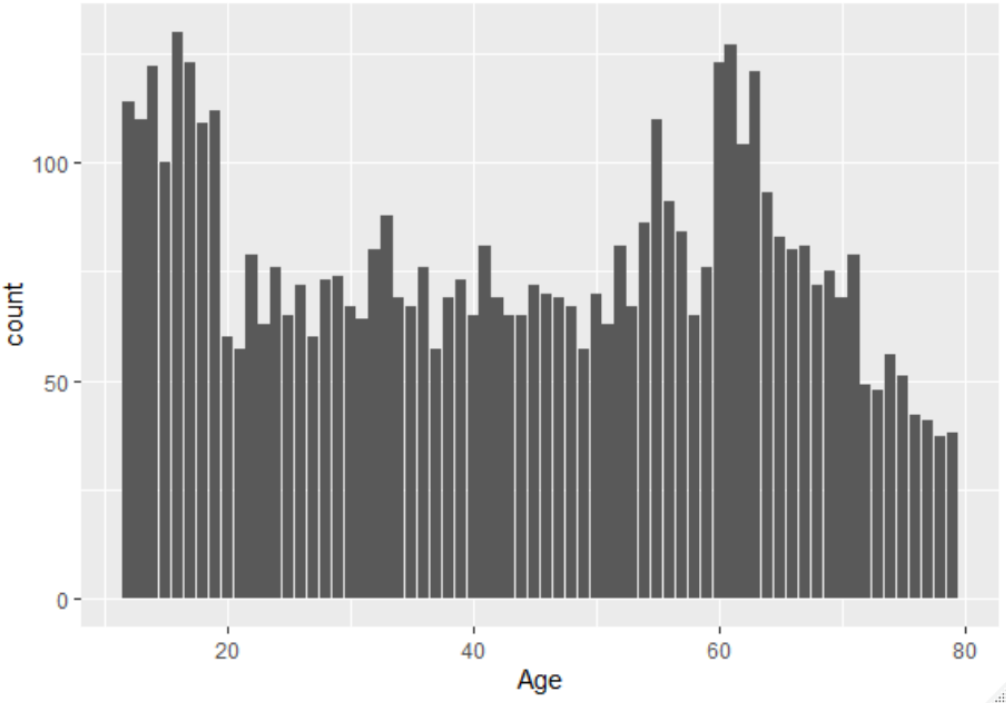
We filtered our data by removing any individuals lacking an entry for any of the given variables. Individuals in the dataset ranged from 12 to 80 years old. However, all individuals 80 or older were classified as 80. We removed these individuals as their inclusion would introduce individuals whose true age was not known and interfere with our ability to visualize how biomarkers change with age. Cleaning reduced our dataset from 6401 individuals to 5251. We created a normalized version of our dataset for use in our models in order to avoid unequal preference for values with larger scales.

Exploratory Analysis

Age Distribution

In the first part of our exploratory analysis we visualized the distribution of ages represented by our dataset, as shown in **Figure 1** below. Ages in the dataset were clearly not normally distributed. Nevertheless, we decided to continue our exploration of the data without removing additional individuals. While the dataset was not perfectly suited for analysis it did possess two positive characteristics: it contained individuals of various ages and did not contain an overwhelming number of individuals of a single age.

Figure 1: Histogram representing age distribution



Biomarker Relationship with Age

In the second part of our exploratory analysis, we created scatter plots for each biomarker and age, fit a regression line to the plot, and computed the adjusted R^2. A few of the resulting plots are shown in **Figures 2-5** below and their corresponding adjusted R^2 values are shown in **Table 1** below.

Figure 2. Systolic Blood Pressure and Age

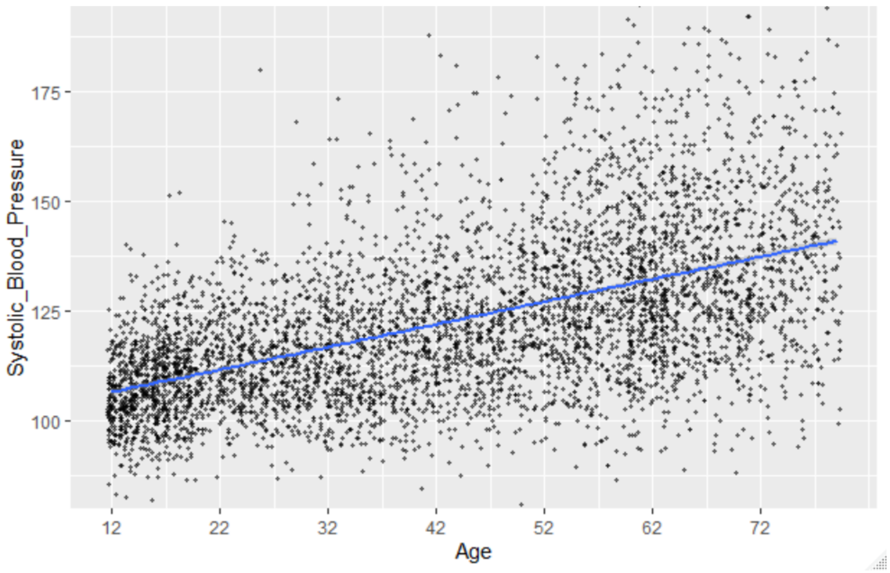


Figure 3. Glycohemoglobin and Age

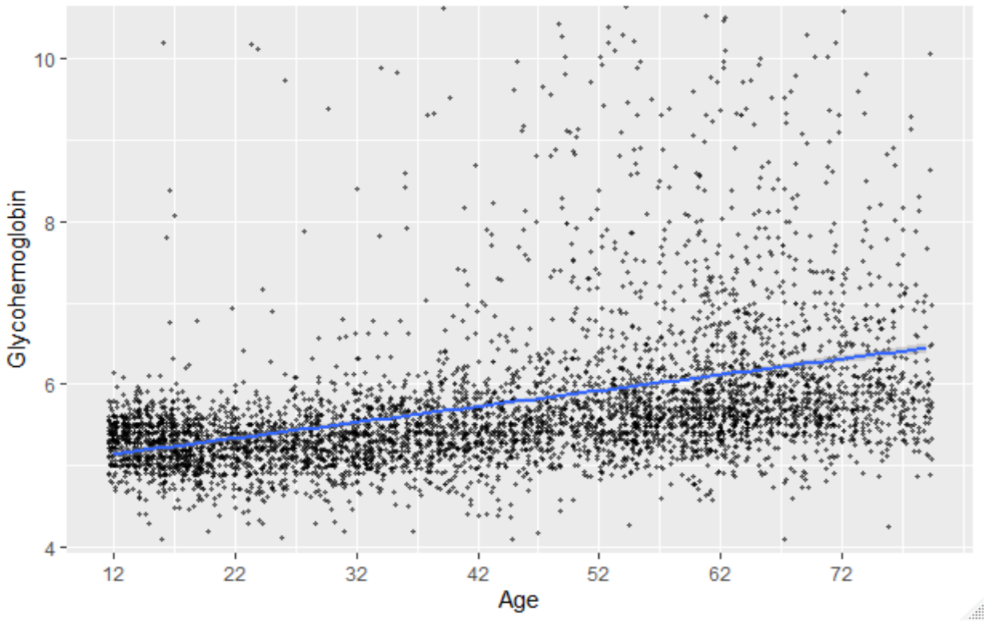


Figure 4. Blood Urea Nitrogen and Age

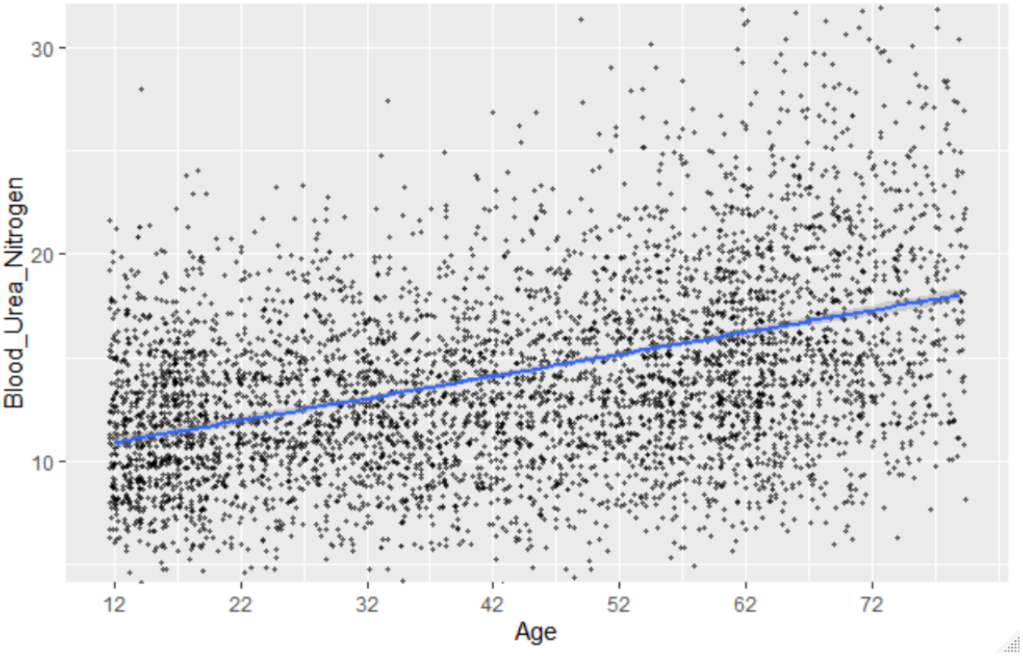


Figure 5. Alkaline Phosphatase and Age

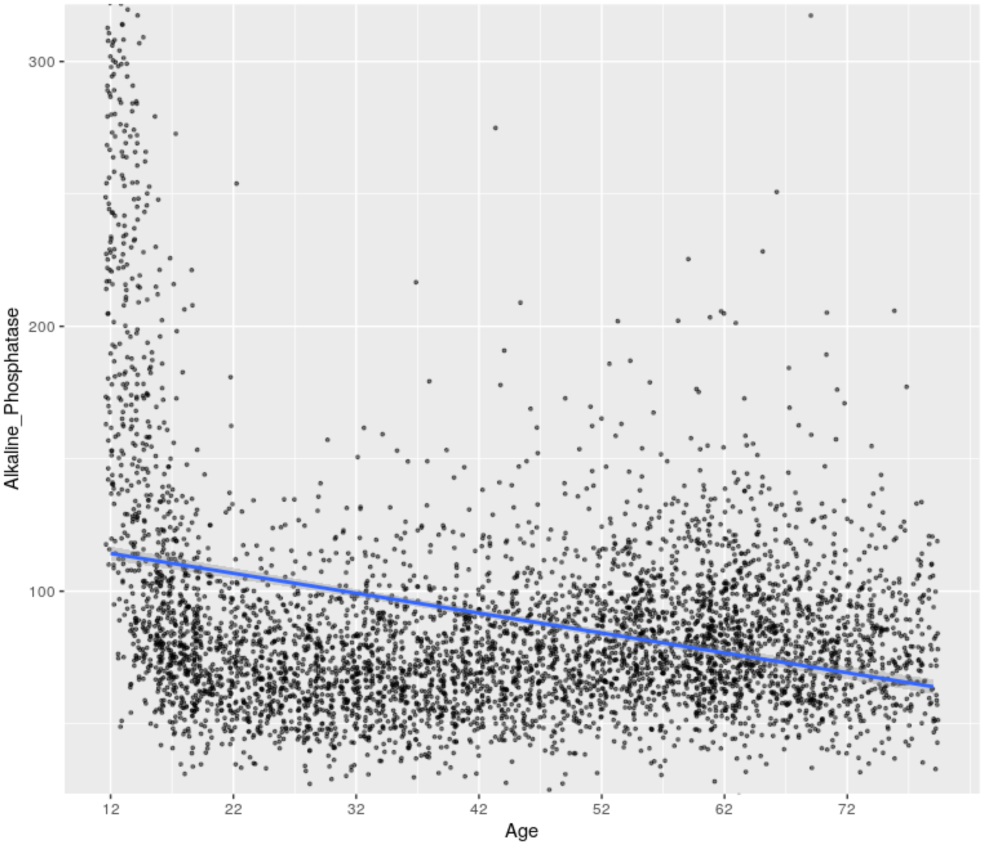


Table 1: Adjusted R^2 for selected biomarkers

Systolic blood pressure	0.2841
Blood urea nitrogen	0.1381
Glycohemoglobinn	0.1332
Alkaline phosphatase	0.07818

As seen in **Figure 2** systolic blood pressure appears to increase with age. Systolic blood pressure is a measure of the force the heart exerts on the walls of the arteries each time it beats. An age-related increase in blood pressure is known to be a universal feature of human aging. Among Westerners over age 40 years, systolic blood pressure increases by roughly 7 mmHg per decade [1]. Epidemiological surveys show a progressive increase in systolic blood pressure with age, reaching an average of roughly 140 mmHg by the eighth decade [1]. Higher blood pressure is associated with cardiovascular and renal disease across diverse populations, even controlling for other factors [1].

We observed an increase in glycohemoglobin with age as shown in **Figure 3.** Higher levels of glycohemoglobin indicate higher levels of sugar in the blood. Glycohemoglobin is a primary predictor of diabetes mellitus and is an important biological marker of health and body function [2]. We hypothesized that as adults age, their bodies become less efficient at cellular import of sugar due to reduced insulin secretion, resulting in a higher level of sugar in the blood and increased glycohemoglobin. At least part of this hypothesis is supported by Muller et al., which states, "Insulin secretion, on the other hand, seems to decrease with age even after adjustments for differences in adiposity, fat distribution, and physical activity [3]."

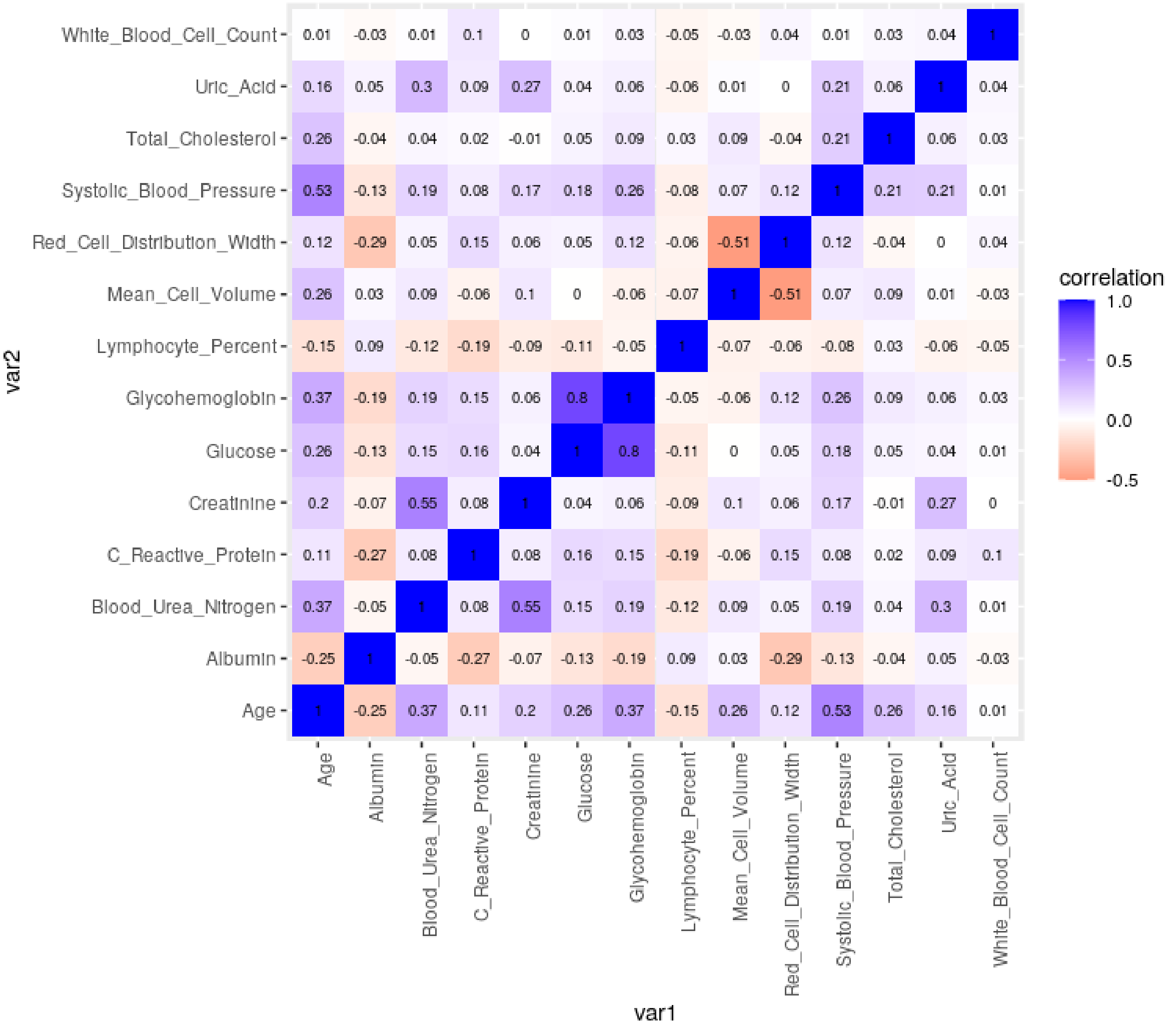
We observed an increase in blood urea nitrogen with age as shown in **Figure 4.** A blood urea nitrogen test measures the amount of nitrogen in the blood that comes from the waste product of metabolism, urea. Blood urea nitrogen levels are dependent on the rate of glomerular filtration and tubular reabsorption occurring in the body [4]. Therefore, blood nitrogen urea levels play a vital role in diagnosing and evaluating renal function and higher levels of blood urea nitrogen indicate deteriorating kidney function. We hypothesized that as adults age, their kidneys function less efficiently and therefore their blood urea nitrogen levels rise. This hypothesis is sound as reduced kidney function with age is well documented and the glomerular filtration rate (GFR) declines at a rate of approximately 1.0 mL/min per 1.73 m^2 per year in elderly subjects [5].

As you can see in **Figure 5**, alkaline phosphatase changed differently with age than the previously mentioned biomarkers**.** Levels dropped steeply from age 12 to approximately 22, when they leveled off. We decided to remove alkaline phosphatase from further exploration as it showed drastically different levels before and after 22. Alkaline phosphatase is an enzyme that breaks down proteins in the body and abnormal levels in the blood can indicate malnutrition, kidney tumors, serious infection, or problems with the intestines or pancreas [6].

The second part of our data exploration yielded several valuable observations about our dataset. Although the data points of the biomarkers shown in **Figures 2-4** at least partially group around their respective regression lines, systolic blood pressure is clearly the best fit of the three. Glycohemoglobin has a significant number of outliers that skew the regression line away from the bulk of the data points. There are several biomarkers that suffer from this same problem. Blood Urea Nitrogen suffers from a different problem, namely, the data points are far less tightly packed around the regression line than those of systolic blood pressure. Both these problems result in lower adjusted R^2 values, as you can see in **Table 1**. We opted to continue to use the data as it was, reasoning that the inclusion of multiple biomarkers in our model would lessen the impact of any individual biomarker. We may remove outliers in future models but for now, we merely noted their inclusion as a limitation of our model.

In addition to the scatter plots, we created a correlation matrix for the variables as shown in **Figure 6** below. The variables most correlated with age were systolic blood pressure, glycohemoglobin and blood urea nitrogen. White blood cell count showed very little correlation with age. We included it in our model but may remove it later on in order to simplify the model.

Figure 6: Correlation Matrix



Modeling

Clustering

Having thoroughly explored the data we began to design our model. In the first part of our model design, we determined if a model based on clustering or regression would be more effective. The former possibility was tested by observing if our biomarkers clustered based on age. Since we planned to use numerous biomarkers in our model we could not simply use a 2-D or 3-D scatter plot but needed to employ a dimensional reduction technique. We opted to use Principal Component Analysis (PCA). The resultant Scree Plot, seen in **Figure 7** below, showed that 32.1% of the variance in the data was explained by principal component 1 (PC1) and principal component 2 (PC2). The data was plotted in PC1 vs. PC2 space, seen in **Figure 8**, but did not show obvious clustering based on age. In order to be certain whether or not clusters would form, we used k-means clustering to generate a silhouette plot, seen in **Figure 9**, which called for the use of two clusters. These clusters shown in **Figure 10** appear arbitrary. This is due to the fact that the algorithm that generates the silhouette plot produces a minimum of two clusters. In short, our dataset does not cluster on age so we decided to design a regression-based model.

Figure 7: Scree Plot

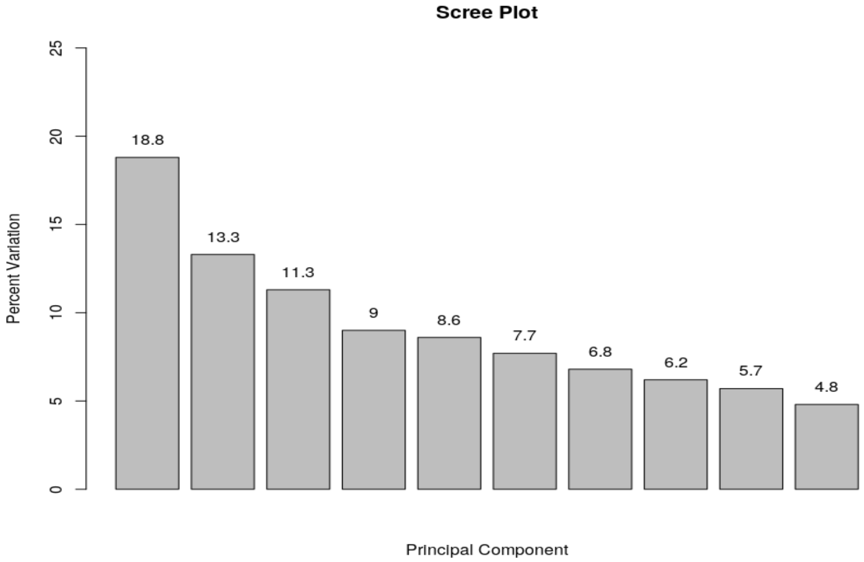


Figure 8: PC1 vs PC2

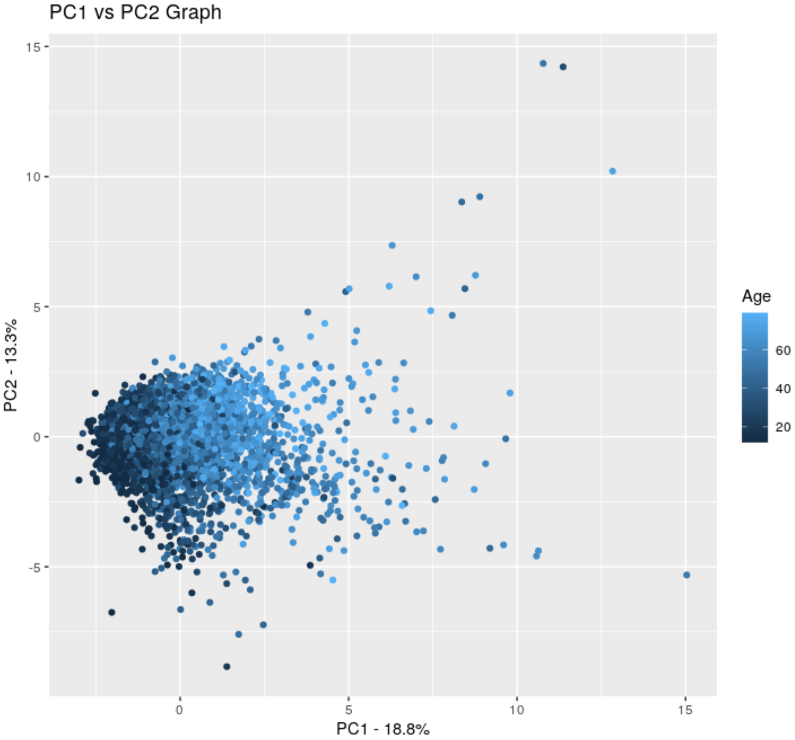


Figure 9: Silhouette plot

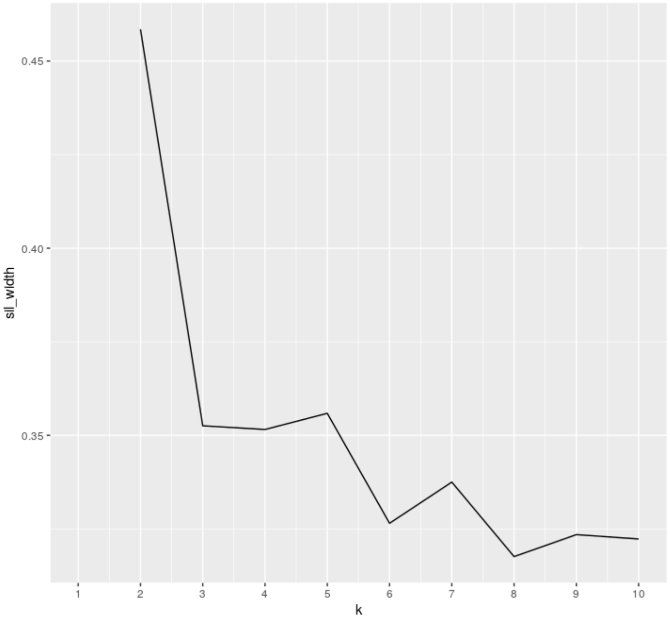
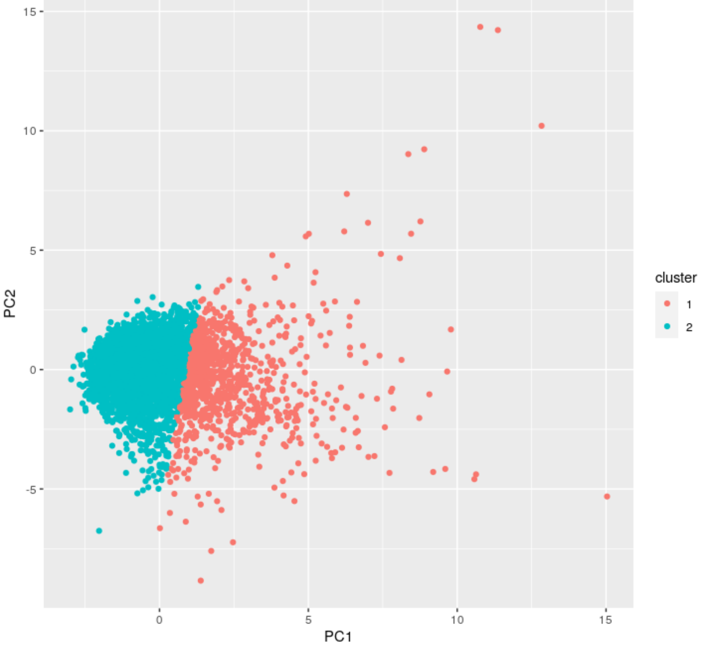


Figure 10: PC1 vs. PC2 with k-means clusters



Regression

We performed a regression analysis to create a model for predicting biological age by using chronological age as the target in our training data. We then used the LinearRegression model in Python available from Scikit-learn to fit a linear regression to our data, using the biomarkers in our dataset and targeting chronological age. This produced a model that generated a line through multi-dimensional space, where the intercept and slope for each of the input variables have been optimized for the data provided.

In order to visualize our results and judge the correctness of the regression line, we created two diagnosis plots **Figure 11 and 12** using Matplotlib to compare the actual and predicted age of each sample based on our linear regression model. The former used all biomarkers—except alkaline phosphatase which we had previously removed—while the latter included only Blood Urea Nitrogen, Glucose, Glycohemoglobin and Systolic Blood Pressure due to their higher correlation with age. Next, we developed a model based on the RandomForestRegressor and generated the diagnosis plot shown in Figure 13. No attempt at subsetting was made with the Random Forest Model as subsetting had not improved the accuracy of the Linear Regression Model.

Figure 11: All Biomarkers Age Diagnosis Plot

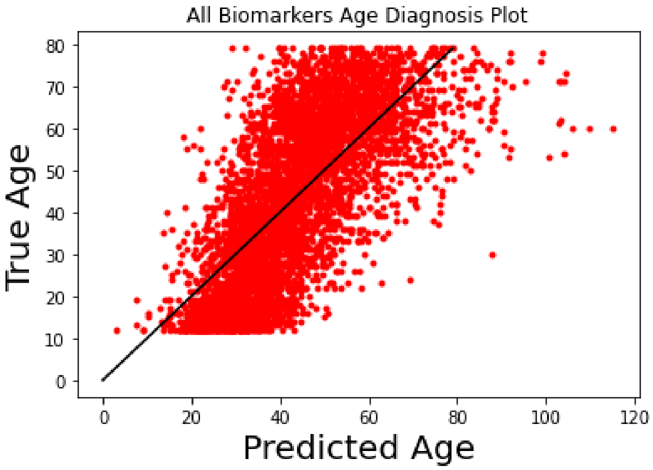


Figure 12: Subset of Biomarkers Age Diagnosis Plot

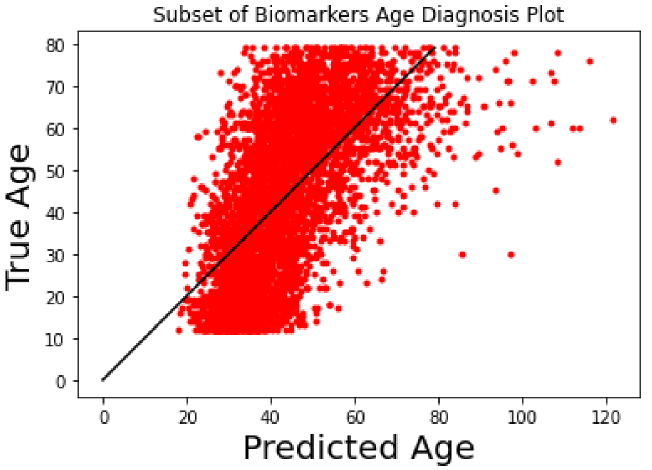


Figure 13: Random Forest Plot

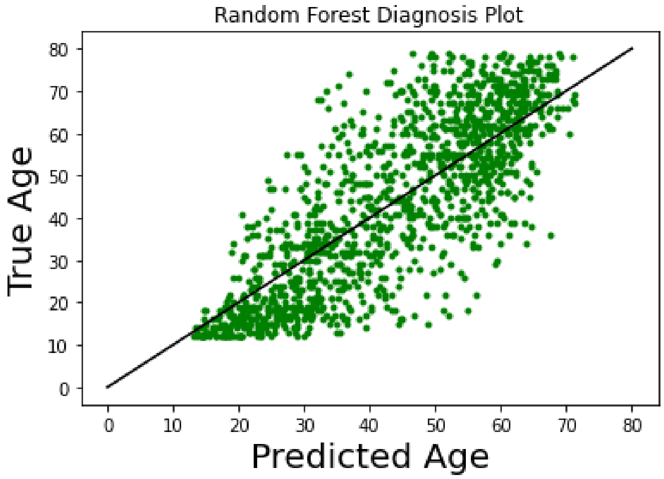


Table 2: Model Comparison

Linear Regression (All Biomarkers)	**Linear Regression (Select Biomarkers)**	**Random Forest Regression**
MAPE Accuracy: ~64%	MAPE Accuracy: ~59%	MAPE Accuracy: ~73%
MSE: ~192	MSE: ~231	MSE: ~136
Absolute Error: 13.9 years	Absolute Error: 15.2 years	Absolute Error: 11.7 years

Discussion

Model Performance

Visual inspection of **Figures 11-13** suggests the Random Forest Regression Model was the best of the three regression models as its data points packed more tightly around the regression line than either Linear Regression Model. It was our hope that the model could be simplified to include fewer biomarkers but the data points in **Figure 12** appear less tightly packed around the regression line than those in **Figure 11**, suggesting that subsetting the biomarkers did not improve model performance. The summary statistics in **Table 2** confirmed our visual inspection. Subsetting the biomarkers in the Linear Regression Model decreased MAPE accuracy by 5%, increased MSE by 39, and increased absolute error by 1.3 years. However, switching from the Linear Regression (All Biomarkers) to the Random Forest

Regression Model reduced improved MAPE accuracy by 9%, decreased MSE by 56 and decreased absolute error by 2.2 years.

In summary, our Random Forest Regression Model was able to use the biomarkers we selected to predict an individual's biological age to within 11.7 years. This large absolute error, together with the model's limitations, leads us to believe that the model could be improved upon. Nevertheless, the model's predictive capability is likely superior to that of an average human. It is easy to imagine an individual looking in the mirror and thinking himself in good health when he actually is not. The objective analysis this model provides makes it a valuable diagnostic tool.

Limitations

Our model suffers from several limitations but the most important one relates to the difference between chronological age and biological age. We trained our model to use biomarkers to compute an individual's biological age. The flaw in our methodology was we used the chronological ages of individuals to train our model to calculate an individual's biological age. Ideally, we would have used confirmed biological ages to train the model. We assumed that the majority of individuals in our dataset have a biological age close to their chronological age and that variations between biological and chronological age among individuals—i.e. some individuals are biologically younger than their chronological age while others are biologically older than their chronological age—would be attenuated by the large size of our dataset. In other words, although we were technically training our model with chronological age, we assumed that we are effectively training it with biological age. There are several ways the dataset could invalidate our assumptions. For example, if all individuals in the dataset were in poor health and thus biologically older than their chronological age, then our assumptions would be invalid. We could improve future models by using more complex training algorithms, perhaps by investigating those used in other biological age predictors such as Klemera-Doubal and Levine.

Several previously mentioned limitations of the dataset affected the accuracy of the model. The data was truncated below age 12 and above age 80, leaving out a large number of individuals from those age groups and calling for additional modification of the models to have this factor incorporated. Also, a significant number of outliers disrupted the linear relationship between age and several biomarkers. Other biomarkers, when plotted against age, show data points that are not tightly grouped around their respective regression lines. Some biomarkers showed a different relationship with age from 12 to approximately ~22 than they showed for ~22 and older. This phenomenon was so extreme in Alkaline Phosphatase that we opted to exclude it from our models. However, other biomarkers, such as Glycohemoglobin, also exhibited this behavior to a lesser extent. We opted not to remove individuals younger than 22 as they totaled 1,116 individuals and their removal would significantly reduce the size of our dataset. Deciding to include these individuals may have caused our linear regression model to perform poorly. This difference in the rate at which certain biomarkers change could be due to the fact that individuals are not fully physically developed until somewhere between 18 and 25. Likewise, perhaps older individuals, as some of their organs and bodily processes begin to slowly degrade due to age, may have very different levels of certain biomarkers compared to the average adult as well.

Furthermore, we also filtered out any individuals with the age of 80, since the data is top-coded to 80 (as in individuals over 80 years old are marked as 80). This means that our data did not include anyone 80 years old or older. However, in the linear regression model, some of the predicted ages were over 80 years old. Those predictions may not be reliable, as the model was not able to look at true data for individuals over 80 years old to support these predictions. However, the Random Forest Regression model did not make any predictions above the age of 80 years old.

Conclusion

Main Findings

In conclusion, using the NHANES 2017-2018 dataset we developed a Random Forest Regression model that can determine an individual's biological age to within approximately 11.7 years. The purpose of the model was to use diagnostic laboratory tests to assess if an individual's biological age was significantly older than his chronological age, indicating the individual may be at risk for age-related diseases and in need of medical intervention, a change in lifestyle, or both. This model was subject to several limitations which may be overcome through future work.

Next Steps

Several improvements could be made to dataset preparation and our model. We could combat the issue of data truncation by using synthetic data. We could use a modified dataset to train our model, one in which the biological age of individuals in the dataset is more certain. We could divide our dataset based on gender, as was done in Hastings et al. However, were we to divide the dataset in this way it would be necessary to check for confounding variables in order to determine if gender was actually causing any observed difference. We could test other models to see if they improved accuracy. For example, we could try other regression ensemble methods available in Python like KNeighborsRegressor and DecisionTreeRegressor. Alternatively, we could use an entirely different approach like Neural Networks to tackle the problem.

Acknowledgments

- **James Groh**: Clustering, PCA, Hidden Forest, heavy coding and editing other teammates code, organizing and commenting code, writing and editing report, slidedeck, Q&A (asks salient questions to course instructor to move the project forward)
- **Erin Pennington**: Modeling, heavy coding, documentation, reporting, slidedeck, selected data and reference publication, point-person/resource on coding and programming and Github for computing, (brought years' experience computing to team work)

- **Anthony Tang**: Exploratory - correlation matrix part, heavy coding, documentation, reporting, slidedeck, Q&A (asks salient questions to course instructor to move project forward), identified point-person/resource on medical terminology/applications
- **Bella Vincent**: Exploratory - scatter plot part, research on biomarker background information, formatting/transferring between report/slides, heavy coding, documentation, reporting, slide deck, identified resource/point person for documentation
- **Janice Oh**: Discussion, heavy coding, documentation, reporting, slidedeck, Q&A (Team) (asks really pertinent clarification questions in team to move project forward), identified point person/resource for results interpretation
- **Jilliane Lagus**: Scheduling/coordinating team meetings, copy-editing documentation inputs (Introduction, Figures/Tables, Formatting Sections, Conclusion) to publication-grade, identified point-person/resource on geriatrics/aging for industry/domain context, final layout/upload of slidedeck, minor coding, Conclusion

References

1. Wolf-Maier K, Cooper RS, Banegas JR, Giampaoli S, Hense H-W, Joffres M, Kastarinen M, Poulter N, Primatesta P, Rodriguez-Artalejo F, Stegmayr B, Thamm M, Tuomilehto J, Vanuzzo D, Vescio F. **Hypertension prevalence and blood pressure levels in 6 European Countries, Canada, and the United States**. JAMA. 2003; 289:2363–2369.
2. Krishnamurti, U., & Steffes, M. (2001). **Glycohemoglobin: A Primary Predictor of the Development or Reversal of Complications of Diabetes Mellitus**. Clinical Chemistry, 47(7), 1157-1165. DOI: 10.1093/clinchem/47.7.1157
3. Muller DC, Elahi D, Tobin JD, Andres R. **The effect of age on insulin resistance and secretion: a review. Semin Nephrol**. 1996 Jul;16(4):289-98. PMID: 8829267.
4. Y. Xue, L.B. Daniels, A.S. Maisel, Navaid Iqbal, **Cardiac Biomarkers**, Reference Module in Biomedical Sciences, Elsevier, 2014
5. Li, G., Chen, Y., Hu, H., Liu, L., Hu, X., Wang, J., Shi, W., & Yin, D. (2012). **Association between age-related decline of kidney function and plasma malondialdehyde.** Rejuvenation research, 15(3), 257–264. https://doi.org/10.1089/rej.2011.1259
6. **An Alkaline Phosphatase (ALP) Test: Levels and More.** (2021). Retrieved 17 November 2021, from https://www.healthline.com/health/alp