Predicting age group from health and nutritional status of Americans

Forgive Agbesi, Jason Lee and Michael Hewlett

Table of contents

1 Summary

In this analysis we explored the use of several classification models to predict whether a respondent is an adult or senior (essentially below or above age 65) based on their health and nutritional data. Our most promising model used Support Vector Classification (SVC). While it appeared promising, much of the model's accuracy was achieved by classifying most respondents as adults, since this was the majority class. Precision and recall for predicting the senior

class was quite low. This suggests that the model has considerable room for improvement, which could be achieved through optimizing the hyperparameters and selecting models based on precision, recall, or f1 scores, rather than general accuracy. With the goal of correctly classifying each group, false positive and false negative errors were both equally important for our analysis, and applying class weighting is worth exploring in future research. Once the model performs better on those metrics, it would be worth exploring which health and nutritional features are most predictive of age, which could provide suggestions for strategic public health programs.

2 Introduction

While taking care of elders is a core value of many cultures, this is not a hallmark of many western societies, including the Unites States (Healthy Aging Center (2022)). Is it possible that this is reflected in different health measures? Put another way, could we use health measures to predict whether an American is a senior or not?

Formally, the question this project seeks to answer is: Can information about the health and nutritional status of Americans be used to predict whether they are adults or seniors?

The dataset used to answer this question is the National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset ((**national_health_and_nutrition_health_survey_2013-2014_?**)(nhanes)_age_prediction_subset_887). It was originally prepared for a research paper on predicting diabetes and cardiovascular disease in patients (Dinh and Miertschin (2016) and Mukhtar and Al Azwari (2021)). The dataset's stated purpose was to assess the health and nutritional status of adults and children in the United States (Papazafiropoulou (2024)), however respondents were classified as either Adults (respondents under 65 years of age) or Seniors (respondents 65 years of age or older). Respondents were located in the United States and provided data through interviews, physical examinations, and laboratory tests to the National Center for Health Statistics (NCHS) (part of the Centers for Disease Control and Prevention (CDC)).

The dataset has 10 variables and 2278 rows, with each row representing a respondent. The variables are:

1. SEQN - The respondent ID aka sequence number

2. age_group - The respondent's age group (adult or senior)

3. RIDAGEYR - The respondent's age in years

4. RIAGENDR - The respondent's gender (1 represents Male, 2 represents Female)

5. PAQ605 - Whether the respondent engages in weekly moderate or vigorous physical activity (1 means they do, 2 means they don't)

6. BMXBMI - The respondent's body mass index

7. LBXGLU - The respondent's blood glucose after fasting

8. DIQ010 - Whether the respondent is diabetic (1 is yes, 2 is no)

9. LBXGLT - A measure of the respondent's oral health

10. LBXIN - The respondent's blood insulin levels

According to the dataset description, there are no missing values, though EDA found some unexpected values for physical activity and diabetic. Since no explanation was provided for these codes, we omitted these respondent's from our analysis.

3 Methods & Results

3.1 Description of methods

We loaded and cleaned the data, first renaming columns for clarity. We then found values for physical activity and diabetic variables that were not explained in the dataset's documentation and decided to remove observations with those values. Next we confirmed the that the dataset's description of no missing values was accurate, then split the data into training and test, and conducted EDA on the training set - including examining summary statistics of each variable and plotting their distributions.

For our analysis, we first transformed categorical variables with one hot encoding, and standardized the scales of numeric variables. Because there were no missing values, it was not necessary to do transformations for this. We then fit 3 models (a dummy classifier, a logistic regression, and SVC) to the training data, and selected SVC for our final analysis because it had the best mean cross-validation accuracy. Finally, we used our SVC model to predict the test data and visualized how the model performed on this data.

3.2 Inspecting errors

The dataset source stated that "gender", "physical_activity", and "diabetic" are binary features. However, "physical_activity", "diabetic" contained three unique values instead of two. According to the dataset's documentation, 'physical_activity' should only have 1 or 2 as values so rows containing 7 should be omitted. Similarly, 'diabetic' should only have 1 or 2 as values so rows containing 3 should be omitted.

As a result, we removed 59 observations from the dataset during validation (1 case where physical activity was "7" and the remaining cases where diabetic was set to "3").

3.3 Renaming columns and glancing at their values

We first renamed the columns of the data set to be more meaningful and easy to understand. Below is a short description of each column in the data set.

- RIDAGEYR: Respondent's Age
- RIAGENDR: Respondent's Gender (1 is Male / 2 is Female)
- PAQ605: Does the respondent engage in weekly moderate or vigorous-intensity physical activity (1 is yes / 2 is no)
- BMXBMI: Respondent's Body Mass Index
- LBXGLU: Respondent's Blood Glucose after fasting
- DIQ010: If the Respondent is diabetic (1 is yes / 2 is no)
- LBXGLT: Respondent's Oral
- LBXIN: Respondent's Blood Insulin Levels

3.4 Splitting the data set

Prior to conducting EDA, we split the data set to avoid looking at the test data and influence the training of our model. The training data was 80% of the original dataset, and the test data was 20%.

3.5 Conducting EDA on the training set

Table 1: Summary Statistics

	gender	physical_activity	bmi	blood_glucose	diabetic	oral	blood_insulin
count	1775	1775	1775	1775	1775	1775	1775
mean	1.51	1.82	27.78	99.11	1.99	113.89	11.66
std	0.5	0.38	7.15	17.08	0.1	45.79	9.54
min	1	1	14.5	63	1	40	1.02
25%	1	2	22.7	91	2	87	5.8
50%	2	2	26.7	97	2	104	8.89
75%	2	2	31.1	103	2	129	14.26
max	2	2	70.1	405	2	604	102.29

The training data has 1775 observations. Since gender, physical_activity, and diabetic features were categorical, only the mean and standard deviation from the table above were relevant for those columns. Body mass index values below 18 are considered underweight, and values over 40 are considered severely obese. We observed that the middle 50% of values fall between 22.7 & 31.1, though the max was 70.1, which is concerningly high. Blood glucose, oral, and blood

insulin have their own ranges, so it was necessary to standardize these variables before fitting our model.

3.6 Visualization for EDA

The distributions in Figure 1 below show class imbalance, with very few seniors relative to adults in our dataset. Across numeric variables, mode values for seniors were less pronounced than they were for adults, though ranges seemed similar. Seniors seemed to have higher oral values and lower blood insulin values than adults.

3.7 Preprocessing features

We one-hot encoded categorical features (gender, physical_activity, and diabetic), and standardized the scale for numeric features (bmi, blood_glucose, oral, and blood_insulin). Because no features had missing values, we did not do any imputation.

3.8 Comparing classification models on training data

We compared a dummy classifier, logistic regression, and SVC model by mean cross validation score. The cross validation scores for each are below.

3.8.1 Mean Cross Validation Score for all 3 models

Table 2: Dummy classifier cross validation scores

	Unnamed: 0	test_score	train_score
0	Dummy	0.847887	0.847887
1	Logistic	0.696338	0.706056
2	SVC	0.676056	0.703521

3.9 Testing Best Model on Test Data

Since Logistic Regression had the best mean Cross Validation score, we selected it as our final model.
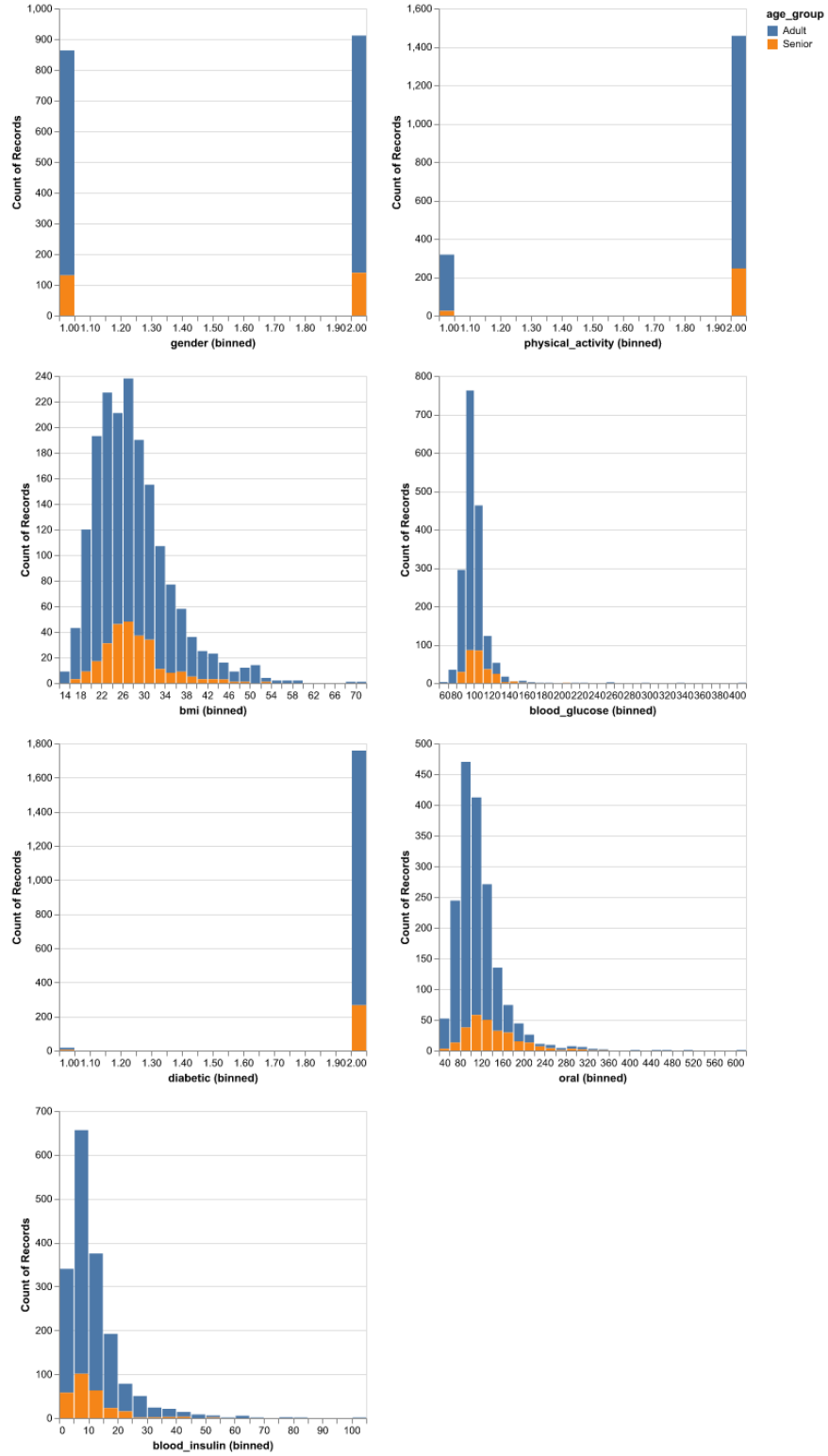
Figure 1: Feature Distributions by Age Group (groups are not stacked)

```
/Users/jasonlee/miniforge3/envs/DSCI522-39-FMJ/lib/python3.11/site-packages/sklearn/base.py:4

X does not have valid feature names, but LogisticRegression was fitted with feature names

/Users/jasonlee/miniforge3/envs/DSCI522-39-FMJ/lib/python3.11/site-packages/sklearn/base.py:4

X does not have valid feature names, but LogisticRegression was fitted with feature names
```

The model's accuracy on test data was 0.725.

3.10 Visualizing model performance

The confusion matrix (Figure 2) showed that while the model score is 0.725, it did very poorly at recall and quite poorly at precision.

This performance was reflected in the ROC curve above (Figure 3). While it could differentiate the positive class "Senior" from the negative class to some extent, the model struggled to achieve both high true positive rates and low false positive rates.

4 Discussion

The question we sought to answer was "Can information about the health and nutritional status of Americans be used to predict whether they are adults or seniors?" Our results indicated that yes, age group can be predicted with moderate accuracy (roughly 73.0%) based on health and nutritional inputs, however there is considerable room for model improvement.

We were initially surprised how high accuracy was without any hyperparameter tuning, and this turned out to be because the classes were imbalanced, meaning accuracy as a metric oversells the model's ability to distinguish the two groups. Since adults were the majority class, classifying most respondents as adults gave a high accuracy, but was not useful for identifying seniors.

In future research, we would use a metric like f1 to account for the class imbalance and conduct hyperparameter optimization to improve the model's recall and precision.

One question for future research is to identify which health and nutritional factors have the strongest predictive ability for age group. Answering that could indicate which public health interventions have the most potential to balance health outcomes across age groups in America.
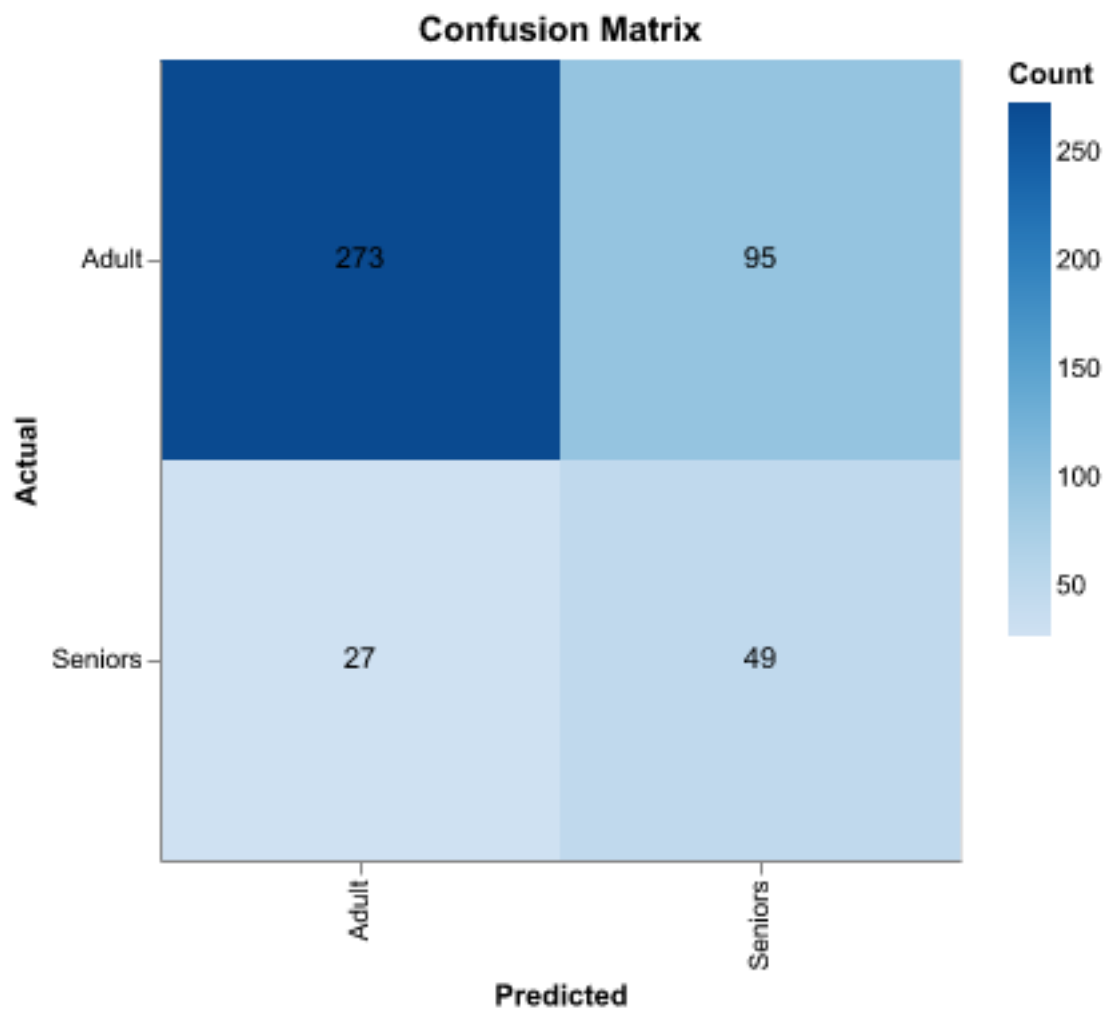
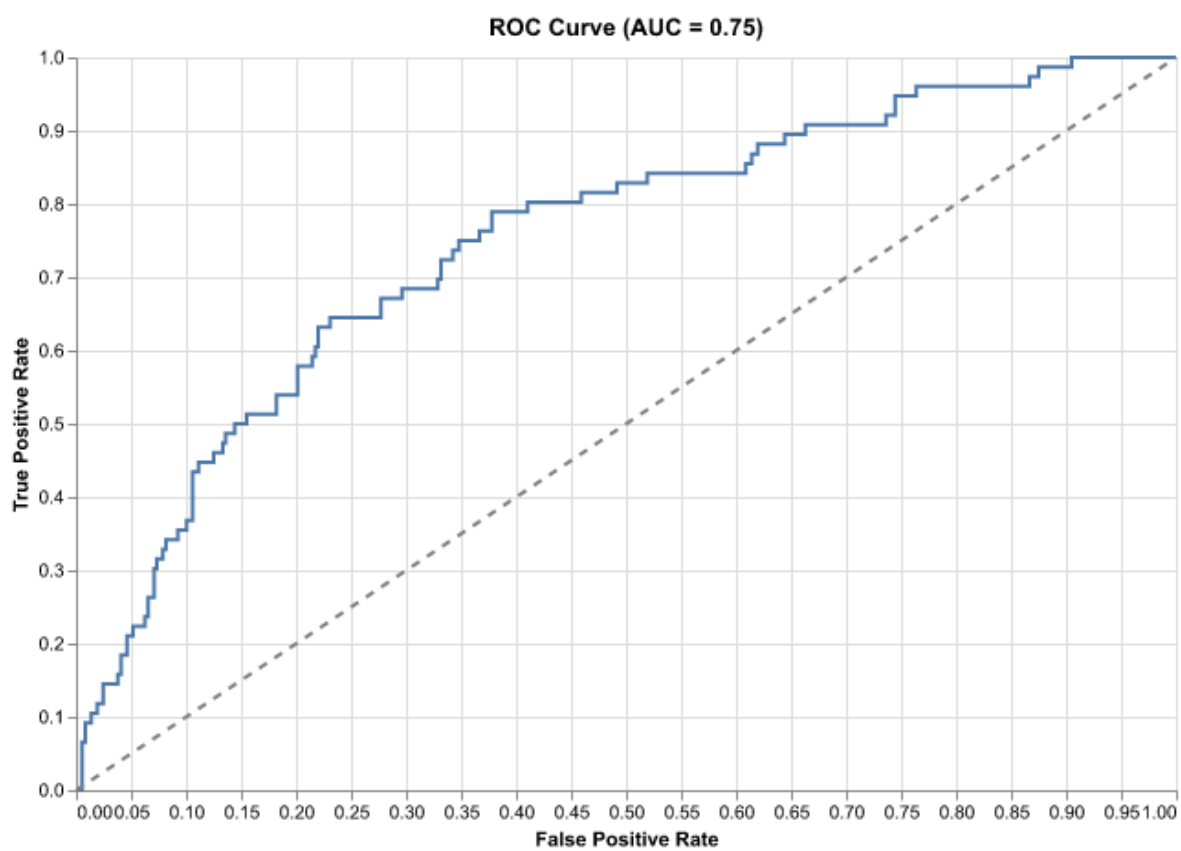Figure 2: Confusion matrix of the best model on test data

Figure 3: ROC curve of the best model on test data

References

Dinh, Andrew, and Susan Miertschin. 2016. "A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning." Semantic Scholar. https://www.semanticscholar.org/paper/A-data-driven-approach-to-predicting-diabetes-and-Dinh-Miertschin/01af1548ff1f3661d8bb813e8c35ee219a79ca9f.

Healthy Aging Center. 2022. "Aging Around the World." Colorado State University. https://www.research.colostate.edu/healthyagingcenter/2022/01/28/aging-around-the-world/.

Mukhtar, Hamid, and Sana Al Azwari. 2021. "Investigating Non-Laboratory Variables to Predict Diabetic and Prediabetic Patients from Electronic Medical Records Using Machine Learning."

Papazafiropoulou, Athanasia K. 2024. "Diabetes Management in the Era of Artificial Intelligence." *Archives of Medical Sciences. Atherosclerotic Diseases* 9: e122–28.