HSR.*health*+

Health Risk Analytics - What's the Prediction?

By Saiajay Chigurupati and Paul A Churchyard, HSR.*health*

HSR.*health*+

Health Risk Analytics - What's the Prediction?

Table of Contents

Table of Figures

Table of Tables

HSR.*health*+

Health Risk Analytics - What's the Prediction?

Executive Summary

In an exploration of predictive modeling, state and ZIP Code level data on mental health outcomes from the GeoMD Platform was used to fit multiple models to compare predictive performance. Models included an ordinary least squares regression, a principal components regression, a lasso regression, and a decision tree. The performance of the models all fared similarly, with only a four-point percentage error on average in predicting the response variable. This result points to the potential for using these models with smaller subsets of predictor variables for real-life predictive use and for using these models to identify trends for further research.

The Power of Prediction: Predictive Analytics in Health Risk Analytics

While the awareness has grown, mental health continues to be a major problem across the country. In 2022, more than 1 in 5 US adults, or 59.3 million, faced mental health problems, with only 50.3%, or 30 million, of those adults receiving mental health treatment in response1. To address the mental health crisis in our country, it is important to understand the social and demographic factors that drive these outcomes. Specifically, using these factors to predict mental health outcomes can inform future analyses and policy actions.

To investigate the potential of predictive modeling in the context of mental health outcomes, I've implemented multiple models using ZIP Code level data. In each model, the response variable was the "Percent Mental Health Not Good" variable, with a variety of social and demographic factors incorporated depending on the model. In this report, I detail and compare the results of the various predictive models implemented on the dataset.

Data Preparation

Before conducting various analyses through different models, a few things were done with the data in preparation. First, health condition variables, like the prevalence of asthma and diabetes, were excluded to focus on social determinant variables. Second, while the ZIP Code and more specific location could be incorporated in future analyses, they were excluded to remain in the scope of the project, which seeks to implement prediction models at the national level. Finally, all variables were standardized through z-score standardization. This subtracts the overall column mean and divides by the overall column standard deviation and is done to prevent one variable from affecting the model more than another simply because the scale of their units differs.

Once standardized, the data was split into training and test data sets, stratified by state, by a 70/30 split. The stratification was done to ensure that no one state ended up over-represented in either the training or test dataset. About 70% of the ZIP Code data entries from each state would be in the training set and about 30% in the test set.

Simple Linear Regression

The first model examined is the ordinary least squares regression using every predictor variable established in Section 1. The regression was fitted on the training dataset and then tested with the test dataset. The summary results are listed below:

Table 1: Simple linear regression results.

OLS Regression Results	
Dep. Variable: Mental Health Not Good	R-squared: 0.353
Model: OLS	Adj. R-squared: 0.351
Method: Least Squares	F-statistic: 128.9
No. Observations: 22745	Prob (F-statistic): 0.00
Df Residuals: 22648	Log-Likelihood: -27457.
Df Model: 96	AIC: 5.529e+04
Covariance Type: nonrobust	BIC: 5.607e+04
Omnibus: 12327.745	Durbin-Watson: 2.008
Prob (Omnibus): 0.000	Jarque-Bera (JB): 278631.511
Skew: -2.135	Prob (JB): 0.00
Kurtosis: 19.606	Cond. No. 77.2

As listed above, the R-squared was 0.353, which means that the regression can predict about 35% of the variation in the percent mental health not good outcome. The root mean square error was 0.788 when tested with the test dataset. Since the data is standardized, we can interpret this as the regression being on average 0.79 standard deviations off in predicting the mental health not good variable.

Principal Components Analysis and Linear Regression

The next modeling technique used on the dataset is the principal components analysis (PCA) and linear regression on a specified number of components. In general, the goal of PCA is dimension reduction. This is achieved by compressing and rotating the data into a few principal component eigenvectors (or linear combinations of the original predictors), which can limit factors like multicollinearity and redundancy. PCA was applied to the dataset and then split into the same training/test split as before by using the same random seed. Then, the ordinary least squares regression was fitted for the first n principal components starting with the first principal component, followed by the first and second, then the first, second, and third, and so on until all principal components were accounted for. Below are the graphs for how the root mean square error (Figure 1), the R-squared changes (Figure 2), and explained variances (Figure 3) each change with the number of principal components:

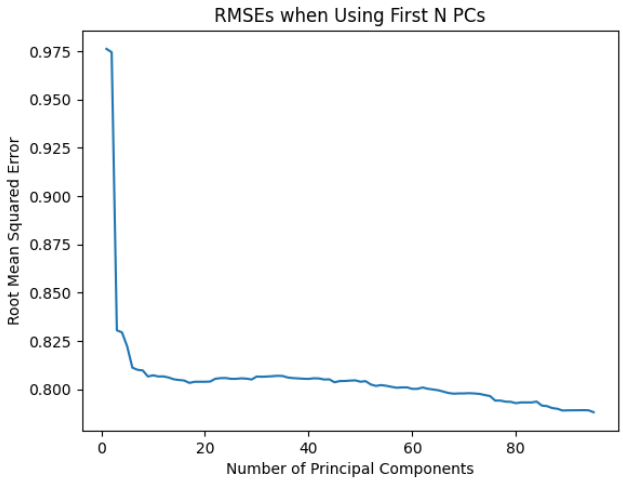

Figure 1: RMS error by number of principal components.

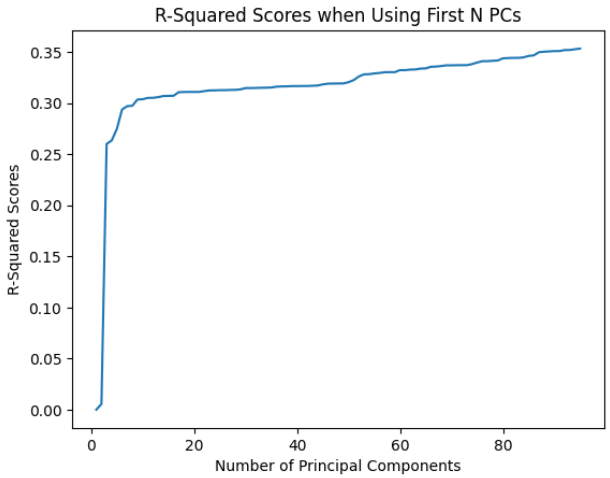

Figure 2: R-squared scores by number of principal components.

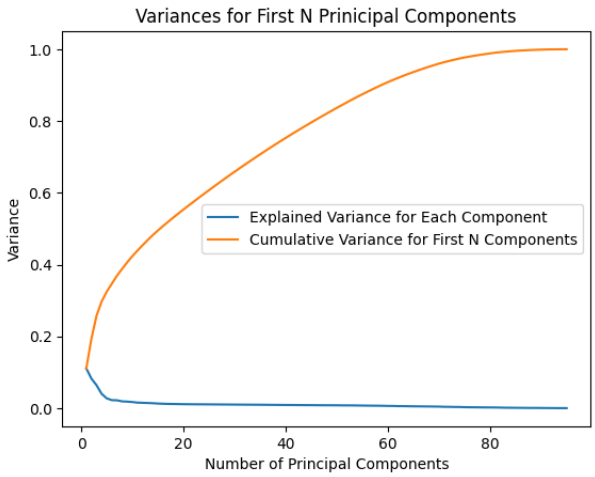

Figure 3: Variance by number of principal components.

As we include more components, the R-squared value continues to rise without any dip. If we look at the adjusted R-squared value, which incurs a penalty for including extraneous terms, we may see dips towards the end of the graph. The largest improvements in RMSE occur within the first 20 principal components. Zooming into that portion of the graph, we see below that sharp corners occur at three, six, and nine principal components, with a slight uptick in the RMSE shortly after nine principal components.

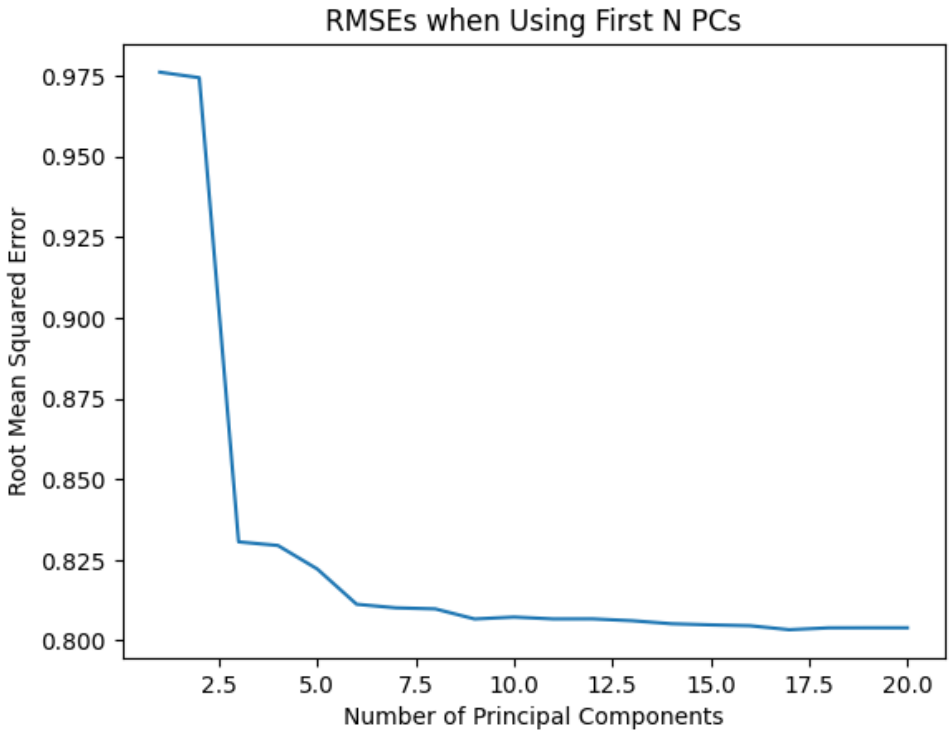

Figure 4: RMS error for first 20 principal components.

Below is the table summarizing the R-squared values, RMSE's, and cumulative variances for three, six, and nine principal components. Compared to the 0.788 RMSE from using every predictor, the 0.807 RMSE from nine principal components is impressive. Furthermore, given the sheer number of predictors and the social factors they span, it is also impressive that three principal components can explain over a fourth of the variance, six components can explain over a third of the variance, and nine components can explain over two-fifths of the variance:

Table 2: R-squared, variance, and RMSE by number of PC's.

Number of PC's	R-Squared	Cumulative Variance	RMSE
3	0.260	0.257	0.831
6	0.294	0.347	0.811
9	0.303	0.406	0.807

To see what original variables contribute the most to these components, we can analyze the eigenvectors by squaring and summing them and then finding the relative proportion of each variable in that sum. Doing this for the first three, six, and nine eigenvectors, we find the following variables rank top 10 for each set of principal components:

Table 3: Top 10 variables for each set of PCs.

Rank	First 3 PCs	First 6 PCs	First 9 PCs
1	Percent Low Occupant Density (3.42%)	Percent Hispanic (2.91%)	Percent Hispanic (2.93%)
2	Percent Have Public Health Insurance (2.97%)	Percent Other Race (2.58%)	Percent Between the ages of 10 and 19 (2.59%)
3	Percent Households With Social Security (2.76%)	Percent Not Fluent in English (2.47%)	Percent Male (2.43%)
4	Percent Households with Smartphone Access (2.62%)	Percent Renter Occupied Housing (2.40%)	Percent Other Race (2.38%)

Rank	First 3 PCs	First 6 PCs	First 9 PCs
5	Percent Owner Occupied Housing (2.59%)	Percent Over the age of 65 (2.35%)	Percent Military (2.29%)
6	Percent Households with Earnings (2.59%)	Percent Owner Occupied Housing (2.18%)	Percent Under the age of 10 (2.18%)
7	Percent Have Health Insurance (2.58%)	Percent Housing Units built before 1950 (2.08%)	Percent Female (2.18%)
8	Percent Over the age of 65 (2.56%)	Percent Rent greater than 30 Percent of Income (2.05%)	Percent Housing Units built before 1950 (2.09%)
9	Percent Renter Occupied Housing (2.48%)	Percent Have Public Health Insurance (1.87%)	Percent Veteran (2.06%)
10	Percent Households with Broadband Access (2.48%)	Percent Some College or Above (1.81%)	Percent Not Fluent in English (2.04%)

Lasso Regression

The lasso regression uses a penalty function that incorporates an alpha value to eliminate predictors. The higher the alpha value the more predictors are cut out of the regression. At an alpha of zero, the regression runs as an ordinary least squares regression and includes every predictor. At a large enough alpha, the predictors are all eliminated, leaving the null model. To evaluate which alpha value works best, we first implemented a cross-validated lasso regression model spanning a spectrum of alpha values. Along with finding the optimal alpha value, we plotted the minimum mean square error for each alpha value across the five folds analyzed at each alpha:

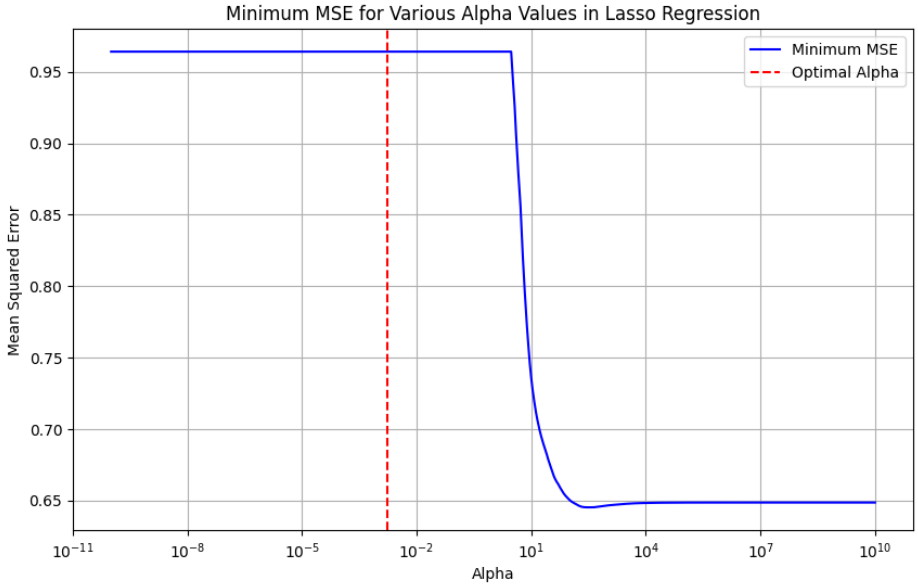

Figure 5: Lasso regression results.

Alphas were tested from a range of 10^{-10} to 10^{10}. The optimal alpha determined by the cross-fold validation on the training data was 0.00165. There is also a minimum in the blue curve at an alpha of 404.13. For a third alpha value to try, we will also use the maximum alpha of 10^10. The corresponding RMSE's and scores in Table 4 below:

Table 4: Scores and RMSEs for each alpha.

Alpha	Score	RMSE
0.00165	0.351	0.789
404.13	0.0	0.976
10^10	0.0	0.976

For the two larger alpha values, the null model is returned. While the RMSE and score for the optimal alpha regression are similar to the original least squares regression, a few predictors are excluded through the lasso penalty term, including:

- Percent Over the age of 65
- Percent Graduate Degree
- Percent Hispanic
- Percent In Labor Force
- Percent Low Occupant Density
- Percent Median Housing Value
- Percent Have Public Health Insurance

- Percent Renter Occupied Housing
- Percent Households with Smartphone Access

Decision Trees

The final model analyzed was the decision tree. Ideally, I would have been able to test various decision trees across a wide range of cost-complexity pruning alpha values. The cost-complexity pruning alpha determines how many nodes in a tree the model will prune. A higher alpha value assigns a higher cost to more complex trees, thus leading to more pruned trees with fewer nodes. Conversely, a lower alpha value leads to more complex and less pruned trees with more nodes. Due to limits in processing power, I instead opted to test various depths of trees, from 1 to 20, with the results shown in Figure 6 below.

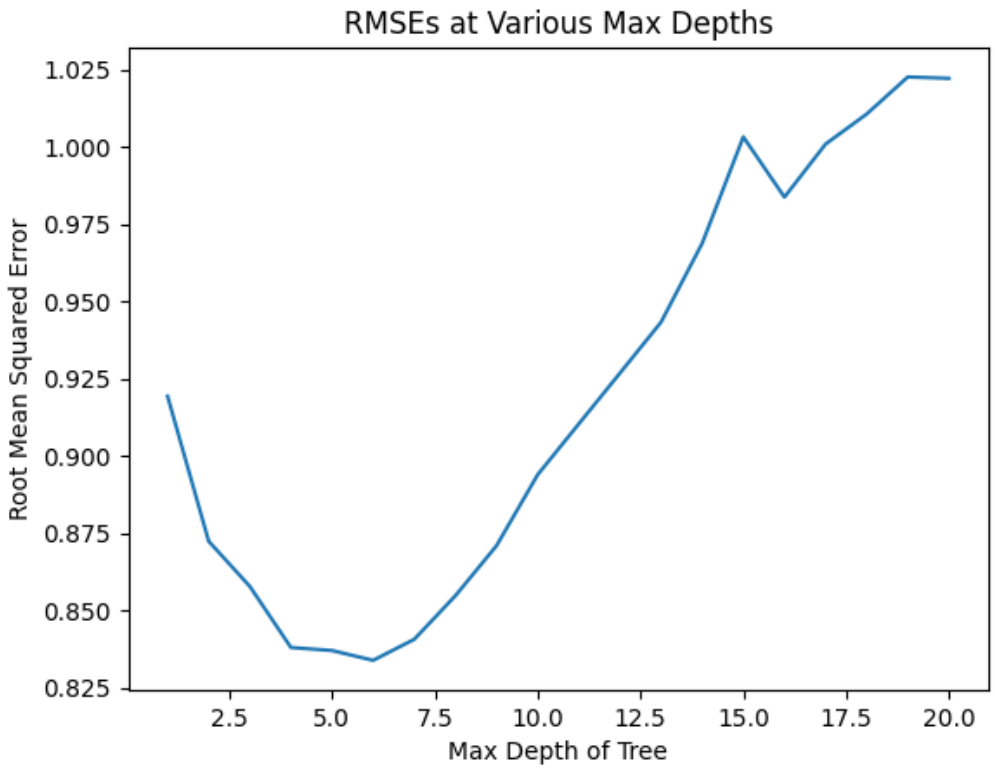

Figure 6: RMSE by tree depth.

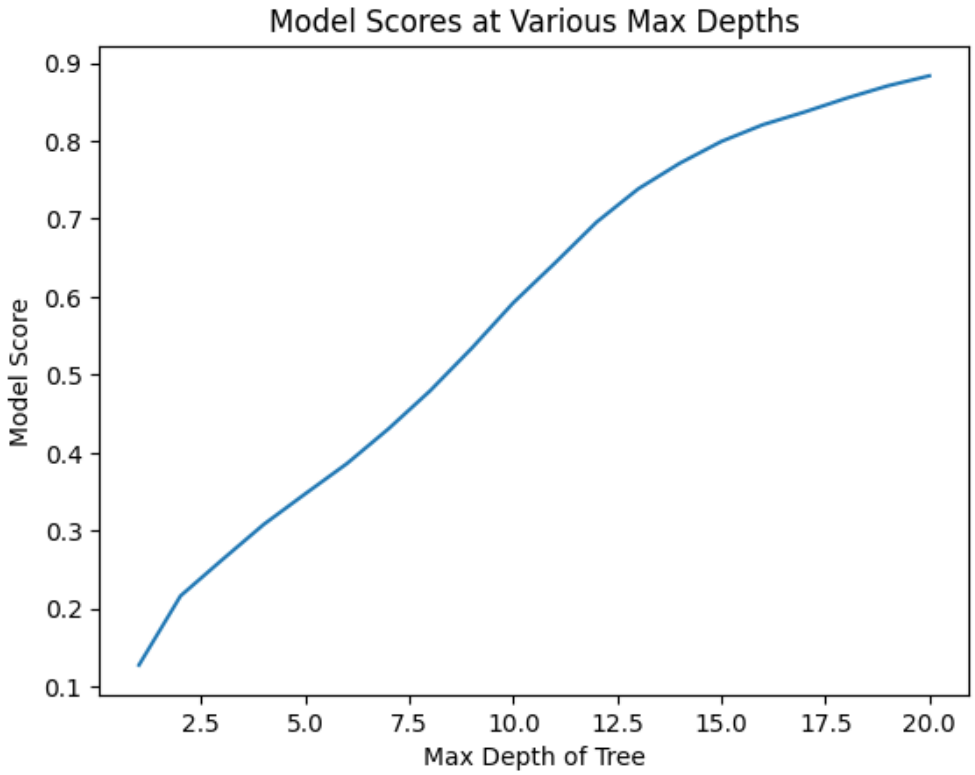

Figure 7: Model score by depth of tree.

In Figure 7, we unsurprisingly see the model score better with the training data as max depth is increased. However, the RMSE rises after a max depth of 6 nodes, indicating overfitting on the training data at higher depths. At the max depth of 6, the score is 0.386, while the RMSE is 0.834. The Python module used to generate the tree diagrams could not export the full tree in a readable document format. Instead, Figure 8 presents the cut off tree a limited depth for better readability and presentability.

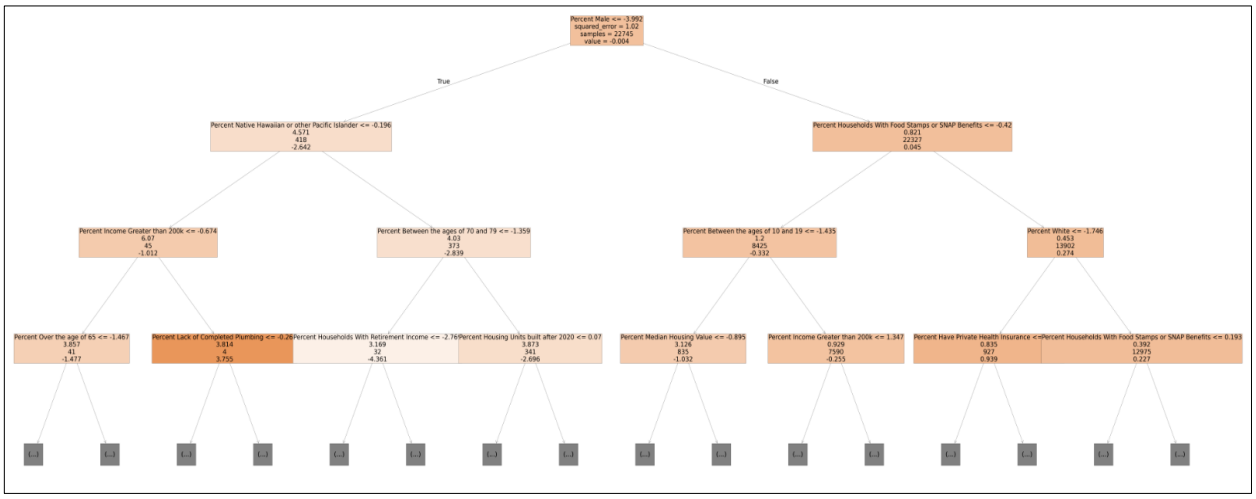

Figure 8: Cut-off decision tree.

Looking at the top of the tree, the first node divides based on gender, predicting better mental health outcomes for significantly high female proportion ZIP Codes (4 standard deviations or lower from the mean male proportion). Looking to the right half of the tree, which covers most ZIP Codes, the tree then incorporates variables such as:

- "Percent Households with Food Stamps or SNAP Benefits", which predicts better mental health outcomes for ZIP Codes with a smaller percent
- "Percent Between the ages of 10 and 19", which predicts better mental health outcomes for ZIP Codes with a smaller percent
- "Percent White", which predicts better mental health outcomes for ZIP Codes with a larger percent.

Refining the Model: Insights & Next Steps in Risk Prediction

Overall, looking at the scores and RMSEs for the best performer of each model as determined by lowest RMSE value, we see that the models all perform similarly with the decision tree standing out as a better fitter of the training data. However, at the cost of potentially overfitting and offering worse predictive power. Given the RMSEs are of the z-score standardized form of the response variable, we can interpret most of these models as being off by about 0.8 standard deviations on average. Given the standard deviation in the response variable was 4.94 percent points, this would translate to being off by 3.95 percentage points.

Table 5: RMSEs and score values by model.

Model	RMSE	Score
Simple OLS Regression	0.788	0.353
PCA Regression [9 Components]	0.807	0.303
Lasso Regression	0.789	0.351
Decision Tree [Max Depth 6]	0.834	0.386
Null Model	0.976	0

Of note, the variable "Percent Hispanic" was a top contributor to the nine principal components in the PCA regression but was dropped in the lasso regression. With predictive modeling, it is important to note that these are not causal determinations: just because one variable is weighted more in the model or kept in the decision tree does not mean that the variable is a direct cause of better or worse mental health outcomes. However, there is potential for these models to point in future directions.

For example, the decision tree surprisingly predicted that ZIP Codes with an extremely high proportion of women face relatively better mental health outcomes. This could point to a hidden trend that warrants future research.

Additionally, these models along with considerations of predictive power should be evaluated in terms of ease of comprehension and number of variables used. Ranked in terms of their ease of explainability to someone less familiar with these models, (from easiest to hardest):

1. OLS regression
2. Decision trees
3. Lasso regression
4. Principal components analysis

In applicable uses, users will rarely have all social and demographic factors on hand to plug in and generate a prediction. Aside from the decision tree, the other regressions use almost every variable in some sense or another. Even the PCA regression components are generated from some linear combination of every predictor variable in the dataset.

Thus, multiple future directions can be taken with this work. First, various smaller subsets of predictors can be tried in these models to see how large the tradeoff in predictive power is with fewer inputs. A small loss in predictive power can be accepted if it means significantly fewer predictors are needed. On the other hand, more predictive modeling like k-nearest neighbors, ridge regression, and random forests can be implemented with this data. While models like the random forest lose the ease of interpretability but gain in predictive power making it a potentially worthwhile tradeoff depending on context and use case.

In addition, different geographical clustering can be used. Rather than just using each ZIP Code as its own data point, it may be worth clustering into counties, states, or regions (e.g., Midwest, New England, etc.).

While the scores were somewhat low (all under 0.4), an RMSE that translates to less than four percentage points provides a promising start to the application of predictive analyses with this immense dataset towards anticipating instances of poor mental health.

References

1. National Institute of Mental Health. (2024, September). *Mental illness*. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/mental-illness