Exploring Social Inequity Through Machine Learning: Analyzing Bias in Healthcare

Madelyn Boaz

Department of Math and Computational Sciences

CS499: Senior Seminar

Dr. James Rauff

December 12, 2024

Racial and ethnic health disparities are considered some of the major challenges for public health in the United States. Research indicates that minority populations typically report their general health being poorer than that of their White counterparts. These worse reports by minority groups compared with Whites result from a broad range of underlying causes that include socioeconomic disadvantage, institutional racism, and access and availability of healthcare services. This project applies machine learning techniques, like Linear Regression, Random Forest, and Decision Tree models, to understand the relationship between race/ethnicity and self-reported poor health. The dataset contains health outcomes for different demographic groups. The aim is to find predictive patterns that indicate which demographic variables are most strongly related to poor health status. These findings have implications for public health strategy and policy and can inform priorities on targeted interventions that would positively impact health outcomes.

Racial and ethnic health disparities stand out as one of the most serious challenges to public health, in which there is a difference in access to quality care, ultimately affecting the general outcome of health. Minorities, such as Blacks, Hispanics, Native Americans, and other groups, experience higher rates of chronic illnesses, like diabetes, hypertension, and heart disease, compared to Whites. Systemic factors, such as socioeconomic inequality, limited access to healthcare facilities, and discrimination within medical systems, are sometimes responsible for these disparities. Other demographic factors that affect health outcomes include education, housing, and employment opportunities. Targeted interventions, like increasing access to preventive care, improving the cultural competency of healthcare providers, and promoting policies designed to decrease the socioeconomic gaps that affect health challenges, are required to reduce these inequities. There are ways to prioritize health equity for improved outcomes in the quality of life of all individuals, especially in the most vulnerable populations.

One of the problems surrounding this topic is that health disparities persist due to limited tools for analyzing demographic predictors. The project aims to identify how demographic factors affect health outcomes. Despite efforts to alleviate these differences in health outcomes, identifying and addressing the root causes remains challenging. There is a lack of tools to effectively analyze how demographic factors influence health outcomes. Using machine learning, demographic predictors of poor health will be identified. We hope to gain actionable insights through this analysis.

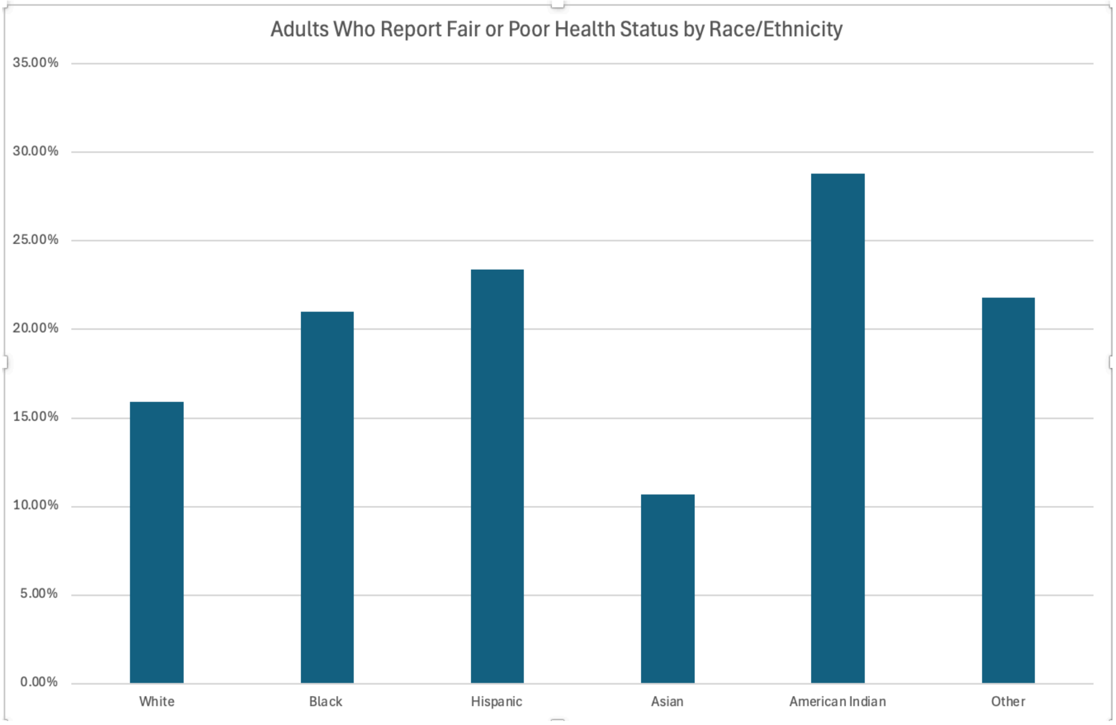
The overall objective of this project was to apply machine learning algorithms such as Linear Regression, Random Forest, and Decision Tree for analysis. The tools I used to do so include Python programming language, the data manipulation library Pandas, the machine learning library scikit-learn, and JupyterLab for coding and visualization.

The data used for this analysis was titled “Adults who Report Fair or Poor Health Status by Race/Ethnicity.” This data comes from Kaiser Family Foundation, which is a trusted source for health statistics and trends in the U.S. It shows self-reported health status by race and ethnicity in the U.S. as well as in every state. One key trend taken from the data is there are higher rates of poor health reported among Black, Hispanic, and Native populations, while there are lower poor health rates reported among White and

Asian populations. The data highlights disparities that reflect systemic inequities in access to care.

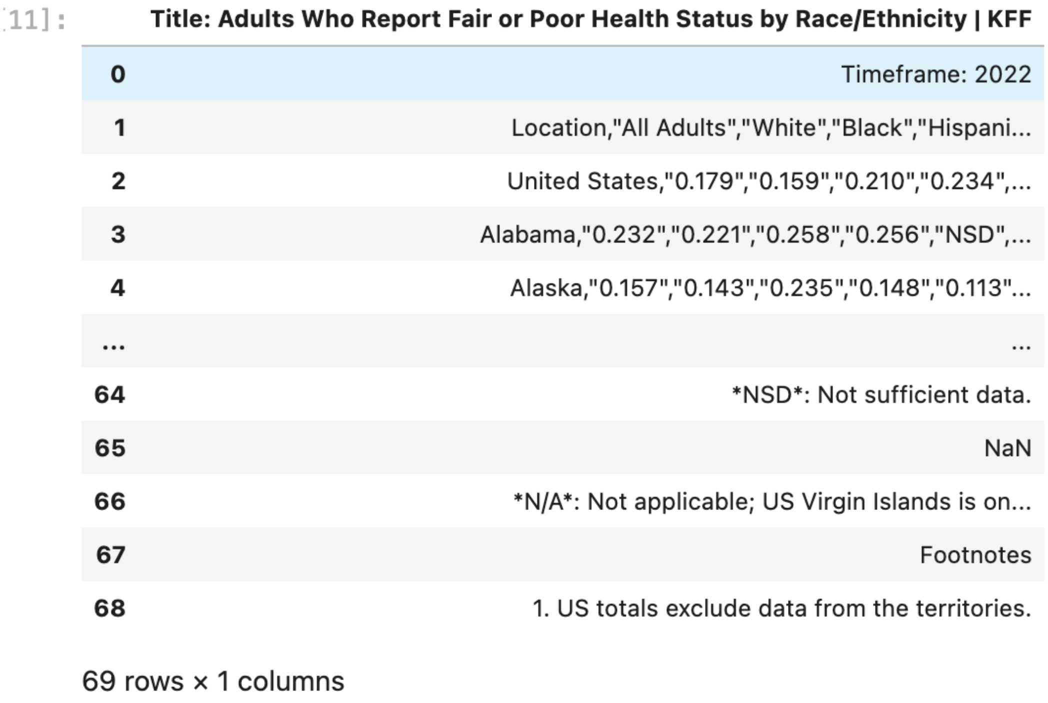
The data comes from the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a state-based, ongoing, random-digit-dialed telephone survey targeting non-institutionalized civilian adults aged 18 years and older. The BRFSS is administered and supported by the Centers for Disease Control and Prevention (CDC). It collects data about health behaviors, chronic diseases, and use of health care services. Its large-scale design ensures comprehensive, nationally representative data, with annual sample sizes of over 400,000 respondents across all 50 states, the District of Columbia, and U.S. territories. The large sample size makes it easier to conclusively find insights about the data.

Shown below is a simple visualization of the raw data.



The bar chart illustrates self-reported health status by race/ethnicity in the U.S. as a whole. It shows the percentage of U.S. adults of each race that report poor or fair health. As you can see, populations such as Black, Hispanic, and Native Americans report higher rates of poor health, which aligns with broader research that suggests they face systemic barriers. White and Asian populations, on the other hand, report lower rates of poor health. This could be due to better access to resources or a healthcare system that serves them more effectively.

Now that the raw data is understood, it is time to start working on it. First, the data had to be imported before cleaning and preprocessing could begin. This is what the raw data looked like.



11]:

	Title: Adults Who Report Fair or Poor Health Status by Race/Ethnicity | KFF
0	Timeframe: 2022
1	Location,"All Adults","White","Black","Hispani...
2	United States,"0.179","0.159","0.210","0.234",...
3	Alabama,"0.232","0.221","0.258","0.256","NSD",...
4	Alaska,"0.157","0.143","0.235","0.148","0.113"...
...	...
64	*NSD*: Not sufficient data.
65	NaN
66	*N/A*: Not applicable; US Virgin Islands is on...
67	Footnotes
68	1. US totals exclude data from the territories.

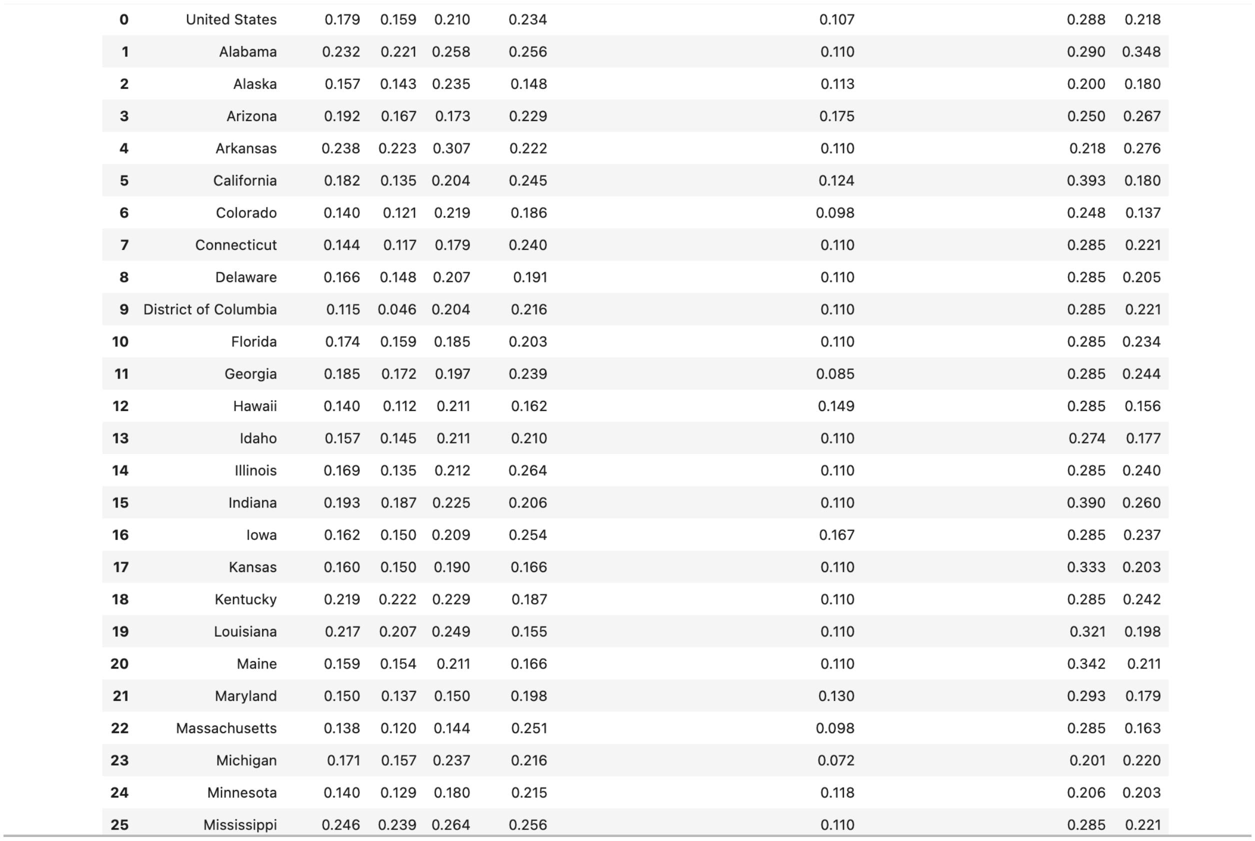
69 rows × 1 columns

There are issues with this data. In this format, it is a mess, which makes it unusable. All the data is in one column, just separated by commas. There are also unnecessary rows and columns. There are quotes around values, which makes the data unusable for analysis. There are also N/A or empty values that must be handled. It is critical for analysis to prepare the data into a usable format.

Next up was preprocessing the dataset. First, I imported the dataset using Pandas and specified delimiters and quotes for accurate reading. This ensured the dataset was imported correctly, which is necessary to avoid issues with messy data. Then, I split a combined column into individual demographic columns to isolate the columns by variables, enhancing the data's usability. I then removed unnecessary rows and columns, ensuring that only relevant information was retained. Lastly, I reset row indices for a clean dataset structure. The clean structure allows for smoother navigation and processing of the data in later stages.

Then, it was necessary to handle missing values and clean the data. The first step in this process was replacing inconsistent values like *Unknown* and empty strings with NaN. This makes handling data more consistent. Then, I converted demographic columns to numeric data types for analysis which was crucial for performing meaningful analyses. Then, I imputed missing values in numeric columns using the median of each column. This maintained the data's central tendency without being overly influenced by outliers. Last, I removed unwanted quotes around values to ensure uniformity which helps avoid errors during further processing.

Pictured below is the cleaned data.



0	United States	0.179	0.159	0.210	0.234	0.107	0.288	0.218
1	Alabama	0.232	0.221	0.258	0.256	0.110	0.290	0.348
2	Alaska	0.157	0.143	0.235	0.148	0.113	0.200	0.180
3	Arizona	0.192	0.167	0.173	0.229	0.175	0.250	0.267
4	Arkansas	0.238	0.223	0.307	0.222	0.110	0.218	0.276
5	California	0.182	0.135	0.204	0.245	0.124	0.393	0.180
6	Colorado	0.140	0.121	0.219	0.186	0.098	0.248	0.137
7	Connecticut	0.144	0.117	0.179	0.240	0.110	0.285	0.221
8	Delaware	0.166	0.148	0.207	0.191	0.110	0.285	0.205
9	District of Columbia	0.115	0.046	0.204	0.216	0.110	0.285	0.221
10	Florida	0.174	0.159	0.185	0.203	0.110	0.285	0.234
11	Georgia	0.185	0.172	0.197	0.239	0.085	0.285	0.244
12	Hawaii	0.140	0.112	0.211	0.162	0.149	0.285	0.156
13	Idaho	0.157	0.145	0.211	0.210	0.110	0.274	0.177
14	Illinois	0.169	0.135	0.212	0.264	0.110	0.285	0.240
15	Indiana	0.193	0.187	0.225	0.206	0.110	0.390	0.260
16	Iowa	0.162	0.150	0.209	0.254	0.167	0.285	0.237
17	Kansas	0.160	0.150	0.190	0.166	0.110	0.333	0.203
18	Kentucky	0.219	0.222	0.229	0.187	0.110	0.285	0.242
19	Louisiana	0.217	0.207	0.249	0.155	0.110	0.321	0.198
20	Maine	0.159	0.154	0.211	0.166	0.110	0.342	0.211
21	Maryland	0.150	0.137	0.150	0.198	0.130	0.293	0.179
22	Massachusetts	0.138	0.120	0.144	0.251	0.098	0.285	0.163
23	Michigan	0.171	0.157	0.237	0.216	0.072	0.201	0.220
24	Minnesota	0.140	0.129	0.180	0.215	0.118	0.206	0.203
25	Mississippi	0.246	0.239	0.264	0.256	0.110	0.285	0.221

It is now in a structured and uniform format. All unnecessary rows and columns have been removed, missing values have been handled, and demographic columns have been converted to numerical types. The data is now free from inconsistencies, making it suitable for exploratory data analysis and modeling. I enhanced the data's quality, reliability, and usability to uncover meaningful insights.

The Machine Learning Models that I used were Linear Regression, Random Forest, and Decision Tree. Linear regression works by predicting a target value by finding a linear relationship between features and the target. It assumes a direct relationship between the input and output variables making it ideal for estimating outcomes where such relationships exist. Decision Trees create a flowchart-like tree structure to make decisions based on feature splits. It splits data into branches, making Decision Trees easy to visualize and interpret. It can overfit, but this can be mitigated when combining multiple Decision Trees into a Random Forest. A Random Forest is a collection of decision trees that improves accuracy by averaging predictions. It uses multiple Decision Trees to vote on predictions. It also offers insights like feature importance which measures what factors have the most influence.

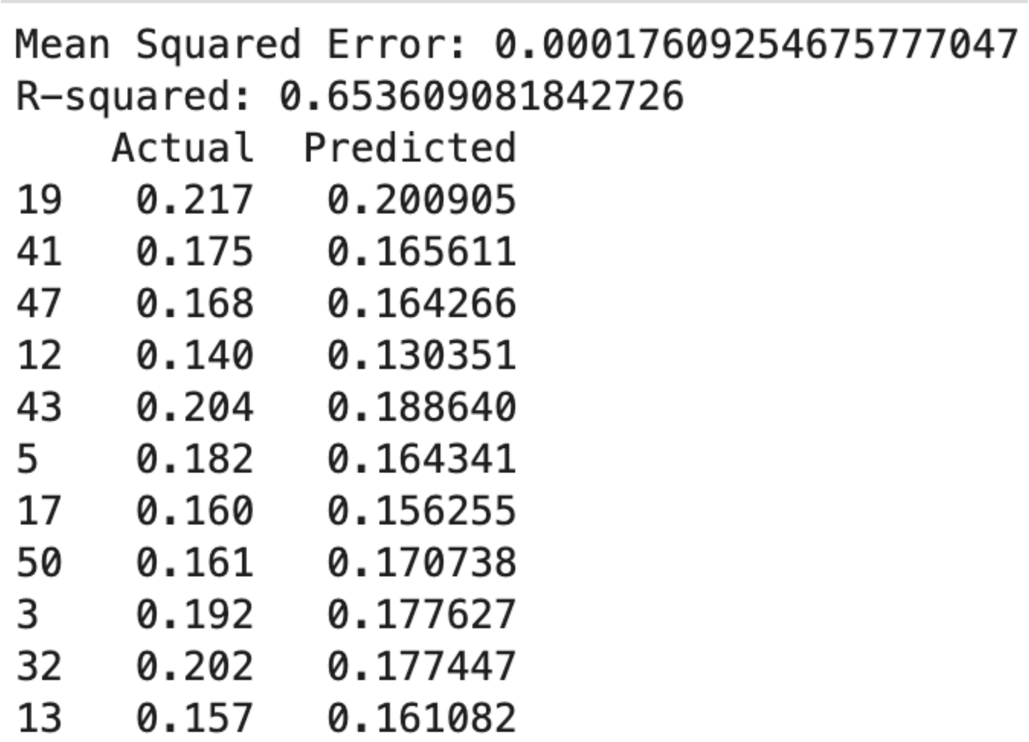
Model evaluation metrics used for the analysis included R-squared and Mean Squared Error. R-squared measures how well the model explains the variance in the target variable. Values closer to 1 indicate a better fit. It helps us understand the proportion of variance in the target variable that is explained by the model. Mean Squared Error averages the squared differences between predicted and actual values. It is the average magnitude of prediction errors. Lower values indicate better performance because they suggest that the predicted values are closer to the actual values, which reduces prediction error. We must use these metrics to ensure we have accurate model performance.

The algorithms were then used to answer the question- what is the relationship between race/ethnicity and the percentage of people reporting poor health in the U.S.? This question explores how race/ethnicity may impact the percentage of people reporting poor health in the U.S. Understanding these patterns can help identify disparities and inform targeted interventions to address health inequalities across different demographic groups.

To implement the machine learning models, target and predictor variables must be defined. The target variable in this case was All Adults (percentage of people reporting poor health), and the predictor variables were the race/ethnicity categories. These included White, Black, Hispanic, Asian/Native Hawaiian or Pacific Islander, American Indian or

Alaska Native, and Other. Predictor variables are independent variables that we use to make predictions on the target variable. The target variable is the dependent variable that we aim to understand or predict.

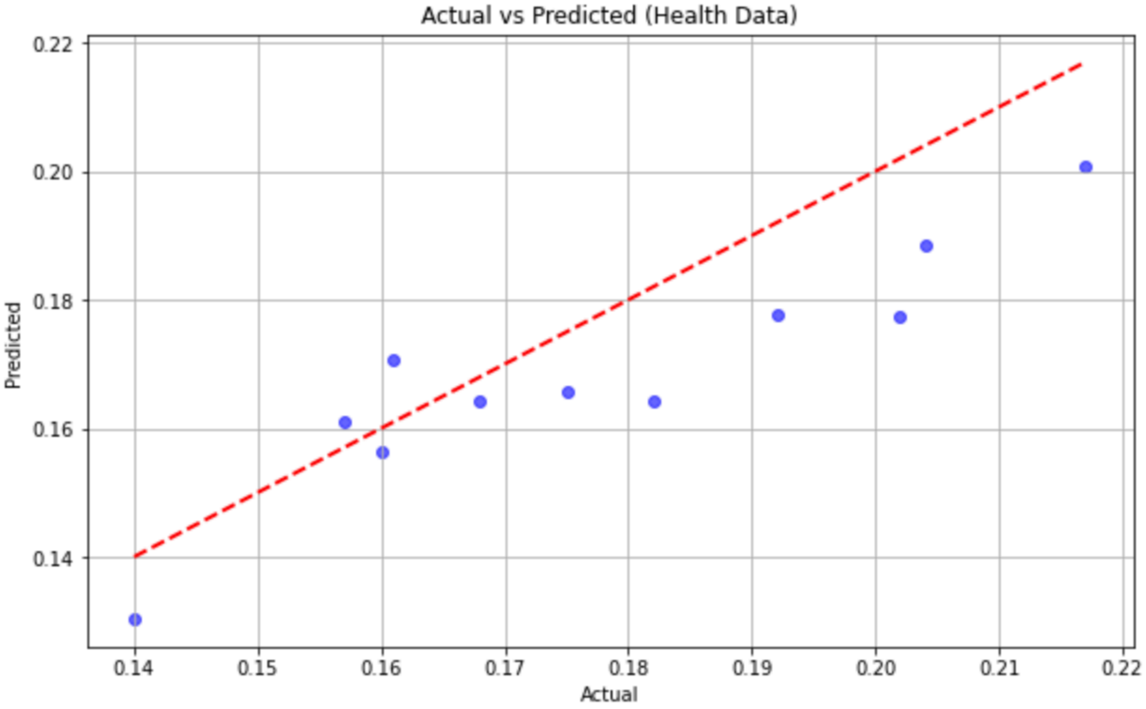
The first model that was employed was Linear Regression. The necessary packages from the sklearn library first had to be imported. Next, the algorithm was run, and the results were displayed as seen below.



```
Mean Squared Error: 0.00017609254675777047
R-squared: 0.653609081842726
    Actual  Predicted
19   0.217   0.200905
41   0.175   0.165611
47   0.168   0.164266
12   0.140   0.130351
43   0.204   0.188640
5    0.182   0.164341
17   0.160   0.156255
50   0.161   0.170738
3    0.192   0.177627
32   0.202   0.177447
13   0.157   0.161082
```

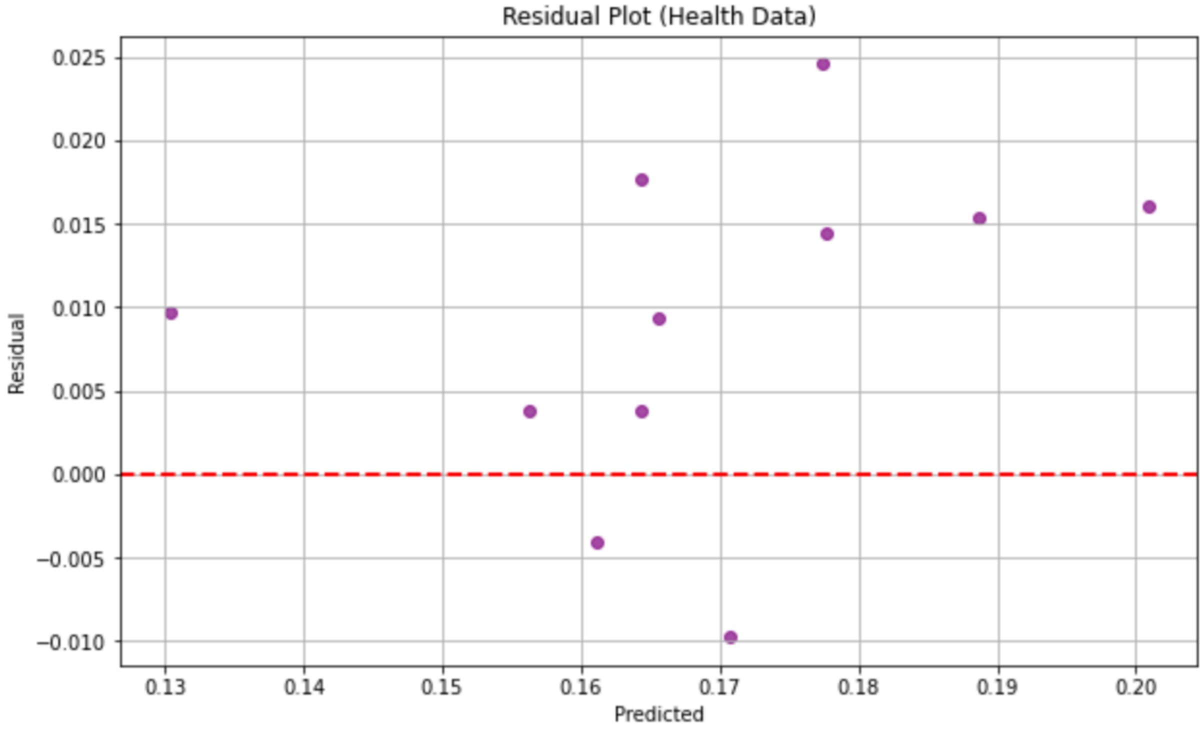
The MSE of about .000176 indicates that the average square difference between the predicted and actual values is very small. Smaller MSE values indicate higher accuracy. The numbers shown in the actual and predicted columns represent the percentages of people reporting poor health as ratios. The R-squared value tells us that about 65% of the variance in the target variable is explained by the features. It is not perfect, but it is a decent indication of the model's explanatory power. This model shows disparities in poor health across racial groups.

Two graphs are provided to visualize the results of the Linear Regression model. The plot below shows the average versus predicted values from the model.



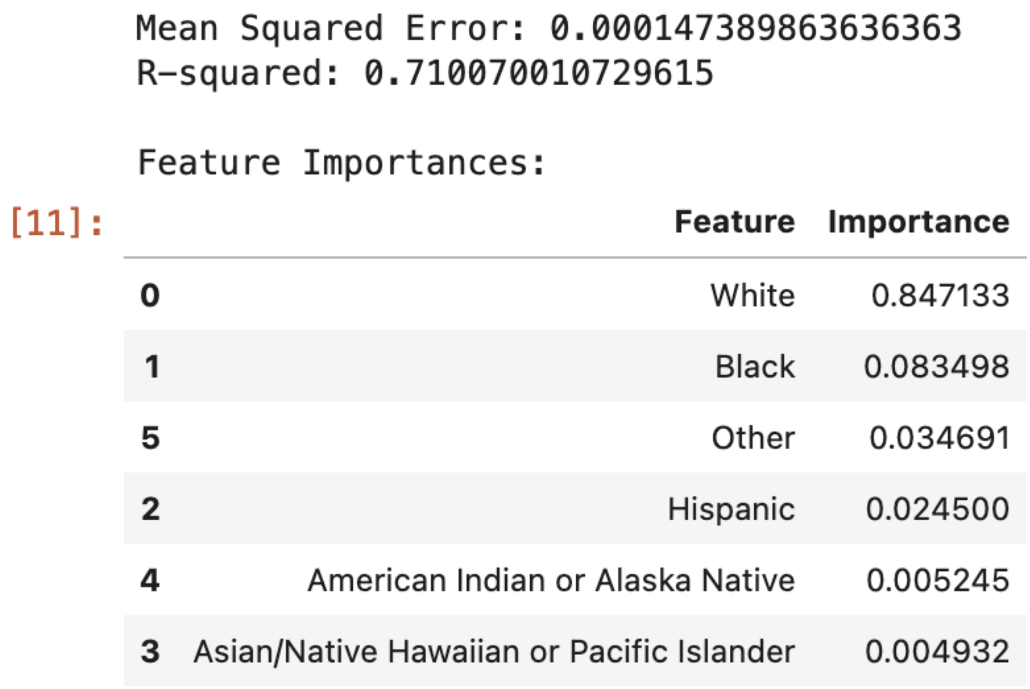
This visualization shows how closely the predicted values align with the actual data points. The closer to the diagonal line that the points lie, the more accurate. As seen, most of the predicted values are close to the actual values. This is consistent with the Mean Squared Error and the R-squared values.

Pictured below is the residual plot which shows how close to zero the errors between the actual and predicted values lie.



The errors generally lie close to zero as seen in the graph. The largest error is under .025. The graphs further support the model's overall accuracy while showing areas of potential improvement.

The second model that was used to answer the posed question was Random Forest. After training and running the model, the results were displayed as seen.



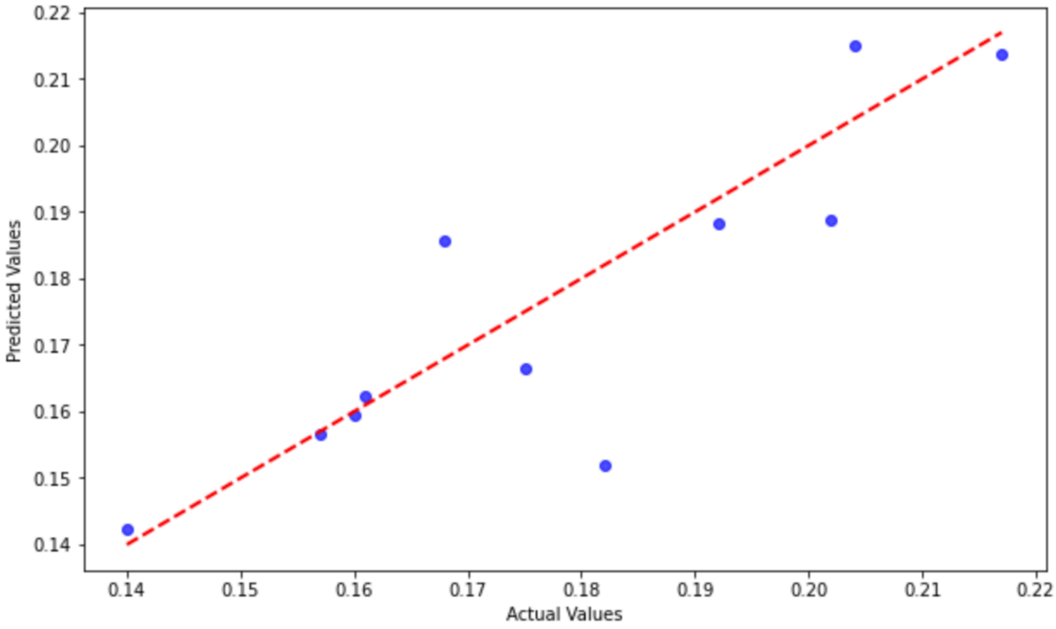
```
Mean Squared Error: 0.000147389863636363
R-squared: 0.710070010729615

Feature Importances:
```

[11]:

	Feature	Importance
0	White	0.847133
1	Black	0.083498
5	Other	0.034691
2	Hispanic	0.024500
4	American Indian or Alaska Native	0.005245
3	Asian/Native Hawaiian or Pacific Islander	0.004932

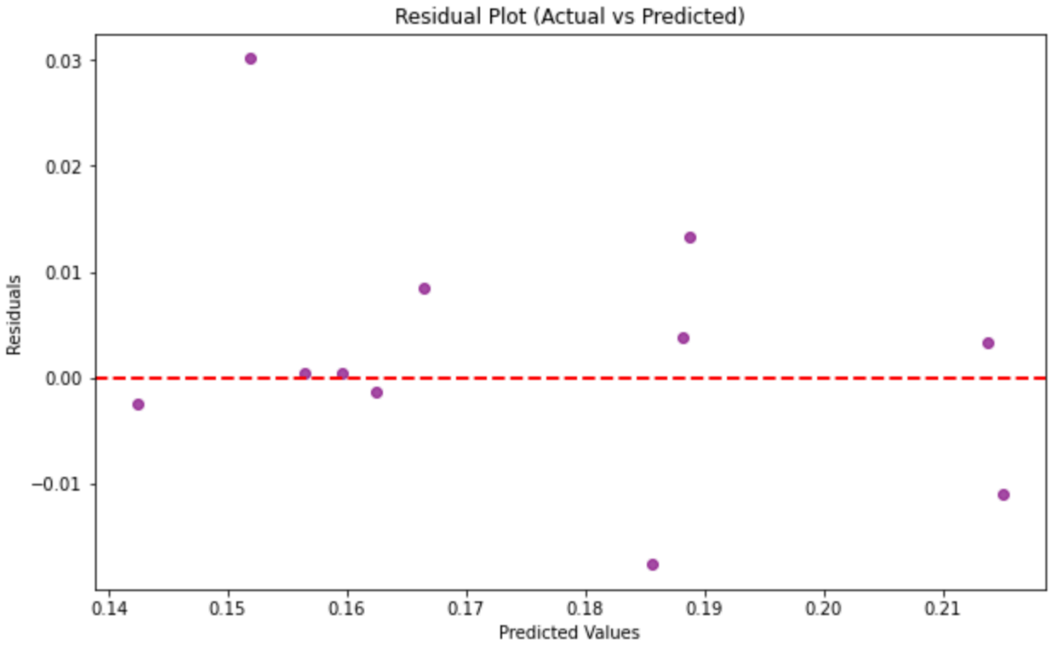
The MSE of about .000147 shows that this model performs well as there is minimal prediction error. The R-squared value of about .7101 suggests that the model explains about 71% of the variance in the target variable, which is a pretty good fit. The better performance of the Random Forest model over the Linear Regression model indicated that the relationship is partly non-linear. “White” being at the top of the Feature Importance shows that white disparities significantly influence health outcomes. This is likely due to high variability in demographic conditions amongst this group. The relatively low importance of smaller racial groups may stem from underrepresentation or lower variance within this group. However, this doesn’t diminish the real-world importance of these groups but may rather highlight limitations in the dataset.

The same two types of visuals were used to illustrate the results of the Random Forest Model. The first visual, once again, shows the predicted versus the actual values.



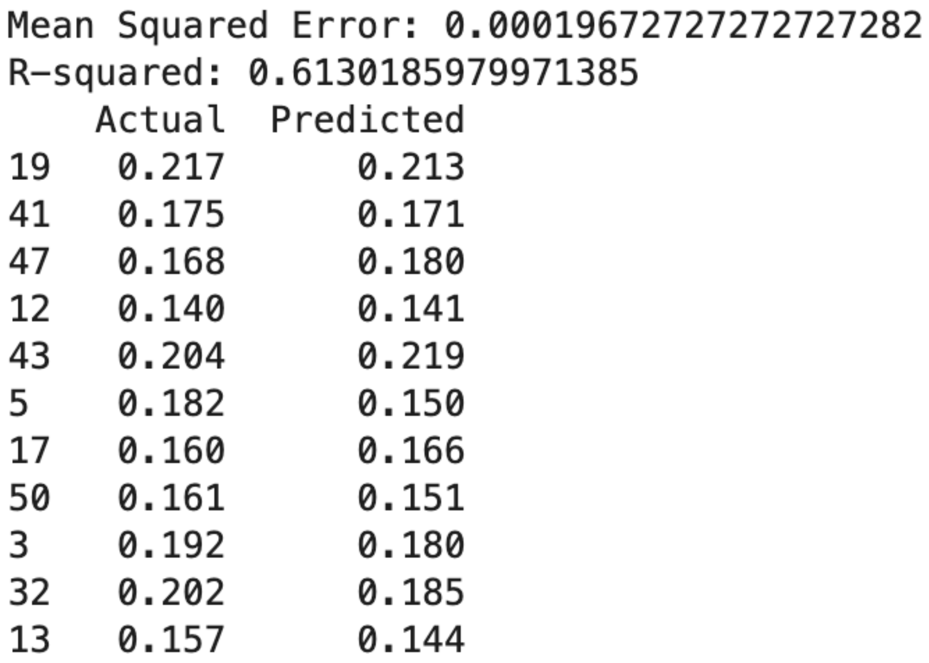
The points generally follow the slops of the line, with only a couple of major outliers. Most predicted values are close to the actual values, which is consistent with the MSE and R-squared values.

Pictured below is the residual plot for the Random Forest model.



The residual plot shows the error between the predicted and actual values. The largest error is just over .03, but the others are small. The graphs further support the model's overall accuracy.

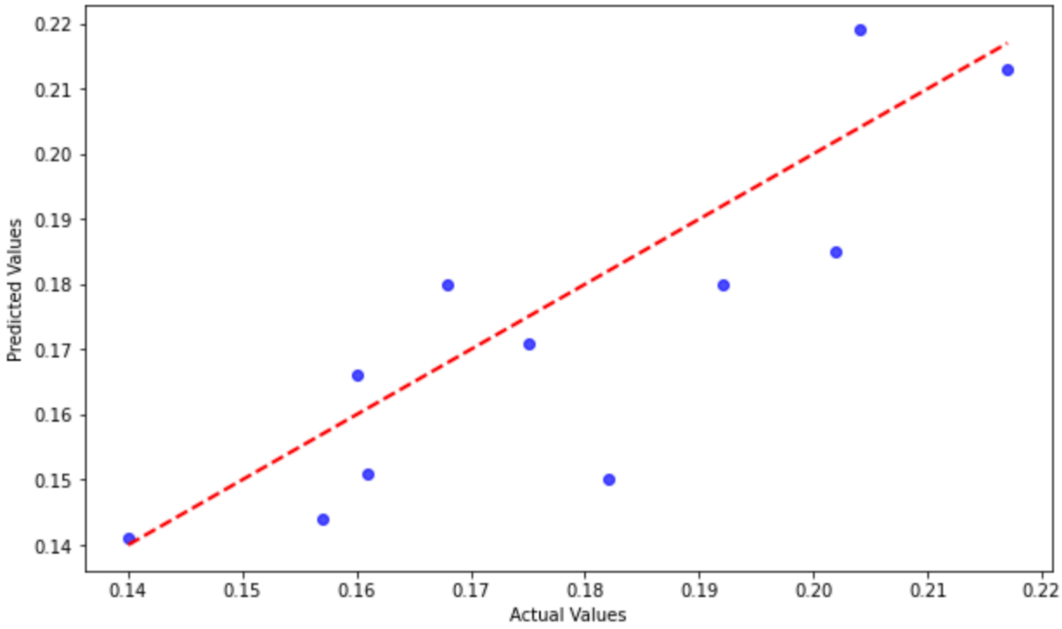
The last model that was implemented was the Decision Tree. The results are pictured below.



```
Mean Squared Error: 0.00019672727272727282
R-squared: 0.6130185979971385
    Actual  Predicted
19   0.217      0.213
41   0.175      0.171
47   0.168      0.180
12   0.140      0.141
43   0.204      0.219
5    0.182      0.150
17   0.160      0.166
50   0.161      0.151
3    0.192      0.180
32   0.202      0.185
13   0.157      0.144
```

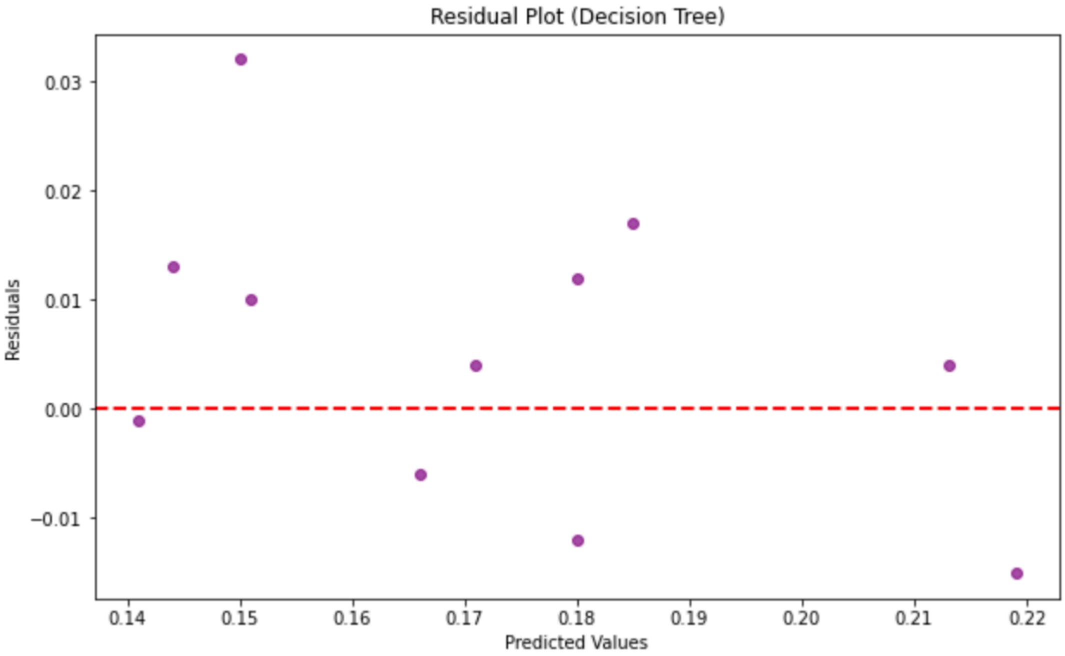
The decision tree model explains about 61.3% of the variation in poor health while capturing meaningful relationships between race/ethnicity and health disparities. The low MSE of about .000197 shows that the model has reasonably accurate predictions with small differences. However, there are some noticeable outliers. The predictions demonstrate how well the decision tree captures the relationships between predictor and target variables, but some predictions show there are still areas where the model can be improved. Though all models pointed to the fact that race/ethnicity is a predictor of poor health, the Decision Tree model had the overall worst performance.

Through the same two kinds of graphs, we visualize the results from the Decision Tree model. Directly below is the plot which shows the predicted vs. actual values.



Most of the points are close to the red line, which indicates that predictions are fairly accurate. There are still some deviations. This is expected because no model can be perfect.

Pictured below is the residual plot for the Decision Tree model.



The points are fairly well-distributed around the zero line. They also seem randomly scattered, which indicates that there are no major issues within the model fit.

The scatter plots show a relatively close alignment between actual and predicted values across the three models, with some discrepancies, especially for Decision Trees. The residual plots indicate that the Random Forest and Linear Regression models have smaller residuals, meaning that they have better model fit. The decision tree exhibits larger residuals which shows that its predictions are more variable and less accurate for certain cases.

Within the Decision Tree, there is a high variance observed. Some predictions deviate from actual values. The Linear Regression model had a better overall performance with more accurate predictions than the Decision Tree. The Random Forest model had the best performance in terms of explaining variance in health outcomes. These results indicate that more complex models like Random Forest tend to perform better than simpler Decision Trees.

Public health interventions are needed for Black, Hispanic, and Other underserved communities. Health disparities usually come from limited access to care, socioeconomic challenges, and systemic barriers that affect these groups at a higher rate. We can use data and analysis for targeted outreach programs. We must increase access to more affordable care and tackle the underlying social determinants of health which include income, education, and geography. This way, interventions become effective. They can also ensure that health is promoted within marginalized communities.

A few main conclusions can be drawn from this research. Race & Ethnicity play a key role in health disparities across the U.S. They are shown to be significant factors. The Random Forest Model captured complex, non-linear relationships effectively. Other models may have overlooked these relationships. There are some limitations in the dataset, such as the underrepresentation of smaller racial groups and the lack of variables like income, education, and access to care. In future work, income, education, & healthcare access could be incorporated. Data Insights must be interpreted within broader social & historical contexts to ensure meaningful and impactful solutions. These points highlight the key findings from my analysis.

References

Centers for Disease Control and Prevention. (n.d.). Behavioral Risk Factor Surveillance System (BRFSS). U.S. Department of Health and Human Services. Retrieved from http://www.cdc.gov/brfss/index.html

Kaiser Family Foundation. (n.d.). Adults who report fair or poor health status by race/ethnicity. KFF. Retrieved from https://www.kff.org