CSE 545 - Social Health

Team - EcoStats (Karan Sheth - 114362694, Wing Au - 112495536, Liting Chiang - 112749075)

1 Introduction

Social health is a broad topic that covers life expectancy, income, and life satisfaction are all intricately related. The goal of this project was to investigate the various relationships between the multiple factors that make up social health. To investigate this area, this project explores various data sets related to economic well-being including poverty rates, mortality rates, educational attainment, and social factors (e.g., trust in religion, faith in politics, etc.).

A significant challenge in investigating this topic is the wide breadth of topics covered - building a singular model to capture all facets of social health could either be too simplistic that it obfuscates import relationships between the social health factors, or it could be too complicated and sacrifice interpretibality. As a result, we conducted multiple analyses on different facets of social health using a mix of techniques including clustering analysis and regression techniques to identify relationships between various social healthy factors and evaluate their predictive performance on life satisfaction.

2 Background

Existing research has shown that increased life satisfaction and overall happiness have lead to longer life expectancy. For example, a 2013 German study [1] found that life satisfaction is a powerful risk-factor for later mortality and is more predictive of mortality than a host of other variables. A similar result was produced in a 2020 longitudinal study [2] performed in the US that found life expectancies at age 18 could be nearly 9 years longer for individuals with high life satisfaction.

Understanding the socio-economic conditions that can contribute to higher life satisfaction and happiness can inform better public policy decisions For example, higher education levels have been linked to increased life satisfaction [3]. Other studies have shown that every ten percent annual income rise increases an individual's happiness by a similar percentage irrespective of the amount of money they earn each year , and a 2018 survey by Purdue University that used Gallup World Poll data, found out that with a yearly financial gain of 95000 USD, people can experience satisfaction in life [4].

3 Data

3.1 U.S. Census Bureau Data [5]

Because social health covers a wide variety of topics, our project required the analysis of multiple sources of data. One important source of data was the U.S. Census Bureau[1]. From the census bureau, we collected data that we believed were key pieces of data related to the overall social health including educational attainment, income, poverty rates, and elder care. Combined, the data sets totaled roughly 6.0 GB in size.

3.1.1 Data Pre-processing

The raw data from the U.S. Census Bureau contained a significant amount of data, and the first step was extracting useful features from the numerous fields present in the raw census data. We use PySpark to aggregate the raw data and perform basic calculations, then filter out zip code with less than 1000 people and some invalid zip codes, and finally convert the raw sums to percentages. By focusing on percentages as opposed to raw counts, we can standardize the data across each county to the same scale. We aggregated the data by matching zip codes and averaged the values of the data over the available years. From the raw data we constructed a feature vector for each zip code that included the following information:

- **Descriptive Data**: Zip code
- **Population Characteristics Data**: Median Age
- **Economic Data**: the overall county poverty rate, the 16+ age group labor population proportion living in poverty, the employment / unemployment rate for the 16+ age group living in poverty, and the poverty rate for the 25+ age group
- **Education Data**: the proportion of adults aged 25+ with a bachelor's degree or higher living in poverty, the bachelor's (or higher) attainment rate
- **Elder Population Data**: The percentage of the 75+ age group living with or without difficulty, the old-age dependency ratio (i.e., the ratio between citizens age 65+ and citizens between 18-65)

3.2 CDC [6] - U.S. Mortality Data

To learn more about how a person's social background influences the causes of death, particularly accidents and suicide, we gathered preprocessed data from the CDC on all aspects of a person's death from year 2005 to 2015. The features that were utilized for our analysis includes manner of death, sex, education history, location of death, etc. More than 4.5 GB of data was collected and preprocessed using pyspark for analysis.

[1] https://www.census.gov/

3.3 World Values Survey Data [7]- Life Satisfaction in Different Countries

World Values Survey gather academic data of social, political, economic, religious and cultural values of people in the world. As a result, we gathered data (more than 2 Gb) from all 110 countries questioned from 1995 to 2021 to gain insight into how people's life satisfaction has changed over time in different countries. More than 300 questions are asked to people during this survey. Following the discovery of features related to life satisfaction, roughly 70 features were chosen to predict people's life satisfaction based on social factors. Features based on value of friends and family, trust in religion, faith in government policies, economic and educational background, neighborhood, and others were chosen. Dataset was pre-processed using techniques like label encoding, standard scalar, dealing with imbalanced data features, etc. The dataset on which prediction was done consisted of 396041 rows and 55 columns.

4 Methodology & Evaluation

Our goal is to understand the social conditions of the people at country level in the world as well as county levels in the US. So, different methods were executed to understand social health. In this section, we describe the various approaches used to analyzing the various different sources of data that we encountered.

4.1 Clustering Analysis Based on U.S. Census Data

4.1.1 Clustering Analysis

After constructing our feature vectors, we clustered our data into 5 clusters by calculating the Euclidean distance between the points based on the 12 numerical features . Since there was no data available at the zip code level that tracked life satisfaction, we use poverty as a heuristic for measuring overall happiness and well-being, as studies have shown these two features are linked.

After clustering, we found that 74.3% of the 4000 lowest poverty zip were categorized as High life satisfaction group, while 77.1% of the 4000 highest poverty zip were categorized as Low life satisfaction group, which suggests that our clustering method was able to distinguish groups based on poverty. We assigned each cluster the label 0 through 4, where 0 is the highest life satisfaction group and 4 is the lowest.

As a starting point to investigate which features were linked to poverty, we first calculated the correlation between the county poverty rate and select features that were not based on populations already living in poverty. This was to examine the relationship between poverty and other social factors such as health care and living conditions. Reported correlations are included in the table below:

Feature	Correlation
Poverty Rate	1.000
16+ Unemployment Rate	0.638853
Educational Attainment: Bachelor's or Higher	-0.465730
% Senior Citizen Living with Difficulty	0.133314
% Senior Citizens Living without Difficulty	-0.156783
% Old Age Dependency Ratio	-0.105983

To further investigate these relationships, samples from each cluster that plotted each specific features against the poverty rate of each zip code. Figure 1 shows the plot comparing educational attainment and unemployment rates against select poverty measures. The first figure compares the poverty rate of adults 25 and over against educational attainment, and the second graph plots the overall poverty rate vs. the unemployment rate for 16+ adults. The distribution of the data from each cluster suggests that there is indeed a relationship between educational attainment unemployment with poverty. Furthermore, plotting the unemployment rate among adults living in poverty vs. the educational attainment in those counties also suggests an inverse relationship between these two features. Interestingly, there seemed to exist little relationship between poverty and senior living conditions.

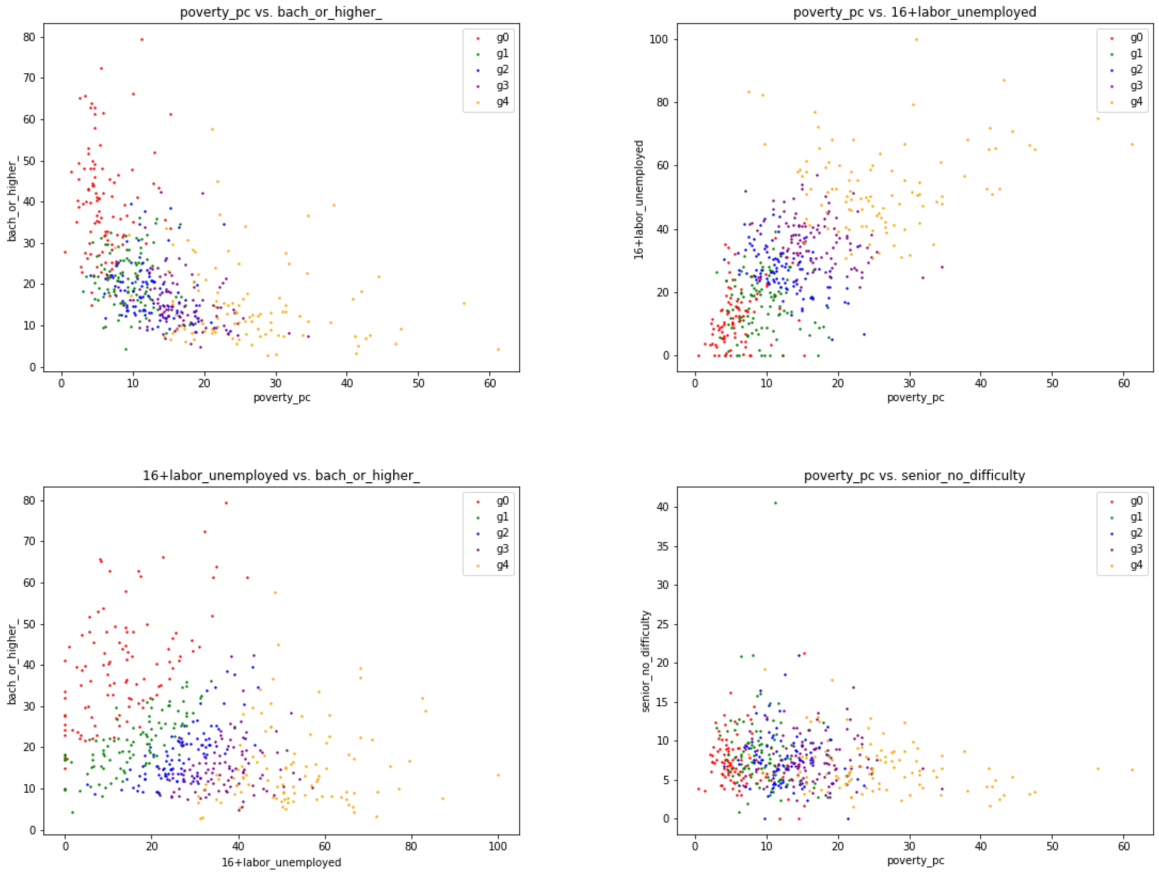

Figure 1: Poverty vs. Educational Attainment; Poverty vs. Unemployment Rate for 16+; Unemployment Rate vs. Educational Attainment; Poverty vs % Senior Citizens Living with No Difficulty

4.2 Understanding Cause of Death in United States

It's critical to comprehend how social background can influence the cause of mortality. The goal is to determine how many people die in accidents or by suicide in the United States. For this, we used pre-processed data from the Centers for Disease Control and Prevention (CDC) on deaths in the United States. Pyspark is used to extract key features from data, integrate data from 2005 to 2015, and do analysis as the data is very large. Plot-1 depicts the gender distribution of causes of death in the United States. 4.9% deaths due to accidents and 1.5% deaths due to suicide in the US is very signifcant. As a result, more research into the backgrounds of persons who died in accidents or by suicide is being conducted in order to determine what may be done to prevent deaths related to this. Plot-2 depicts how suicide instances are distributed according to the person's gender and educational background. Plot-3 depicts how suicide incidents are distributed according to the person's age and educational background. The plot-4 depicts how accident deaths are divided according to the person's age and educational background.

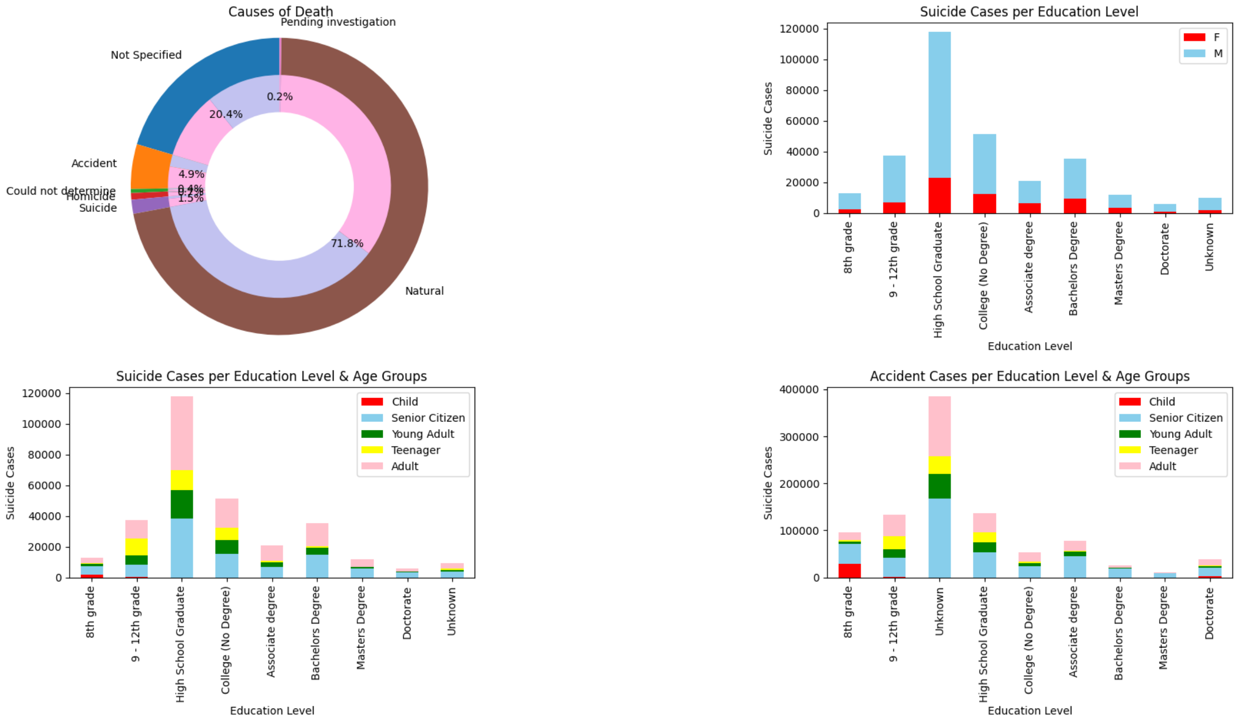

Figure 2: Analyzing the manner of death based on social condition

4.3 Predicting Life Satisfaction of country based on Social Factors

People all throughout the world are asked questions about their family, neighborhood, values, beliefs in political systems, confidence and faith in various procedures, and so on in the World Values Survey. We're figuring out how people's life satisfaction has changed in different countries from 1995 to 2021 based on responses from people all over the world. Figure-3 depicts how life satisfaction among people has changed in various countries (156-China, 458-Malaysia, 643-Russia, 702-Singapore, 756-Switzerland, 840-United States, 862-Venezuela) across years.

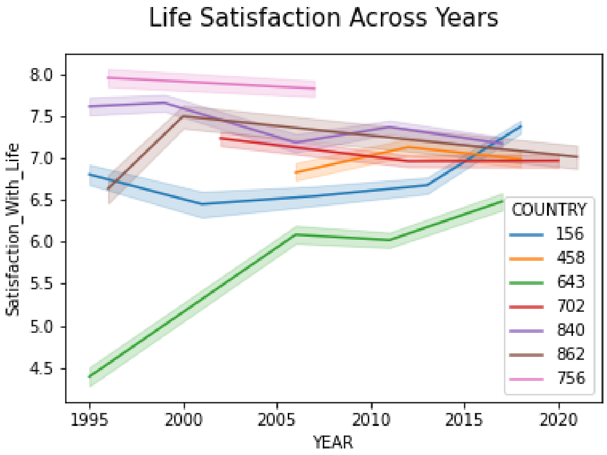

Figure 3: Life Satisfaction Across Years

We're also employing large-scale machine learning to forecast people's life satisfaction depending on their socio-economic circumstances. After doing preprossing of dataset, train-test split of dataset was done. The models were trained on data from 1994 to 2018 and it was tested on the data from 2019 and after. Prediction was done using models like Linear Regression, Logistic Regression, Random Forest, and XG Boost classifier and their accuracy is mentioned in Table-1. Also, hyperparameter tuning was done on Random Forest model using GridSearchCV to get best parameters for the model. (Parameters used - n_estimator=200, criterion="entropy", min_samples_leaf=2, min_samples_split=4)

Models	Linear Regression	Logistic Regression	Random Forest	XG Boost
Accuracy	0.5854	0.6514	0.6774	0.6834

Table 1: Accuracy of Different Models while predicting Life Satisfaction

5 Conclusion & Discussion

Our project investigated the relationships between various facets of social health. The result of our analysis suggests a few main conclusions

- An important factor to reducing overall poverty is expanding educational opportunities, as higher educational attainment was linked to lower poverty and unemployment rates
- The high proportion of suicide cases among those with only high school degrees warrants suggests greater need for supporting economically weaker areas in the United States, in particular adult men
- Tracking citizens' social features can be a useful tool in predicting life satisfaction.

References

[1] C. Guven and R. Saloumidis, "Life satisfaction and longevity: longitudinal evidence from the german socio-economic panel," *German economic review*, vol. 15, no. 4, pp. 453–472, 2014.

[2] H. Lee and G. K. Singh, "Inequalities in life expectancy and all-cause mortality in the united states by levels of happiness and life satisfaction: A longitudinal study," *International Journal of Maternal and Child Health and AIDS*, vol. 9, no. 3, p. 305, 2020.

[3] W. Zhang, K. L. Braun, and Y. Y. Wu, "The educational, racial and gender crossovers in life satisfaction: Findings from the longitudinal health and retirement study," *Archives of gerontology and geriatrics*, vol. 73, pp. 60–68, 2017.

[4] J. M., "29 eye-opening can money buy happiness statistics in 2021," 2022.

[5] "Us census bureau: https://www.census.gov/."

[6] CDC, "Centers for disease control and prevention: https://www.kaggle.com/datasets/cdc/mortality?datase
2016.

[7] I. R., "World values survey: All rounds - country-pooled: https://www.worldvaluessurvey.org/wvscontents.jsp," 2021.