A Plan Beyond Retirement

CS418 - Introduction to Data Science

GITHUB LINK FOR CODE:
https://github.com/kashyap-1234/Retirement

TEAM MEMBERS:

- **Munukutla Durga Venkata Kashyap**
- **Sai Mahesh Vemulapalli**
- **Aneesh Potnis**
- **Aditya Pimpley**
- **Neksha Patel**

I. INTRODUCTION

Retiring in comfort is something a lot of people have as their goal. This generally means saving up enough to sustain a life without worry in the later years. But often it is tough to estimate the many expenses that would come about during this stage of life. A majority of these expenses would be medical expenses. In today's world, life expectancy continues to rise, which results in individuals facing the challenge of ensuring their financials for the post-retirement years. By utilizing the power of machine learning algorithms and predictive modeling, our project aims to make this tough decision making process much easier, by providing the individuals with an estimate of their life expectancy beyond retirement, allowing them to longer for a healthier and more fulfilling retirement. This also helps them in making informed decisions about their finances and their expenditure habits before retirement while preparing for their golden years with a proactive and well-prepared approach. Another benefit of our approach to solving this problem is that the solution can also be used to identify at-risk individuals and provide them with the necessary healthcare.

Over the course of constructing the solutions for this problem, we also decided to explore two branches of this problem, forming two subproblems - the probabilities of shift in the health state in immediately next wave, and the mortality status in the upcoming wave.Our hypothesis at the beginning of the project is that individuals with a history of health issues would have a lower than average life expectancy, individuals from low socioeconomic status would have lower life expectancy than ones from high socioeconomic status, and the gender of a person would have almost no impact on their life expectancy.

II. DATA

Sourced from the Health and Retirement Study (HRS) conducted by the University of Michigan, the dataset is comprised of longitudinal population health data that focuses on transitions between age groups ranging from 50 to 99 years old, income levels, and gender. Quantitative variables (age), categorical/nominal variables (gender), and categorical/ordinal variables (income group) are all included in the data since it consists of 301 rows and 32 columns. The quantitative and discrete representation of health transitions is the most important part of the dataset. It detail movements between states of good health, states of poor health, states of nursing home care, and states of home health care. Within the population that was analyzed, this dataset provides a thorough perspective on the intricate interplay of age, gender, income, and health changes with one another.

III. DATA ANALYSIS

The gathered data has been analyzed through the usage of visualizations to gain a better understanding of the relationships between the different variables that comprise the dataset. Understanding these correlations enables us to build better and more intuitive models that can make better predictions with higher accuracies.

We thoroughly examined the complex relationship between age demographics and the process of transitioning from a state of poor health to the ultimate outcome of death.

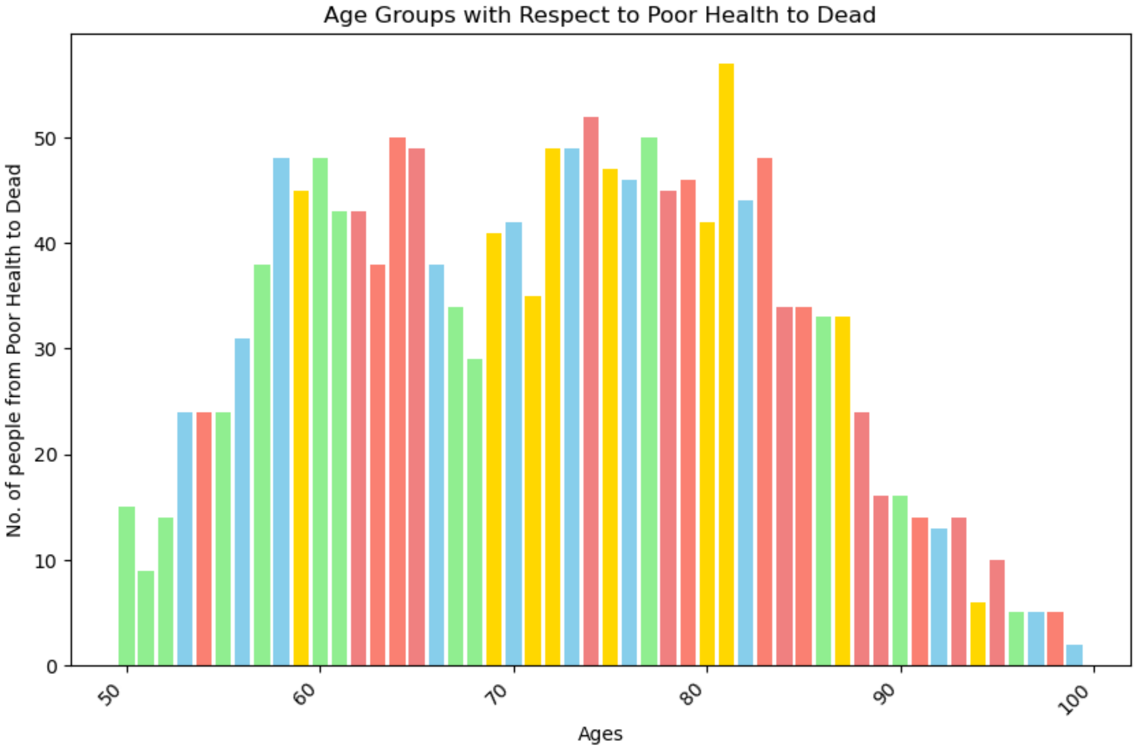

Fig 1: Age Groups with respect to Poor Health to Dead.

The graph provided here visually depicts the intricate relationships between the x and y axes. The x-axis displays a thoughtfully selected range of age groups, providing a thorough description of different stages of life. At the same time, the y-axis accurately measures the significant path of individuals as they move from a state of poor health to the sad end of death at each certain age. As we begin this visual exploration, a clear pattern becomes apparent, revealing a high number of cases occurring between the ages of 50 and 100 years. During the period from 70 to 90 years old, there is a noticeable increase in the number of individuals experiencing this specific health trajectory.

"The graph's narrative surpasses its peak, revealing a gradual decrease in numbers as the age range advances towards the later years, especially within the 90s age group." The progressive decrease of cases eventually reaches a point of complete cessation, which is particularly noticeable at the age of 96, where the graph shows a sharp decline to a total count of zero. This detailed and thought-provoking portrayal vividly illustrates the increasing likelihood of shifting from a condition of poor health to death as one ages. The graph serves not only as a statistical representation, but also as a powerful reminder of how certain age groups are more susceptible to the negative effects of declining health. It prompts us to focus on the significant influence that age has on the complex fabric of life and mortality.

Now, let us examine the complexities of a scatter plot that reveals the correlation between different age groups and the progression from a state of strong physical well-being to the unavoidable outcome of death. This graphic provides a clear view of how individuals of different age groups move from a state of good health to the final outcome of death.

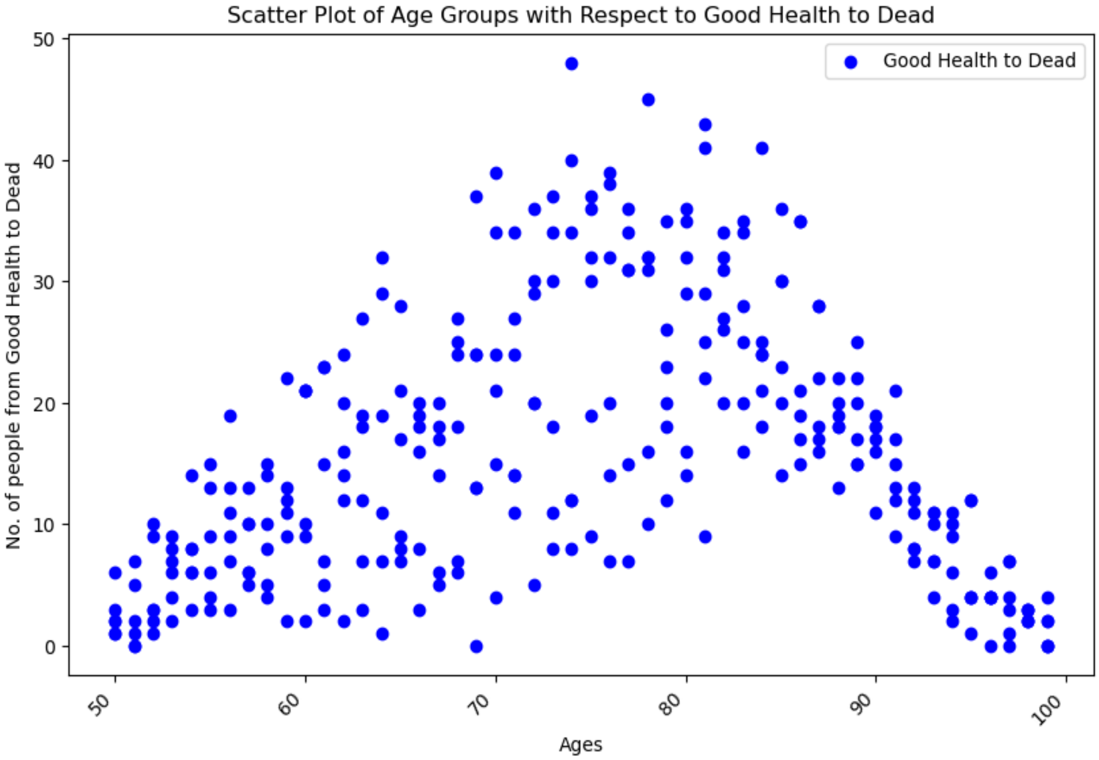

Fig 2: Age Groups with respect to Good Health to Dead.

The scatter plot clearly depicts the relationship between different age groups and the progression from a state of good health to the unavoidable end of death. The x-axis precisely represents a range of age groups, providing insights into different periods of life. Meanwhile, the y-axis accurately measures the number of persons progressing from good health to the solemn outcome of death at each exact age. The graph illustrates a compelling story, showing a significant clustering of instances among individuals aged 50 to 100 years. A notable increase in the trajectory becomes apparent, peaking at the age of 75, which is the stage with the highest number of individuals transiting from good health to dead. After reaching its highest point, the graph noticeably declines until it comes to a complete stop at the age of 99, where the count decreases to zero. This visualization provides a clear understanding of how the prevalence of transitioning from good health to death varies with age. It also highlights the complex trajectory of this health condition at different phases of life. Furthermore, the computed mean age of 65 years serves as a quantitative reference point, providing a unified comprehension of the core trend in this particular situation.

We explored the distribution of health statuses across the data set and aggregated the entries after grouping them by the present health status and the changes in health status in the upcoming wave. This is visualized in the form of a pie chart with branches pointing to the respective health status transition percentages in the next wave.

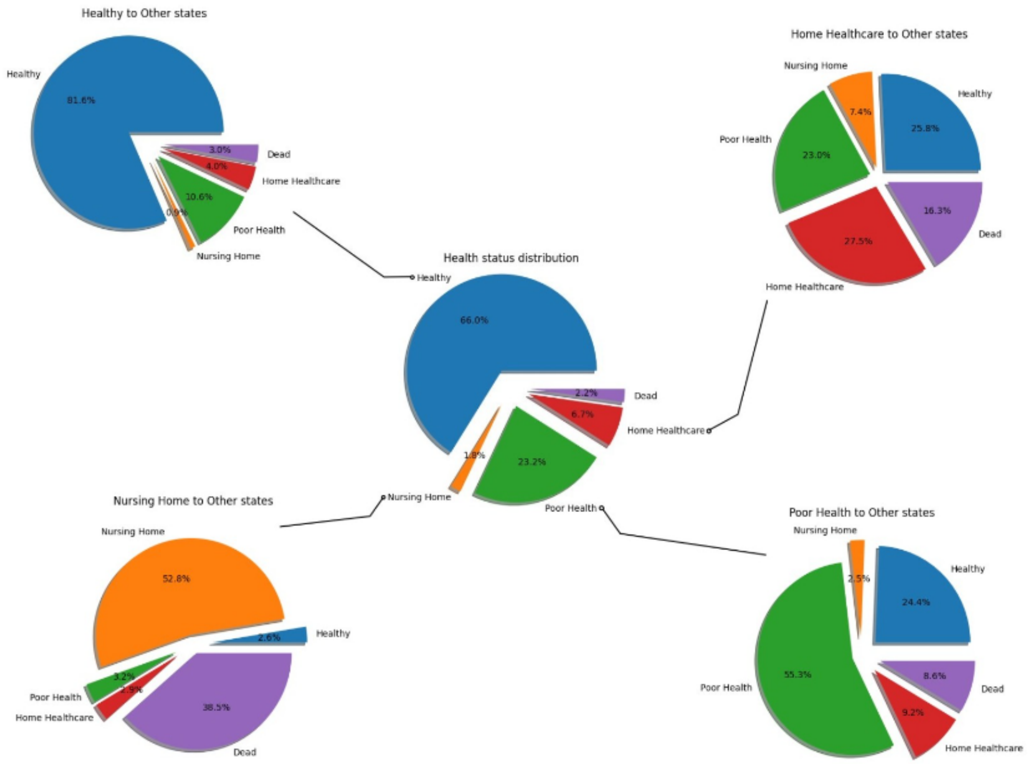

Fig 3: Health Status distribution across the dataset

Immediately observable aspect is that the dataset is dominated by the “Healthy” status with 66% of the entire dataset, followed by “Poor Health” with 23.2%, with “Nursing Home” as the current healthstate having the least number of entries with 1.8% share. This is important to note as it signifies an imbalance in the dataset when constructing models that generate predictions on health states. The hypothesis when constructing this visualization was that the individuals who are healthy tend to have a higher chance of remaining healthy, and the individuals who are in nursing homes would have the highest chance of dying in the next wave. Based on this chart, we can see that the hypothesis was proven correct, with healthy individuals tending to stay healthy with 81.6% chance, and individuals receiving care in a nursing home having an unfortunately high 38.5% chance of dying in the next 2 years.

Another aspect of the data crucial to the models' performance is the distribution over the ages of the individuals. Below is a visual representation of this distribution.

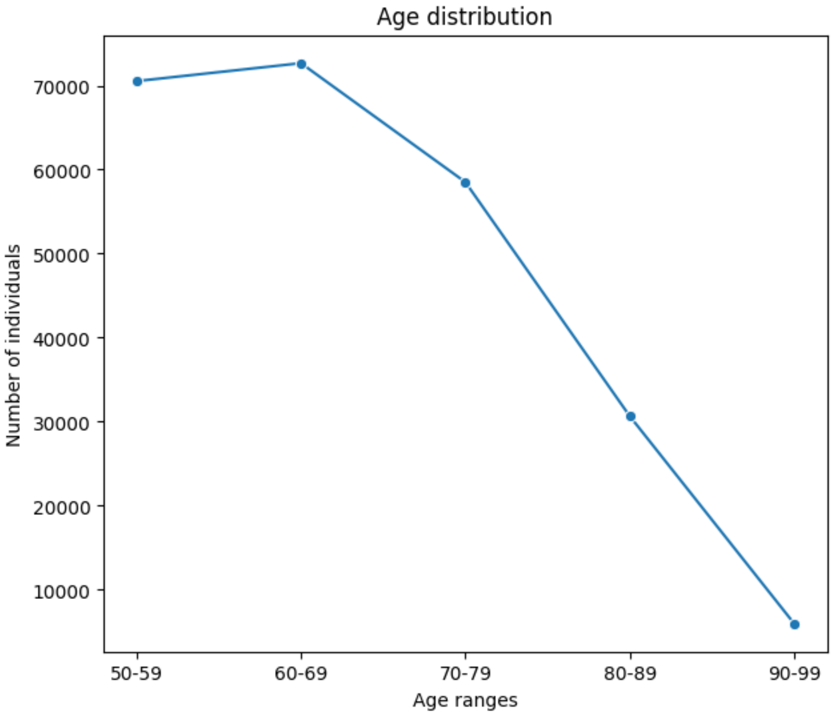

Fig 4: Distribution of data points across Age range.

This visualization highlights another imbalance in the dataset - early ages, between 50 and 70, have a lot more entries in the dataset compared to the later ages. This is an expected observation that is analogous to the distribution of population in the real world - fewer people survive till the 90s due to a vast number of reasons which are outside the scope of this project. So while the models trained on this data would have a bias over understanding the data points from the early ages, this is a favorable outcome for the problem's context, as our goal is to make the model reflect the real world results as closely as possible.

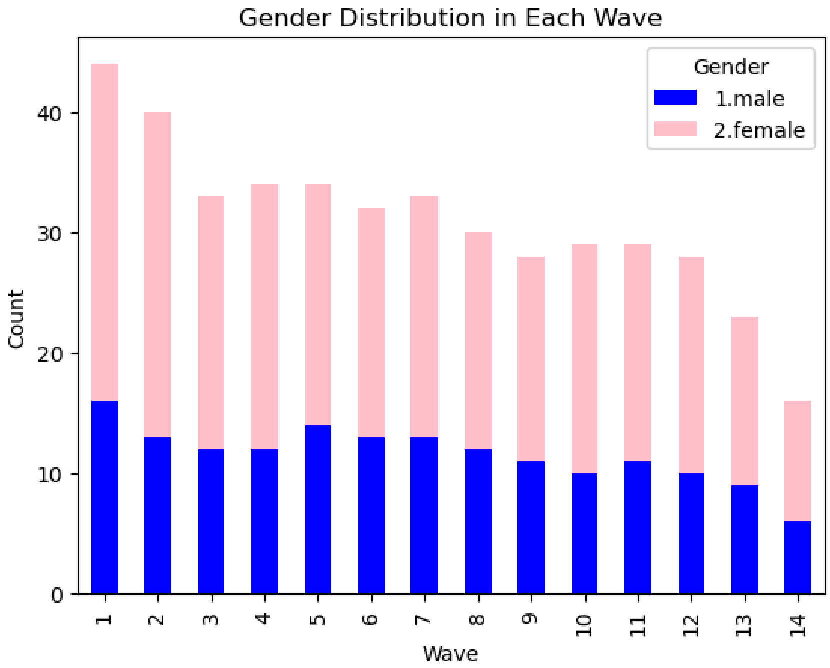

Fig 5: Distribution of gender across waves.

This is a stacked bar chart showing the gender distribution in each wave. There appears to be a noticeable gender imbalance in each wave. The pink segments (representing females) are consistently larger than the blue segments (representing males) in each wave. This suggests that there are more female participants than male participants across all waves. Over the progression of waves, there is a visible decrease in the overall count of both male and female participants. The total height of the stacked bars decreases as we move from left to right, indicating a decrease in the total number of participants in each wave. This temporal trend may suggest a declining participation rate or a decrease in the number of observations over time.

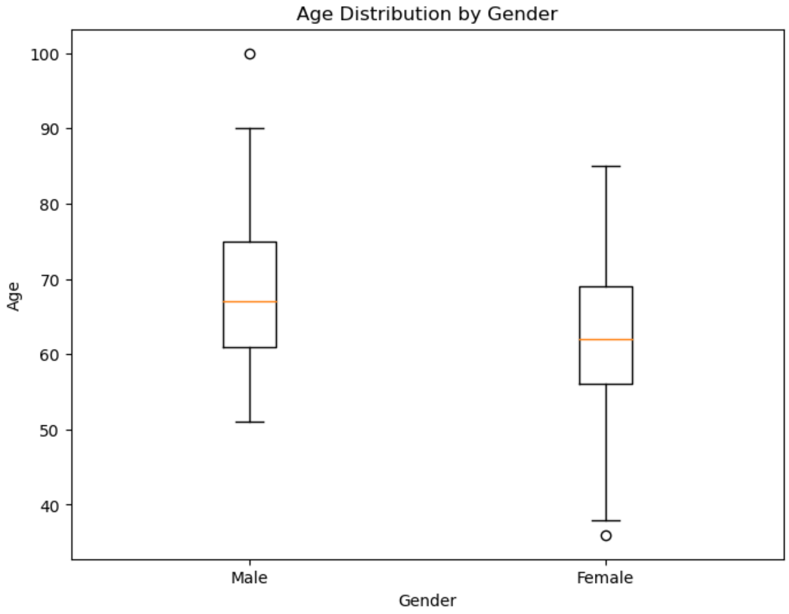

Fig 6: Distribution of age across gender.

The box plot illustrates the distribution of ages for males and females in the dataset, providing insights into the central tendency and variability within each gender. The plot compares the age distributions for males and females, with the box representing the interquartile range (IQR) and the line inside the box indicating the median age. The median age for males is 67 and females is 61. The plot reveals the presence of potential outliers in the dataset. There are males with ages exceeding 100 and females below the age of 30. The whiskers of the box plot extend to the minimum and maximum values within a certain range (usually 1.5 times the IQR). Points beyond this range are considered potential outliers. The identification of potential outliers is important, as they might have a significant impact on predictions and accuracy, especially in models that are sensitive to extreme values. Outliers could influence statistical analyses and models, potentially skewing results or leading to biased predictions, particularly in the context of predicting life expectancies.

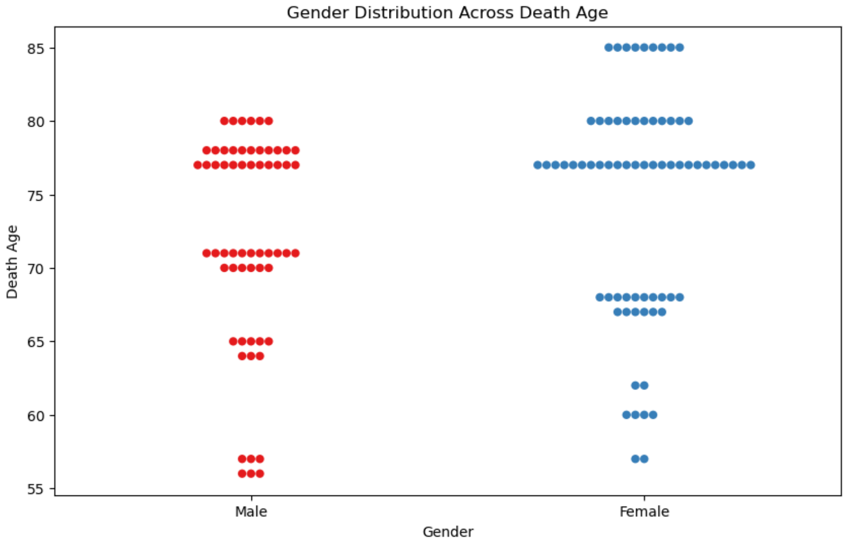

Fig 7: Distribution of gender across Death Age.

The swarm plot visualizes the distribution of death ages across genders in the surveyed individuals. The plot suggests that, on average, females tend to have higher death ages compared to males. The majority of data points for females are positioned at higher ages on the y-axis, indicating a general trend towards longer life expectancy for females in the surveyed population.The observation aligns with the well-known fact that, on average, females tend to live longer than males. This is a common demographic trend seen in many populations.The swarm plot indicates that the maximum occurrence of deaths for both females and males was observed around the age of 77. Interestingly, for females, the plot suggests a broader distribution of death ages, with some individuals reaching ages around 90 years. This implies that while the most common age of death is around 77, there is notable variability, and some females in the surveyed population lived into their nineties.

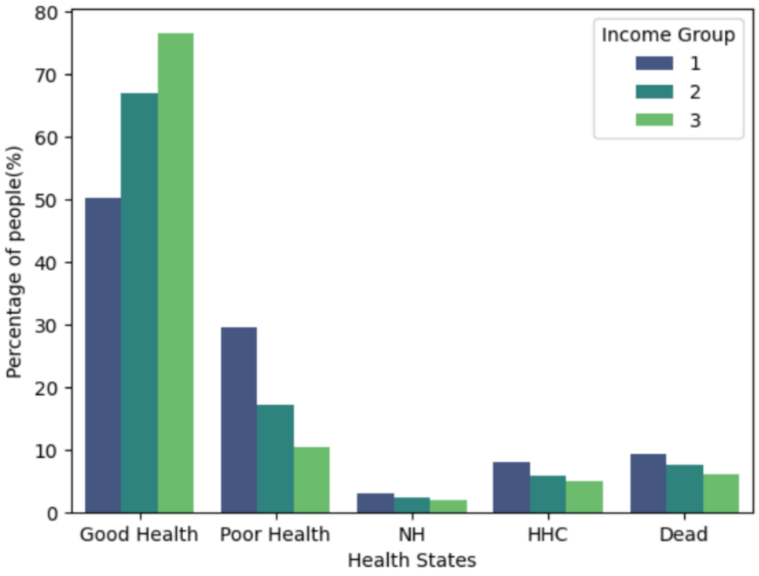

Fig 8: Distribution of Health States across Income Groups.

The above graph plots the percentage of people in each health state based on their income groups.In X axis we can see the health states being mapped whereas Y axis maps the percentage of people. Income group was chosen to be the hue attribute. We can see that good health is dominated by income groups 1 and 2. As we go down the health state order, income group 1 comes out as the one with most people. We can hypothesize by looking at this graph that people with a higher income group end up spending a majority of their life in a good health state and enjoy a better quality of life than those in lower income groups.

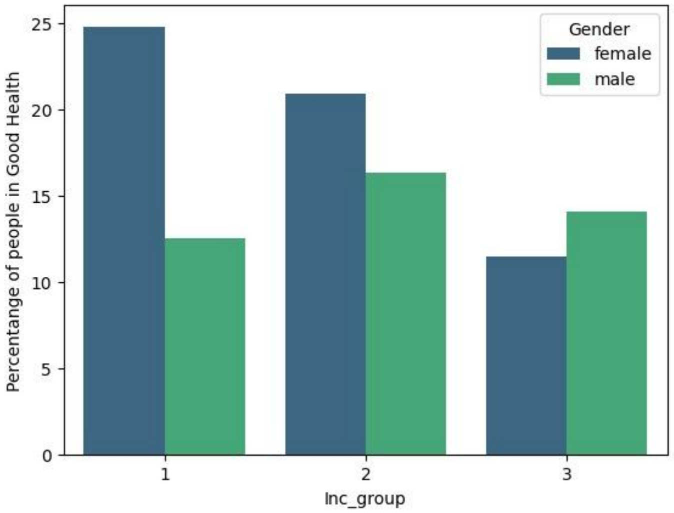

Fig 9: Distribution of Income Groups across Genders.

The above graph showcases the gender split among the various income groups.The Y axis shows the percentage of total people surveyed by gender. The X axis is the income groups. The bar plots have been split by assigning gender as the hue attribute. We can see that a majority of people in income group 2 and 1 are women but this is flipped in income group 3. From our hypothesis in previous visualization you would assume that since there are more men in income group 3, men would live longer than women but we see that it is not the case. Hence we need to also analyze and build models rather than only drawing conclusions from visualizations.

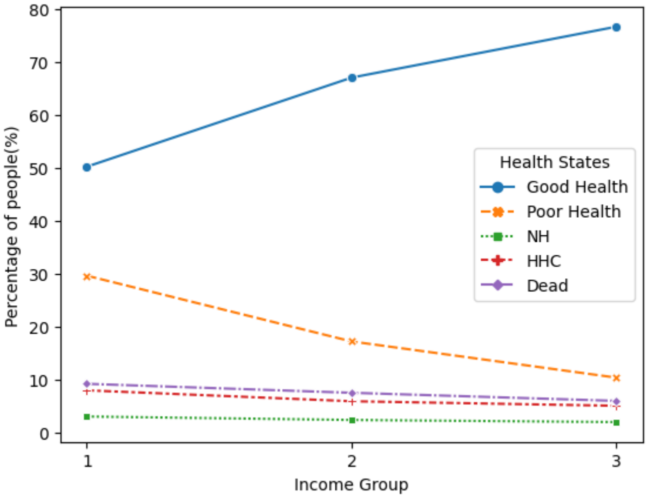

Fig 10: Percentage of People in each income group for a Health State

The visualization in Fig 10 considers the trends in the 5 health states for the 3 income groups over the duration of the study. The hypothesis is that the higher the income group the longer people tend to be in good healthstate. Similarly, the percentage of people dying for every income group tends to decrease as we go up income groups.

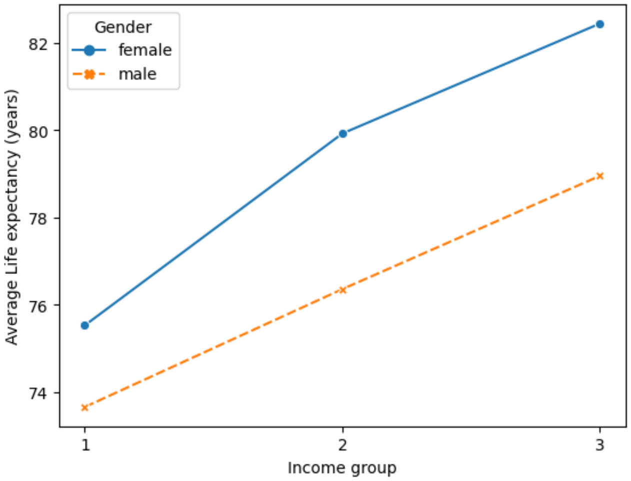

Fig 11: Average Life Expectancy by Gender across Income groups

The visualization in Fig 11 maps the average life expectancy of males and females belonging to each income group. The hypothesis is women always tend to live longer than men in their income group.

Curiously, but for women in income group 2 they live longer than men in income group 3. So another hypothesis could be that if you are above a certain income group then women live longer than men no matter the income and the higher the income for women they live longer still.

IV. MACHINE LEARNING MODELS

We trained 7 models using various machine learning algorithms and methods to solve the task of predicting a person’s life expectancy. For the two subproblems, where we predict the health status of the individual in the next wave and the mortality of the individual in the next wave, we used a classification approach. For the health status prediction, we devised a multi class classification approach, and for the mortality prediction, we used a binary classification approach. The baseline for these is set as selecting the most frequent class in the dataset in both cases, resulting in accuracies of 36% and 50% respectively.

Logistic Regression

Logistic regression has been implemented as the first model.Its task is to predict whether an individual will survive the current wave or not. The data was cleaned, label encoded and downsampled in order to make both the dead and alive classes balanced. Since the data was perfectly balanced, a predictor predicting everything as dead or alive was chosen as the baseline with a 50% accuracy.

The model was initially trained with gender,current income group and age but the accuracy wasn’t impressive. Additional data like Health states and income groups in previous waves was added. The data was then encoded from strings to ordinal integers. A boost in accuracy was seen after retraining the model and an accuracy of 82% was recorded which was a 64% improvement over the baseline.

Following are the graphs visualized from the model:

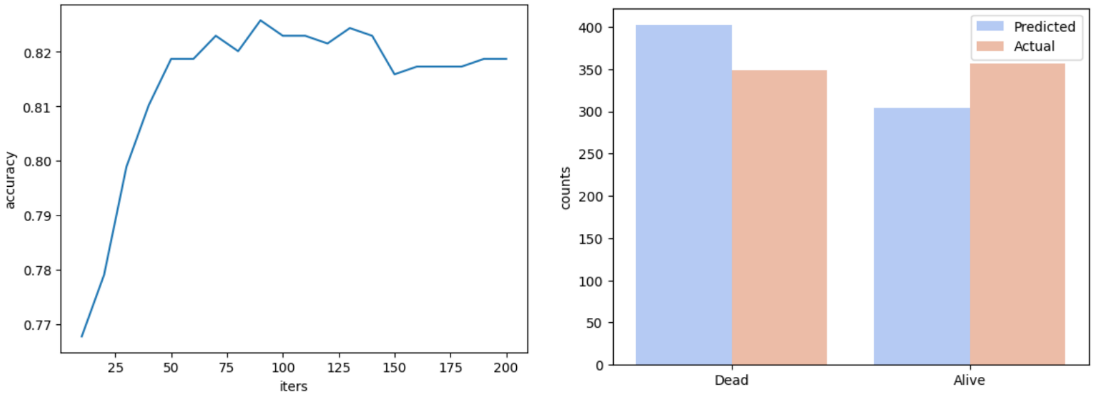

Fig 12: Results of Logistic Regression - Accuracy and Predicted

The first graph plots the accuracy graph of the model with y axis displaying the accuracy and X axis displaying the number of iterations. It essentially is trying to descent the gradient and fit with the perfect weights to obtain the highest accuracy. The second graph showcases the comparison between the counts of actual and predicted labels. X axis and Y axis display status and counts respectively.

Support Vector Machine

When it comes to forecasting an individual's future over the next two years, the Support Vector Machine (SVM) stands out as a powerful predictive model. Compared to a baseline accuracy of 50%, the SVM greatly surpasses it by obtaining an outstanding accuracy of 72%. The model's ability to identify patterns related to life or mortality is demonstrated by a significant 44% enhancement compared to the baseline.

In order to further explore the prediction powers of the SVM, a thorough analysis was performed, graphing the actual number of deaths against the expected number of deaths in the test dataset. The x-axis of this graph represents discrete income categories, classified as 1, 2, and 3. The y-axis, which ranges from 0 to 1000, represents both the predicted and actual number of deaths. Upon further analysis of the graph, it becomes evident that the SVM's predictions outperform the actual values in income groups 1 and 2, providing interesting insights. Nevertheless, the model faces difficulties in precisely capturing trends within income group 3, resulting in deviations between its prediction performance and the actual results.

These observations emphasize the model's effectiveness in specific demographic circumstances and identify areas that should be improved, especially with the intricacies of income group 3. The visual depiction presented is a useful tool for understanding the SVM's prediction complexities and identifying specific areas for improvement to increase accuracy across different income groups.

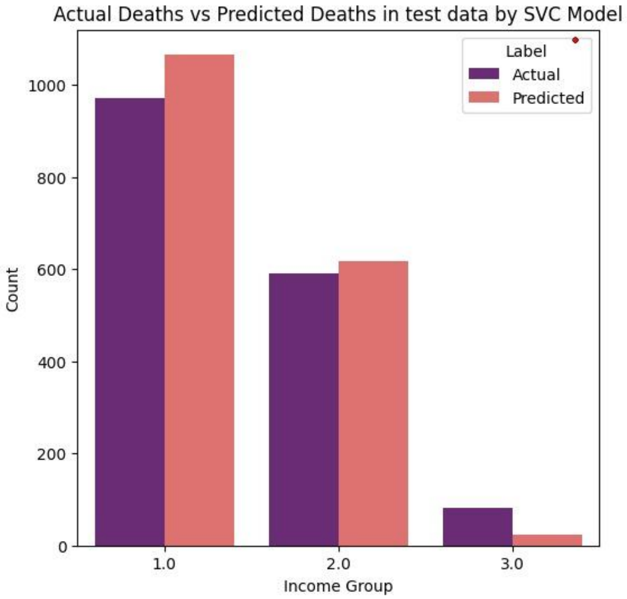

Fig 13: Actual Deaths vs Predicted Deaths in the test data by SVC Model

Random Forest

Another model used for this part of the project is a Random Forest model. Random Forest is an ensemble method that operates by constructing multiple decision trees and taking the output by majority voting. In our case, Random Forest performed better than the baseline with an accuracy of 48% on multi class classification which is a 33.3% improvement over the baseline, and an accuracy of 72% which is a 44% improvement over the baseline. Due to the imbalanced nature of the dataset in terms of health state distribution, the model in multi class classification problem favors learning the patterns of the data points belonging to the health state “Healthy”, as doing so yields the highest accuracy boost. Conversely, for the health states “Dead” and “HC and NH” (Home care and Nursing home), the model failed to grasp the patterns in identifying the data points belonging to these states. One anomaly is the higher number of predictions in the “NH” (Nursing Home) state than the real labels, but this could be attributed to the random sampling nature of the test data we employed, and could be explained through further analysis of the model’s performance. These observations are visualized in fig 14 below.

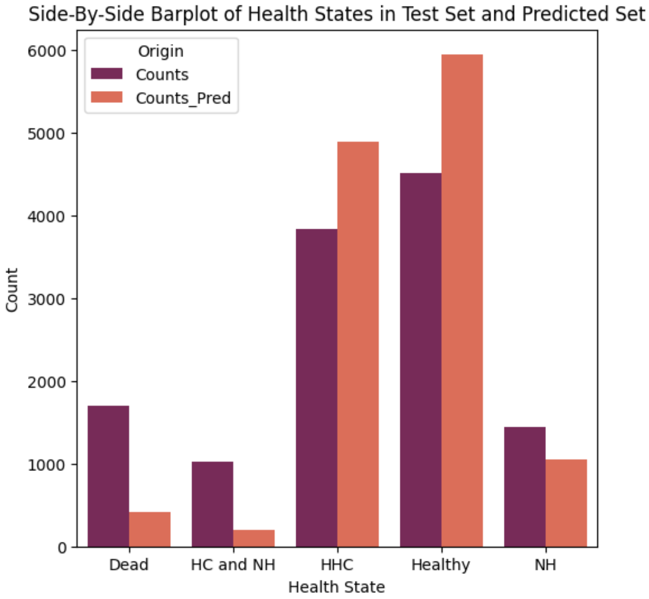

Fig 14: Actual State vs Predicted State by Random Forest Model

Multinomial Logit Regression

One of the ML models used for this part of the project is multinomial logit regression model. Multinomial logistic regression is employed to model the complex relationships between age, gender, income group, and previous health states to predict the probabilities of transitioning between different health states. This type of modeling is particularly useful for understanding how various factors contribute to transitions between discrete categories, such as different health states in this scenario. The resulting transition probabilities can provide valuable insights into the dynamics of health changes within a population over different ages and across various demographic groups. The transition probabilities derived from the multinomial logistic regression

model play a crucial role in predicting the life expectancies of individuals based on their current health states. Understanding these transition probabilities allows for the estimation of how likely an individual is to move from one health state to another over time.
For instance, the probability of transitioning from a healthy state to a dead state is quite low of approx. 0.00266 for an individual at age 65. The dataset was split into training and testing subsets, with 70% of the data randomly assigned to training and the remaining 30% to testing. A multinomial logistic regression model was then fitted using the training data, considering features such as age, gender, income group, and previous health states to predict the current health state. Subsequently, transition probabilities were calculated for different health states at various ages using the test data. To establish a baseline, a dummy classifier was employed, predicting the majority class. The baseline accuracy, determined to be 80%, was compared to the logistic regression model's accuracy of 79%. The confusion matrix provided a detailed breakdown of model predictions, and the log-likelihood served as a measure of the model's predictive performance. Overall, this comprehensive evaluation helped assess the model's efficacy in predicting health state transitions based on individual characteristics.

Linear Regression

Another approach to the problem that we considered was predicting the life expectancy of the person based on the features of age at the start of the study, income group, etc. and predict the age of death from it.
For implementing this, a linear regression model was selected. In the dataset not everyone had taken part in the study from the first wave and hence there were a number of missing values that needed to be imputed in the dataset. Next, one hot encoding was used in the gender column as it is a nominal variable.
The dataset was then split into train and test sets to train the linear regression model. K-fold cross validation was used to create a validation set from the training data and check the accuracy of the model on the training set.
The baseline for the linear regression model is set as the average age of death across all genders and income groups. The baseline had a mean squared error of 123.88 when test data was used to predict the life expectancy.

The linear regression model got a mean squared error of 41.99, which is still high but a 66.15% improvement over the baseline model. The linear regression model also got a r2 score of 0.66 .
We found that predicting a person's age of death would involve numerous factors which could not be accounted for in the dataset. Hence, we decided to predict the life expectancy of a person in a particular age group. In this way we moved from a linear regression problem to a multi class classification problem where to predict life expectancy was much more feasible.

Decision Tree

Decision tree used the same data pre-processing as linear regression but the ages of death were grouped together in age groups as suggested previously. The 4 age groups were less than 65, 65 to 74, 75 to 84 and 85 plus.

The decision tree classifier was then trained on the data after splitting into training and test sets. For the baseline, the age group with the highest distribution of deaths (mode) was selected.
The decision tree classifier was trained with a varying max depth hyperparameter to find out which depth would give the maximum accuracy.
The decision tree classifier with the criterion hyperparameter set as entropy was able to give an accuracy of 56% as compared to the 35% accuracy from the baseline, a 60% improvement.

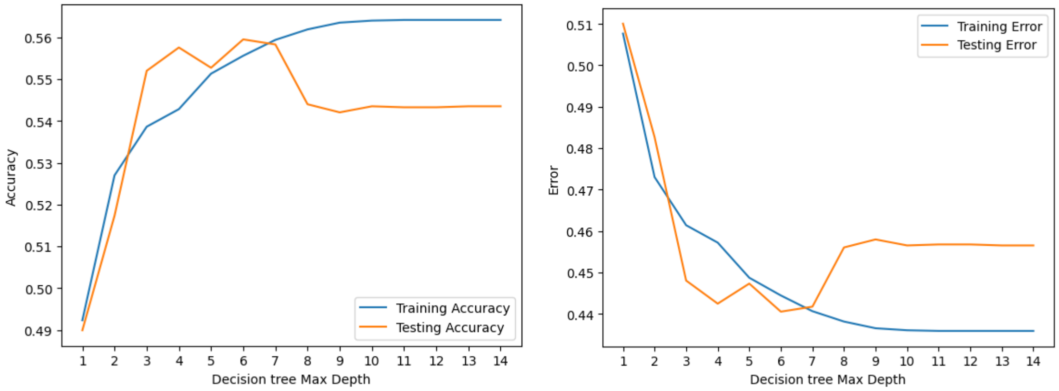

Fig 15: Training vs Testing accuracy and error for decision tree.

From the plot for accuracy we can see that the decision tree starts overfitting to the training data after depth 7 and highest accuracy we received was for depth 5.

Neural Network

A neural network for performing multi class classification was also implemented. The neural network was trained on the pre-processed data. For multi class classification the output label needs to be in 4 classes for the 4 age groups, for doing this one hot encoding was used the age at death column.
A sequential model where the layers are defined one after the other was created to define the neural network. The neural network is made up of the input layer, 2 hidden layers and an output layer. The first hidden layer has 256 nodes and the second layer has 32 nodes with a ReLU activation function. The final output layer has a softmax activation function with 4 output nodes for the 4 classes.

The baseline of the neural network was created with only 1 hidden layer of 32 nodes which gave an accuracy of 46.72%. While the neural network with 2 hidden layers was able to give us an accuracy of 55.17% an 18% improvement over the baseline.

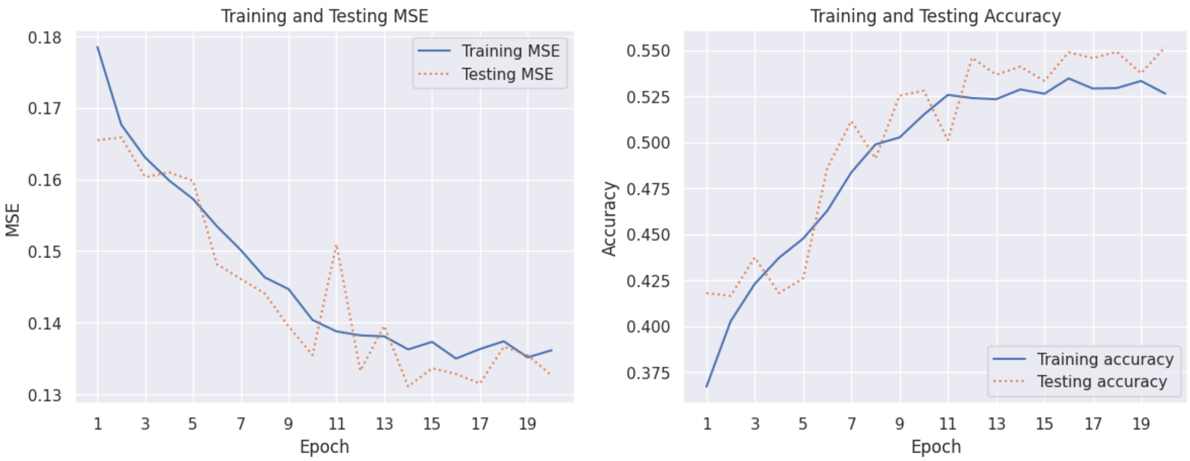

Fig 16: Training vs Testing accuracy and error for Neural Network.