Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 10, No. 3, 2026
www.scholink.org/ojs/index.php/asir

Original Paper

The Social Side of Health: Using Machine Learning to Predict Mortality

Jaiden Chen
The Bishop's School, La Jolla, California, CA 92037, United States

Received: August 3, 2026 Accepted: August 28, 2026 Online Published: September 4, 2026
doi:10.22158/asir.v10n3p31 URL: http://dx.doi.org/10.22158/asir.v10n3p31

Abstract

Why do some people live longer than others? Age and disease obviously play a major role, but factors such as smoking, income, BMI, and diabetes may also be connected to mortality. For this project, I used data from the 2013–2014 National Health and Nutrition Examination Survey (NHANES) linked with CDC mortality records. I focused on six variables: age, gender, income, smoking status, BMI, and diabetes.

After merging and cleaning the datasets, I had 6,100 participants with known mortality outcomes. I trained two machine learning models, Random Forest and Logistic Regression, using an 80/20 training and testing split. Random Forest achieved 92.9% accuracy and an AUC of 0.889, while Logistic Regression achieved 92.2% accuracy and an AUC of 0.856.

Age was the strongest predictor in the Random Forest model, followed by BMI and income. Logistic Regression also showed relationships involving smoking, diabetes, income, and age. Overall, the project showed me that mortality is not connected to just one factor. Both health and social conditions can provide useful information when predicting mortality.

Keywords

mortality, machine learning, artificial intelligence, NHANES, Random Forest, Logistic Regression, public health

1. Introduction

Why do some people live longer than others? Before starting this project, I mostly thought of lifespan as something affected by age, genetics, and disease. However, health can also be connected to factors such as income, smoking, diet, environment, and access to healthcare. Previous research has found major differences in life expectancy across income levels in the United States, showing that longevity is not only a biological issue (Chetty et al., 2016).

This made me interested in whether artificial intelligence could help find these patterns. Machine learning is increasingly being used in healthcare because it can analyze large amounts of information and find relationships between variables that may not be immediately obvious (Yu et al., 2018). Instead of focusing on one specific disease like I did in my previous heart-disease project, I wanted to study mortality more broadly.

For this project, I used the National Health and Nutrition Examination Survey (NHANES), which includes health and demographic information collected from people across the United States. I linked the survey data to mortality records and focused on age, gender, income, smoking status, BMI, and diabetes. I chose these variables because I wanted to include both health factors and social conditions.

I then built two machine learning models, Random Forest and Logistic Regression. The goal was to see whether the models could predict mortality and which variables contributed most to those predictions. This led to the main research question: What can machine learning tell us about who lives longer?

2. Methods

2.1 Dataset

I used data from the 2013–2014 National Health and Nutrition Examination Survey. NHANES is run by the National Center for Health Statistics and collects health information through surveys, physical examinations, and laboratory testing.

I combined several NHANES files. The demographic dataset provided age, gender, and income. The smoking questionnaire provided smoking history, the diabetes questionnaire provided diabetes status, and the body-measurement dataset provided BMI. I merged the datasets using SEQN, which is the identification number assigned to each NHANES participant.

Next, I linked these records with the CDC mortality file. In the final dataset, a mortality status of 0 represented someone who was alive, and 1 represented someone who had died during the follow-up period.

After removing participants without a valid mortality outcome, the final dataset contained 6,100 participants. Of these, 5,633 were alive and 467 were deceased.

2.2 Variables and Data Cleaning

The six variables used to predict mortality were age, gender, income, smoking status, BMI, and diabetes status. I selected them because together they cover demographic traits, behavior, physical health, and socioeconomic status.

Before training the models, I cleaned the data. Smoking and diabetes responses were converted into values of 1 and 0, and gender was also converted into numerical values. Participants without mortality information were removed.

Some of the remaining variables had missing values. Missing income and BMI values were replaced with the median. Missing smoking and diabetes responses were coded as zero for this model, and missing gender values were filled using the most common value. This kept more participants in the dataset. Still, it may have introduced some bias.

2.3 Exploring the Data

Before building the models, I checked a few basic patterns in the dataset. The clearest difference was age. Participants who were alive had an average age of about 45.6 years, while participants who died had an average age of about 68.8 years.

I also created a correlation heatmap to examine how the variables were related to each other and to mortality. Correlation does not show causation, but it gave a useful first look at which patterns might later appear in the models.

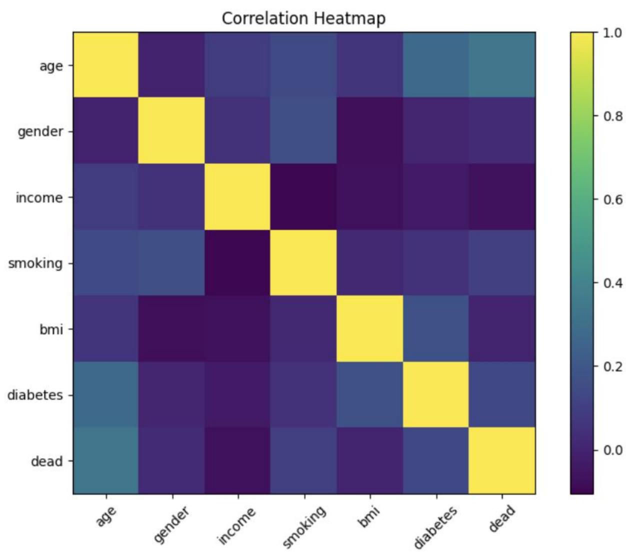

Figure 1. Correlation Heatmap of Health, Demographic, and Mortality Variables

2.4 Training the Models

I divided the dataset into 80% training data and 20% testing data. The split was stratified so that the proportion of deceased participants remained similar in both groups.

The first model was a Random Forest Classifier. Random Forest combines many decision trees to make a final prediction. I used 200 trees with a maximum depth of eight. I picked this model because it can capture more complex relationships between variables and also provides feature-importance values.

The second model was Logistic Regression. Logistic Regression is commonly used when there are two possible outcomes, such as alive or deceased (Schober & Vetter, 2021). Unlike Random Forest, its coefficients also show whether a variable is associated with an increase or decrease in the predicted outcome.

I evaluated both models using accuracy, confusion matrices, ROC curves, and AUC scores. I used several measurements because looking only at accuracy can sometimes give a misleading picture of how well a model actually works.

3. Results

Both models performed well overall. The Random Forest model achieved **92.9% accuracy** with an **AUC of 0.889**, while Logistic Regression achieved **92.2% accuracy** with an **AUC of 0.856**. Random Forest performed slightly better on both measurements.

Random Forest Results

The Random Forest correctly predicted 1,111 participants who survived and 22 participants who died. It incorrectly predicted 16 surviving participants as deceased and missed 71 participants who actually died.

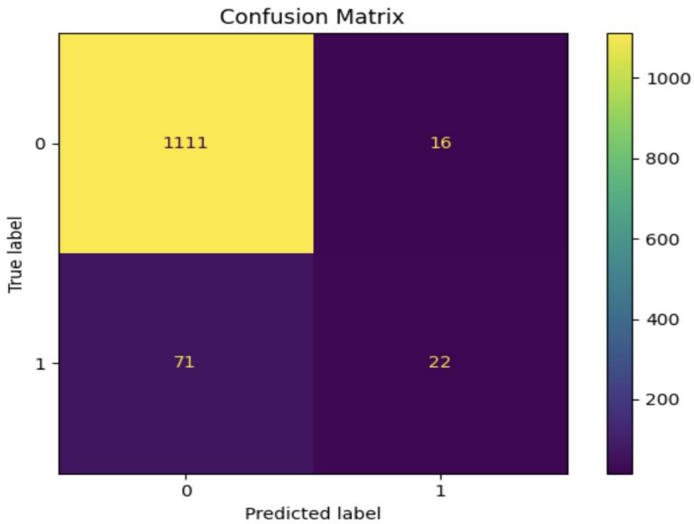

Figure 2. Random Forest Confusion Matrix

The model's AUC of 0.889 showed that it was fairly strong at separating participants with lower and higher mortality risk.

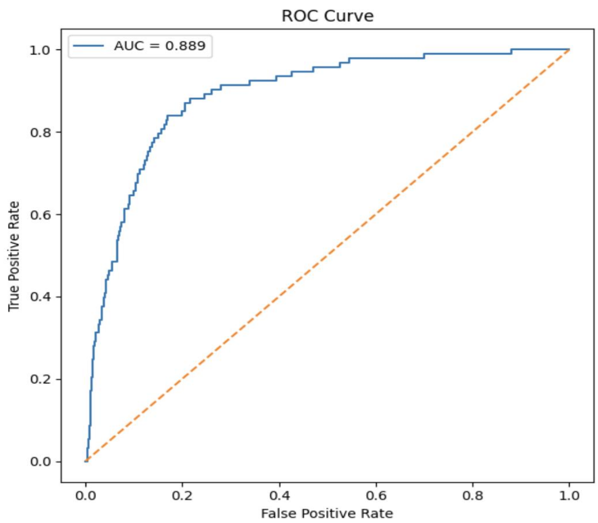

Figure 3. Random Forest ROC Curve

The feature-importance graph was one of the most interesting results. Age was by far the strongest predictor at approximately **50.8%** of total feature importance. BMI was second at **21.4%**, followed by income at **18.6%**. Diabetes, smoking, and gender had smaller feature-importance values.

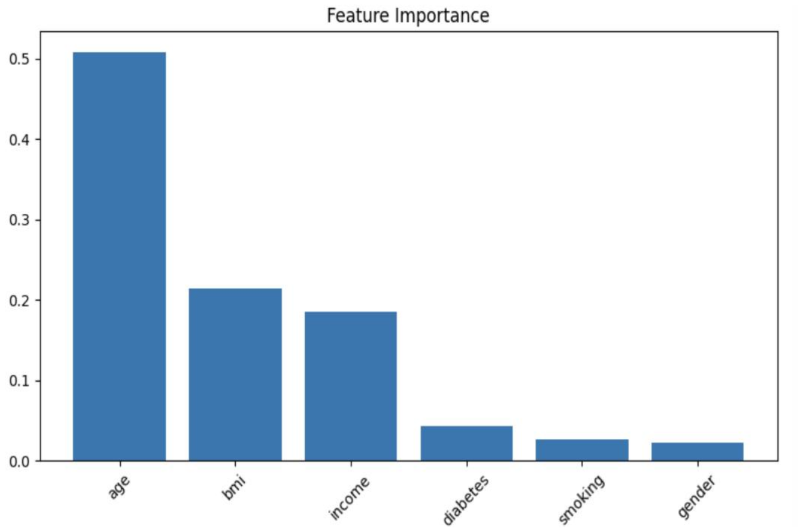

Figure 4. Random Forest Feature Importance

The importance of age was not surprising after seeing the average ages of the two groups. However, I was more interested in seeing BMI and especially income rank so highly. Income is not a medical measurement, yet it still provided useful information for predicting mortality.

Logistic Regression Results

Logistic Regression achieved 92.2% accuracy and an AUC of 0.856.

The model correctly predicted 1,122 surviving participants but only three deceased participants. It missed 90 participants who actually died.

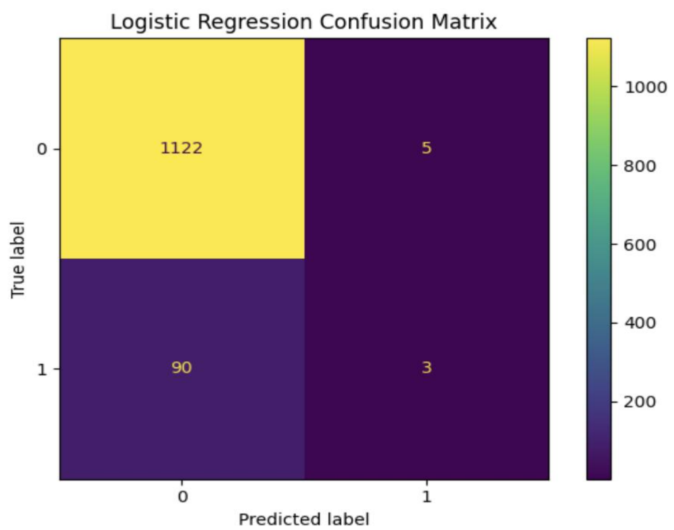

Figure 5. Logistic Regression Confusion Matrix

Even though the accuracy was high, the confusion matrix showed that the model had difficulty identifying the smaller group of participants who died.

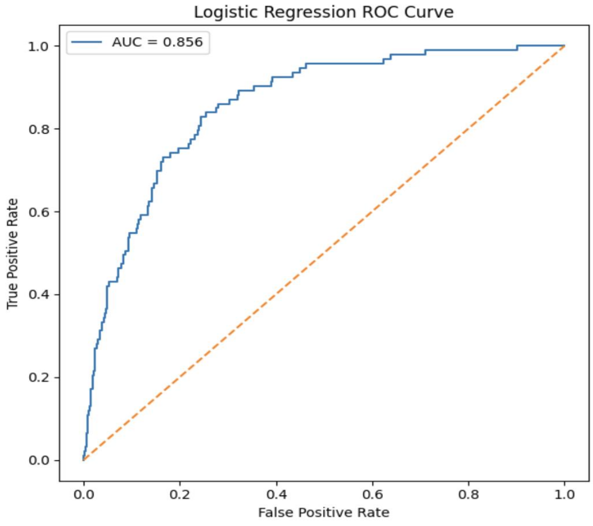

Figure 6. Logistic Regression ROC Curve

The coefficients provided a different way of looking at the variables. Smoking and diabetes had positive coefficients, meaning that they were associated with higher predicted mortality when the other variables were included. Income had a negative coefficient, meaning higher income was associated with lower predicted mortality. Age also had a positive relationship with mortality.

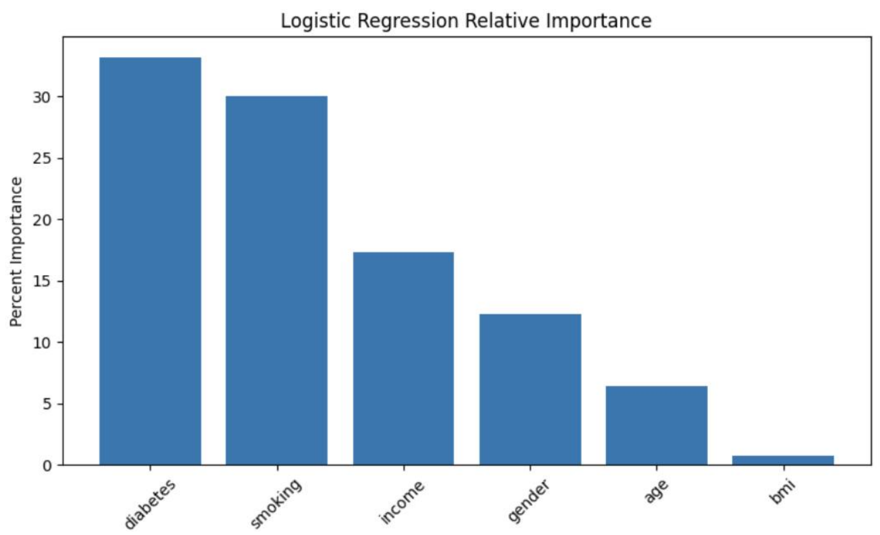

Figure 7. Relative Magnitude of Logistic Regression Coefficients

The two models did not rank every variable in the same way. I expected some differences because Random Forest and Logistic Regression analyze data differently. However, both models showed that mortality could be predicted using a combination of health and demographic information.

4. Discussion

One of the clearest findings was the importance of age. The participants who died were more than 23 years older on average than those who survived, and age accounted for about half of the Random Forest's feature importance. This makes sense because mortality risk generally increases as people get older.

What interested me more was that age was not the only useful factor. BMI and income were also important in the Random Forest model, while smoking and diabetes showed clear positive relationships with mortality in Logistic Regression. Previous research has also connected smoking, diabetes, BMI, and socioeconomic conditions with mortality (Doll et al., 2004; Global BMI Mortality Collaboration, 2016; Wang et al., 2019).

Income was especially interesting to me because it is different from the other variables. BMI and diabetes directly describe someone's health, but income describes part of someone's social situation. Previous research in the United States has found a strong association between income and life expectancy (Chetty et al., 2016). Income may be connected to differences in healthcare access, food, housing, education, stress, and other resources. My model cannot tell me why income was useful, but its importance shows why mortality cannot be viewed only as a biological issue.

Another important result was the difference between accuracy and the confusion matrices. At first, an accuracy above 92% sounded extremely successful. However, only 7.7% of the participants in my dataset were deceased. The Random Forest found only 22 of the 93 deceased participants in the test set, while Logistic Regression found only three.

This changed the way I thought about model accuracy. A model can have a very high accuracy while still failing to identify many of the people we care most about finding. In healthcare, a false negative could mean that a high-risk patient is missed. This is why medical AI should be evaluated using more than one number.

There is also a larger issue with using AI on human health data. Machine learning learns from patterns that already exist in society. If health disparities already exist, an algorithm may learn those disparities as part of its predictions. Researchers have raised concerns about fairness and bias in medical AI for this reason (Chen et al., 2023). AI can identify that income is related to mortality, for example, but it cannot understand by itself why that relationship exists.

This project showed that AI can spot patterns quickly across thousands of people, but human interpretation still matters. In healthcare, knowing why a model produces a result can matter as much as the prediction itself. That is why explainable and trustworthy AI is now a growing area of medical research (Rasheed et al., 2022).

5. Limitations and Future Work

This project had several limits. I used only six variables, even though mortality is influenced by many more factors. Missing data also meant I had to decide how to fill some values, and those choices may have affected the results. The class balance was another issue. Only 7.7% of participants were deceased, which made it harder for both models to learn patterns linked to death.

The models also show associations rather than causation. A variable such as income may rank as important in the model, but that does not mean income itself directly causes someone to live longer. Instead, it may stand in for other differences in people's living conditions, access to care, or general circumstances.

In the future, I would add variables such as blood pressure, cholesterol, physical activity, education, and laboratory measurements. I would also test methods designed to deal with imbalanced data so the models could become better at identifying participants who died. Another possible direction would be using explainable AI to better understand why the model makes an individual prediction.

6. Conclusion

This project started with a simple question: **Who gets to live longer?** Using health and demographic information from NHANES, I built two machine learning models to see whether six variables could help predict mortality.

Random Forest achieved 92.9% accuracy and an AUC of 0.889, while Logistic Regression achieved 92.2% accuracy and an AUC of 0.856. In the Random Forest model, age turned out to be the most important predictor, though BMI and income also played a role. The Logistic Regression results also showed that factors like smoking, diabetes, and age were linked to a higher chance of mortality. Still, one of the main takeaways for me was that accuracy by itself can be misleading. Even though the models seemed accurate overall, they struggled to correctly identify the people who died, partly because deaths made up a much smaller portion of the dataset.

This project also helped me see that mortality is not shaped only by medical conditions. Social factors, such as income, can also carry important information about a person’s health outcome. AI is powerful because it can look through data from thousands of people and detect patterns much faster than I could manually. At the same time, it cannot fully explain the human stories or deeper causes behind those patterns. Instead of giving me a simple answer to the question, “Who gets to live longer?”, this project showed me how complex that question actually is. To me, the value of AI in healthcare is not only that it can predict future outcomes, but that it can help us ask better questions about the many factors that influence health.

References

Centers for Disease Control and Prevention, National Center for Health Statistics. (n.d.). *National Health and Nutrition Examination Survey (NHANES).* U.S. Department of Health and Human Services.

Centers for Disease Control and Prevention, National Center for Health Statistics. (n.d.). *NHANES linked mortality files.* U.S. Department of Health and Human Services.

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*, 719-742. https://doi.org/10.1038/s41551-023-01056-8

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., & Cutler, D. (2016). The association between income and life expectancy in the United States, 2001-2014. *JAMA*, *315*(16), 1750-1766. https://doi.org/10.1001/jama.2016.4226

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ*, *328*(7455), 1519. https://doi.org/10.1136/bmj.38142.554479.AE

Flaxman, A. D., & Vos, T. (2018). Machine learning in population health: Opportunities and threats. *PLOS Medicine*, *15*(11), e1002702. https://doi.org/10.1371/journal.pmed.1002702

Global BMI Mortality Collaboration. (2016). Body-mass index and all-cause mortality: Individual-participant-data meta-analysis of 239 prospective studies in four continents. *The Lancet*, *388*(10046), 776-786. https://doi.org/10.1016/S0140-6736(16)30175-1

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, *149*, 106043. https://doi.org/10.1016/j.compbiomed.2022.106043

Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. *Anesthesia & Analgesia*, *132*(2), 365-366. https://doi.org/10.1213/ANE.0000000000005247

Wang, Y., O'Neil, A., Jiao, Y., Wang, L., Huang, J., Lan, Y., Zhu, Y., & Yu, C. (2019). Sex differences in the association between diabetes and risk of cardiovascular disease, cancer, and all-cause and cause-specific mortality: A systematic review and meta-analysis of 5,162,654 participants. *BMC Medicine*, *17*, 136. https://doi.org/10.1186/s12916-019-1355-0

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*, 719-731. https://doi.org/10.1038/s41551-018-0305-z