**Final Report**

Roshan Chouhan, Yanzheng Liu

**Hypothesis 1:**

To evaluate the hypothesis that individuals with high risk, high BMI, and diabetes are significantly more likely to have experienced a heart attack, using crossing validation, logistic regression and model selection techniques.

Applied 10-fold cross-validation using the caret package to evaluate the model's stability, and the output of the logistic regression performance remained consistent across folds.

Logistic Regression outputs a confusion matrix. The model failed to identify any positive cases of heart attack, despite achieving an overall accuracy of 94.5%. However, this high accuracy is misleading due to severe class imbalance: only 5.5% of the population had a heart attack. The model’s sensitivity was 0. Which means no meaningful agreement and needs adjustment. The improvement might be threshold tuning, or switching to models better suited for imbalanced classification.

Despite an overall accuracy of 94.5%, the confusion matrix (see Figure 1) revealed a major limitation: the model failed to predict any positive (Yes) heart attack cases.

Using the regularized logistic regression model, we analyzed the key predictors retained in the final model. Among the variables, Diabetes emerged as the strongest risk factor, with a coefficient of approximately 1.27, indicating a strong positive association with heart attack incidence. This suggests that individuals with diabetes are significantly more likely to have experienced a heart attack.

We plotted the predicted probability of heart attack (Figure 2). Most probabilities were extremely low, suggesting that the logistic model is under-confident in predicting positive outcomes, likely due to data imbalance.



In contrast, HighBMI showed a mild positive association, suggesting that individuals with a higher body mass index are somewhat more likely to report a heart attack, though the effect size is considerably smaller than that of diabetes. High Risk was associated with a negative coefficient. It implies a potential protective association in this model. This counterintuitive result may reflect biases in how "high risk" is recorded or managed clinically.

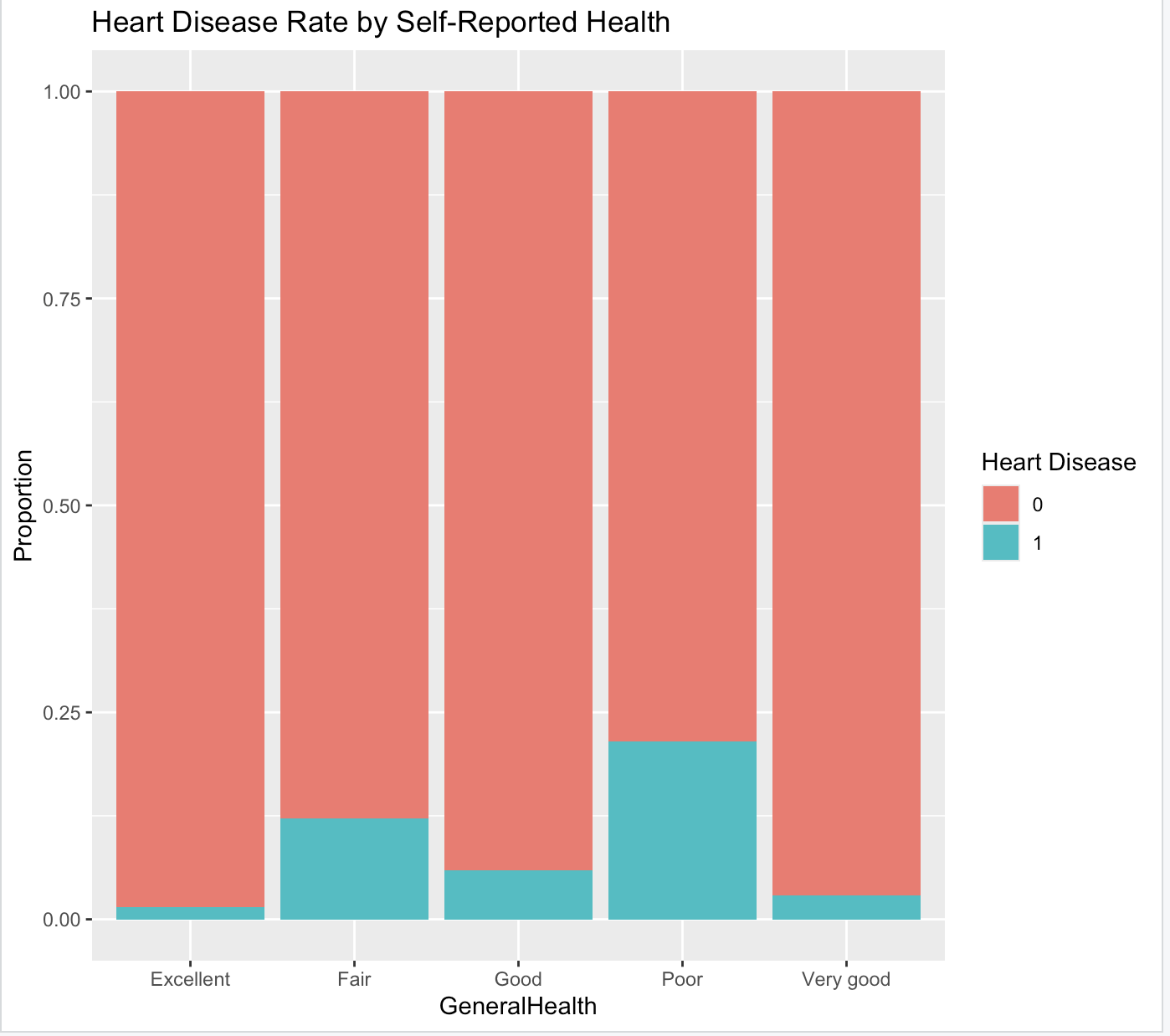
The initial model with 40 features failed to converge and produced an empty model due to overfitting and instability. To improve model stability and interpretability, we narrowed down the feature set to 10 carefully selected variables. After Lasso selection, the final model retained only three predictors: HighRisk, HighBMI, Diabetes.

Conclusion

Our analysis supports the hypothesis that individuals with high risk, high BMI, and diabetes are more likely to have experienced a heart attack, with diabetes emerging as the most significant predictor. Using logistic regression and Lasso-based feature selection, we identified diabetes as having the strongest positive association, while high BMI showed a mild effect. Interestingly, the “high risk” variable had a negative coefficient. These findings emphasize the need for careful modeling in imbalanced datasets and suggest that individuals with diabetes and high BMI should be prioritized for cardiovascular screening and prevention efforts.

**Hypothesis 2:**

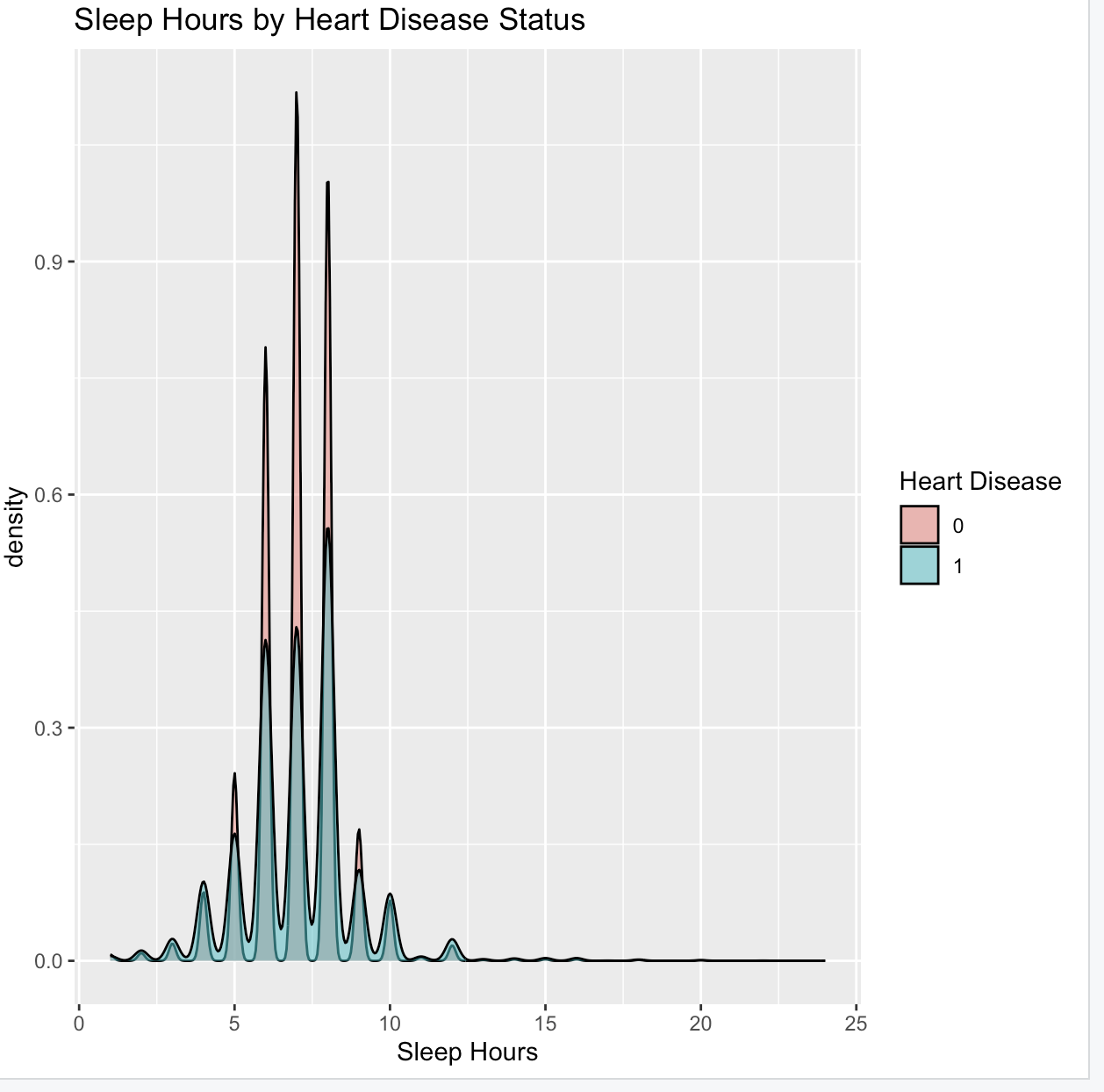
Our aim was to explore whether a small set of self-reported and behavioral features can be used to flag individuals at risk of heart disease. We began by selecting variables that are non-invasive and easily obtainable in non-clinical settings—specifically BMI, sleep hours, smoking status, alcohol consumption, physical activity, mental health days, general self-reported health, age category, and sex.

To understand the relationships between these features and heart disease, we conducted a thorough exploratory data analysis. The bar plots revealed strong associations between the categorical features and higher heart disease rates. It was surprising to note that self-reported mental health status indicated a higher rate of heart disease. Among the numeric variables, BMI and mental health days showed slightly shifted distributions between those with and without heart disease. Sleep hours, while not strongly linearly associated, showed enough variation to warrant further analysis. Correlation among numeric features was low, and VIF scores confirmed there was no multicollinearity issue, making them suitable for use in a multivariable model.

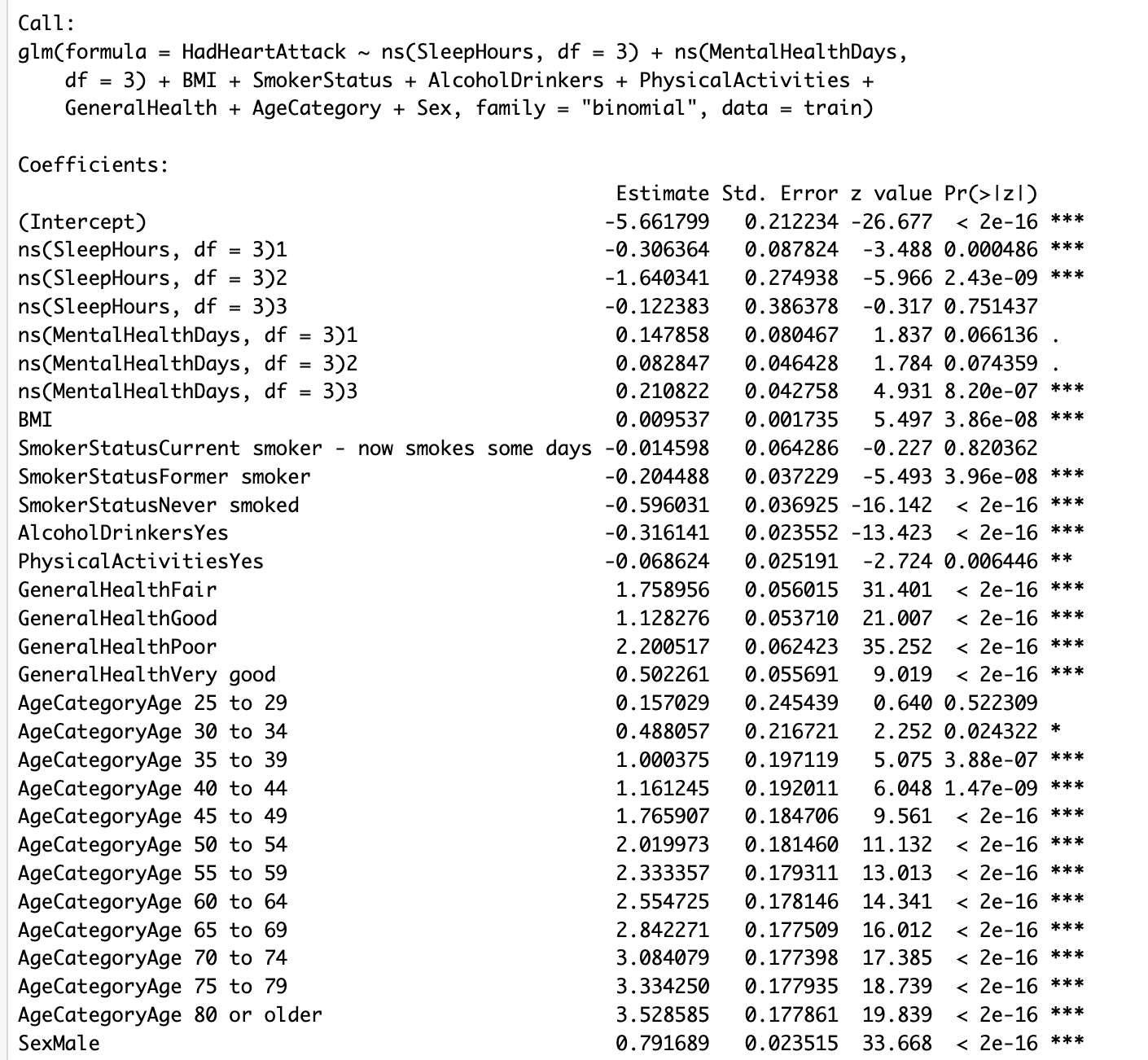
We first trained a baseline logistic regression model using all selected features. While it achieved high accuracy (94.6%) and a strong AUC (0.819), the model had extremely low sensitivity (~0.5%), meaning it failed to detect most true positive heart disease cases. This was attributed to the class imbalance in the data (only ~5% of observations were positive).

To explore model simplicity, we applied LASSO using cross-validation. LASSO confirmed that nearly all of our chosen features were informative—dropping only a few categorical levels—and supported the robustness of our variable selection. This can also be seen from the p-values of the features obtained in our baseline logistic regression model shown in *Fig 2*. However, LASSO did not lead to substantial model reduction or improved classification on its own.

Our next step is to move beyond linearity as non-linear trends were suggested by the ‘SleepHours’ variable as can be seen in *Fig 1*. We plan to model sleep hours and mental health days using natural splines to account for potential nonlinear relationships. This will help us evaluate whether a more flexible model can better capture risk patterns and improve classification performance, especially in borderline cases.



*Fig 1: SleepHours suggests non-linear relation with response*



*Fig 2: Coefficient Estimates and p-values for baseline Logistic Regression Model*

**Hypothesis 3:**

This hypothesis examines whether three key lifestyle-related variables — mental health, physical activity, and alcohol use — show a significant relationship with self-reported heart attacks. If confirmed, these findings can promote targeted interventions that prioritize mental health support and behavioral modifications for cardiovascular prevention.

We fit a logistic regression model with a squared term for MentalHealthDays and an interaction between Inactive and AlcoholUse:

| Term | Estimate | p-value | Interpretation |
| --- | --- | --- | --- |
| MentalHealthDays | -0.044 | < 0.001 | Risk decreases initially |
| I(MentalHealthDays^2) | +0.0019 | < 0.001 | Risk increases after a point (U-shape) |
| Inactive | +0.571 | < 0.001 | Higher risk |
| AlcoholUse | -0.615 | < 0.001 | Slight protective effect alone |
| Inactive × AlcoholUse | +0.149 | 0.0007 | Jointly increases risk |

The effect of mental health days follows a U-shaped curve (Fig.6) moderate mental health distress may have lower reported heart attacks, but very high or very low values increase risk. Being physically inactive increases risk. Interestingly, alcohol use alone is associated with lower odds, but combined with inactivity, the risk significantly increases.



Fig. 6 Mental Health Days vs Heart Attack Risk

Lasso validated the importance of the same variables found in the logistic model. This suggests the selected features are robust, even under regularization constraints, and their contributions are not spurious.

8 x 1 sparse Matrix of class "dgCMatrix"

s0

(Intercept) -2.743607966

MentalHealthDays -0.033984113

Inactive 0.544620079

AlcoholUse -0.559336352

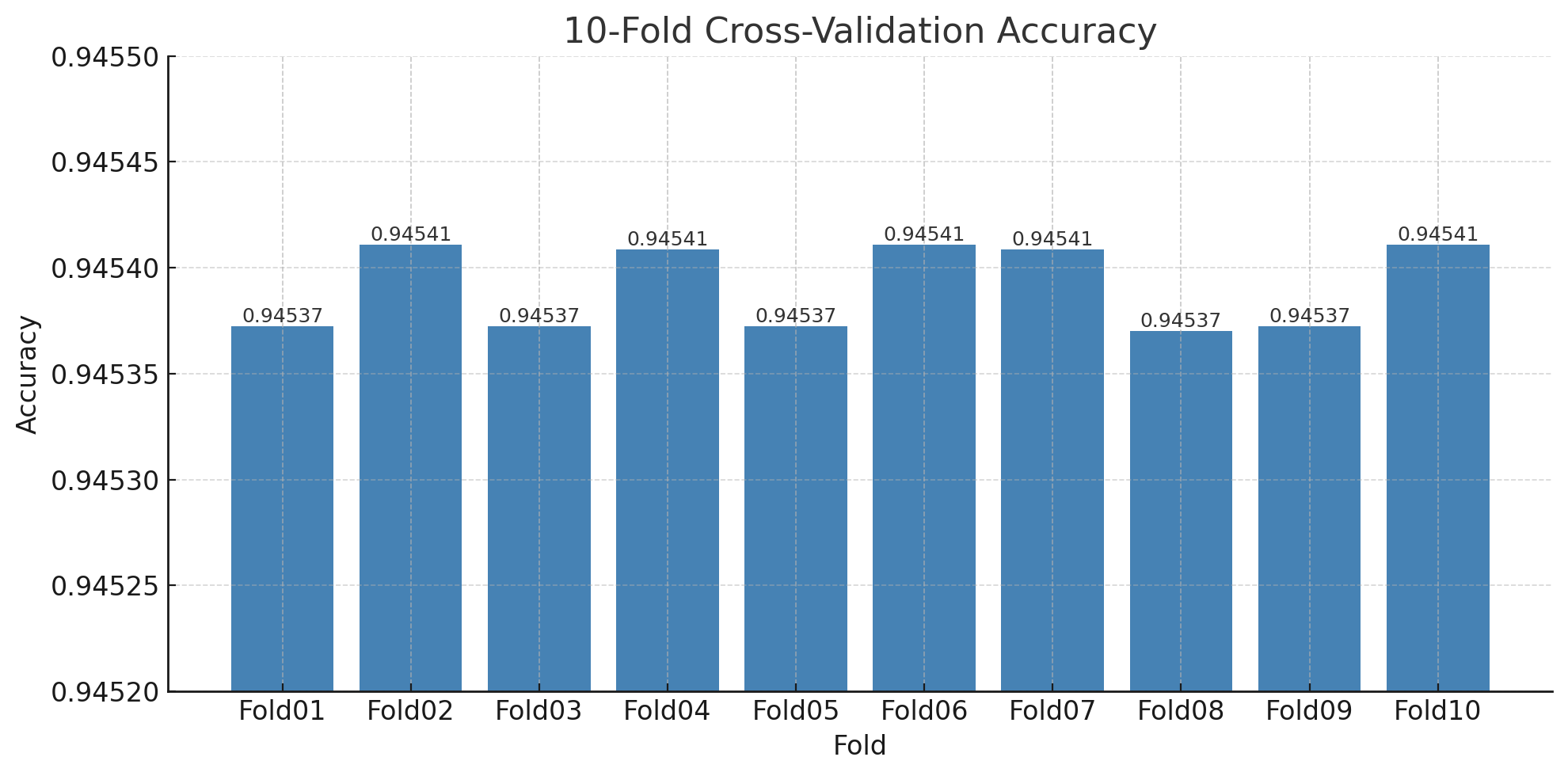
I(MentalHealthDays^2) 0.001529389

MentalHealthDays:Inactive 0.005170267

MentalHealthDays:AlcoholUse -0.007056308

Inactive:AlcoholUse 0.146911051

The output of this model supports the conclusion that behavioral and mental health factors are jointly important in predicting heart attack risk.



While the accuracy was very high and stable across all 10 folds, the Kappa statistic = 0 reveals the model’s inability to capture the positive class This is a sign of severe class imbalance, and reinforces that accuracy alone is not a sufficient metric in this case.

Our analysis provides strong support for Hypothesis 3: that poor mental health, physical inactivity, and alcohol use significantly contribute to heart disease risk. Using logistic regression with nonlinear and interaction terms, we found that the number of poor mental health days had a U-shaped relationship with heart attack risk which suggests that both very low and very high levels of poor mental health may be associated with increased risk. Physical inactivity was consistently associated with higher odds of heart attack, while alcohol use alone appeared slightly protective, though its interaction with inactivity revealed a significant increase in risk when both factors occurred.

Conclusion:

Cardiovascular risk is multifactorial driven not only by clinical conditions like diabetes and obesity but also by mental health, lifestyle, and behavioral factors. Medical institutions and public health stakeholders can use these insights to implement more holistic, preventative screening strategies that combine clinical, behavioral, and mental health indicators—particularly in early or non-clinical settings.