Midterm Project Report

Huanyu Chen, Yifei Liu, Longyu Zhang

Abstract

This study examines the predictive performance of various data science models in estimating recovery time from COVID-19 based on demographic and health-related factors. The techniques used include LASSO regression, Elastic Net, PLS, MARS, GAM, and Random Forest. The results show that MARS has the best predictive performance among all the methods. Our findings underscore the importance of considering both linear and nonlinear relationships in modeling recovery time. Moreover, our findings can have implications for clinical decision-making and resource allocation in COVID-19 management.

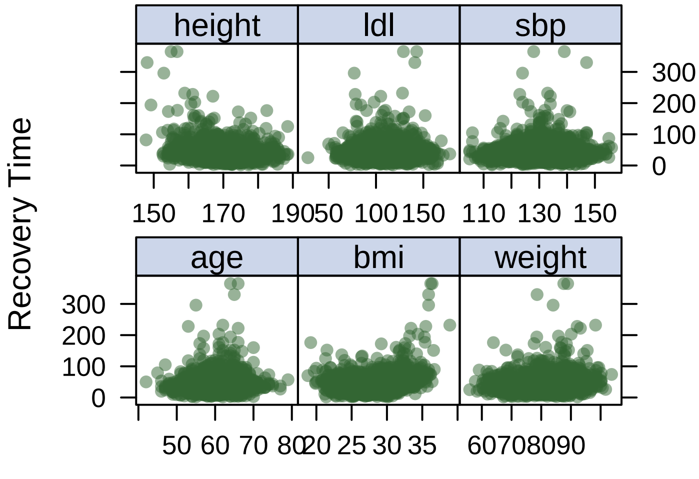
1 Exploratory Analysis and Data Visualization

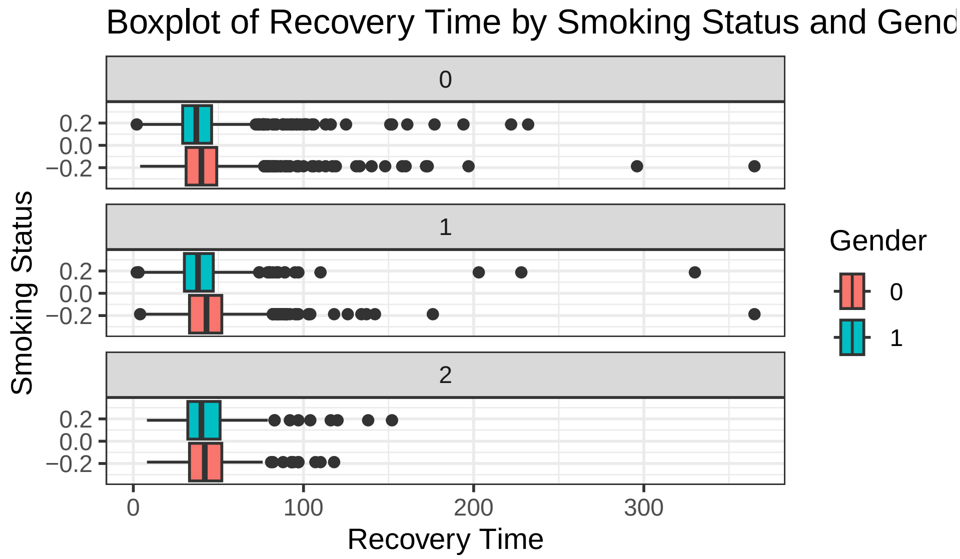
1.1 Variable Types

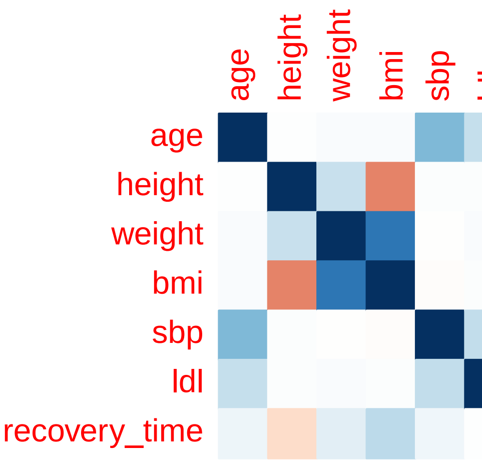
In this dataset, `age`, `height`, `weight`, `bmi`, `SBP`, `LDL`, and `recovery_time` are continuous variables.

1.2 Boxplot of Recovery Time by Smoking Status and Gender

Our analysis reveals a notable trend: across all smoking statuses, females (`gender` = 0) consistently exhibit longer recovery times compared to males. Interestingly, individuals who had never smoked had more outliers on the right side of the boxplot, suggesting a longer recovery time. This counter-intuitive finding suggests that individuals with healthier lifestyles, such as non-smokers, paradoxically require more time to recover from COVID-19.







2 Model Training

3 Result

4 Conclusion

5 Appendix