

An Interpretable XGBoost and GPT-Assisted Data Lakehouse Model for Predicting Community-Level Chronic Illness and Long COVID Vulnerability Scores

Ronish Shrestha and Md Rana

Lamar University

USA

rshrestha14@lamar.edu

Abstract

Despite widespread exposure to COVID-19, the combined influence of social, environmental, and behavioral factors on long-term chronic illness vulnerability remains poorly understood. This study introduces a unified, interpretable framework that integrates multi-domain data to generate actionable vulnerability scores at a county level. A weighted approach was utilized to combine variables in a normalized form to obtain a score for vulnerability. Among various models, Random Forest and XGBoost classification models performed well in classifying the level of vulnerability. Among them, XGBoost had a high accuracy of 93.6%. To further analyze the predictions, SHAP was utilized to identify the most important factors contributing to the prediction. From the findings, it is clear that health conditions, environmental exposure, and behavioral factors are crucial in determining vulnerability.

Keywords

XGBoost, GPT-Assisted Data, Lakehouse Model