Machine Learning Insights into COVID-19 Variant Spread Across US Regions

MSR 6

Hu L, Zhang X, Weimer I, Yapici HO, Shenoy A, Lodaya K, D’Souza F
Boston Strategic Partners, Inc.

BACKGROUND

- The COVID-19 pandemic has prompted global health crises and significant disruptions. [1] A pivotal concern has been the evolution of COVID-19 variants
- While machine learning has emerged as a powerful tool for predicting disease patterns and outcomes, there remains a notable gap in research specifically focused on the regional impacts of COVID-19 variants

OBJECTIVE

- Using machine learning techniques, specifically random forest regression, we evaluated the influence of various regional or temporal factors on the proportion of key COVID-19 variants
- We aimed to identify predictors of variant prevalence and develop a data-driven approach to pandemic management

DATASETS

- The National COVID Cohort Collaborative (N3C) database, which provides the share of variants weekly in each Health and Human Service (HHS) region, was utilized **(Figure 1)**
- Other public data sources used to identify important predictors of variants include:
  - US Department of Transportation - Bureau of Transportation Statistics
  - World Weather Online
  - US Environmental Protection Agency
  - US Census

METHODS

- We utilized and combined data from the aforementioned sources to generate a thorough predictive model
- Training and testing data were separated using a time-series split
- Random forest regression was used to analyze how different factors impact the share of COVID-19 variants such as Alpha (B.1.1.7), Delta (B.1.617.2), Omicron subvariant BA.5, and Omicron subvariant XBB.1.5 across various U.S. regions **(Figure 2)**
- The significance of various predictors was assessed by analyzing their feature importance

Figure 1. Regional Map of HHS 10 Regions[2]

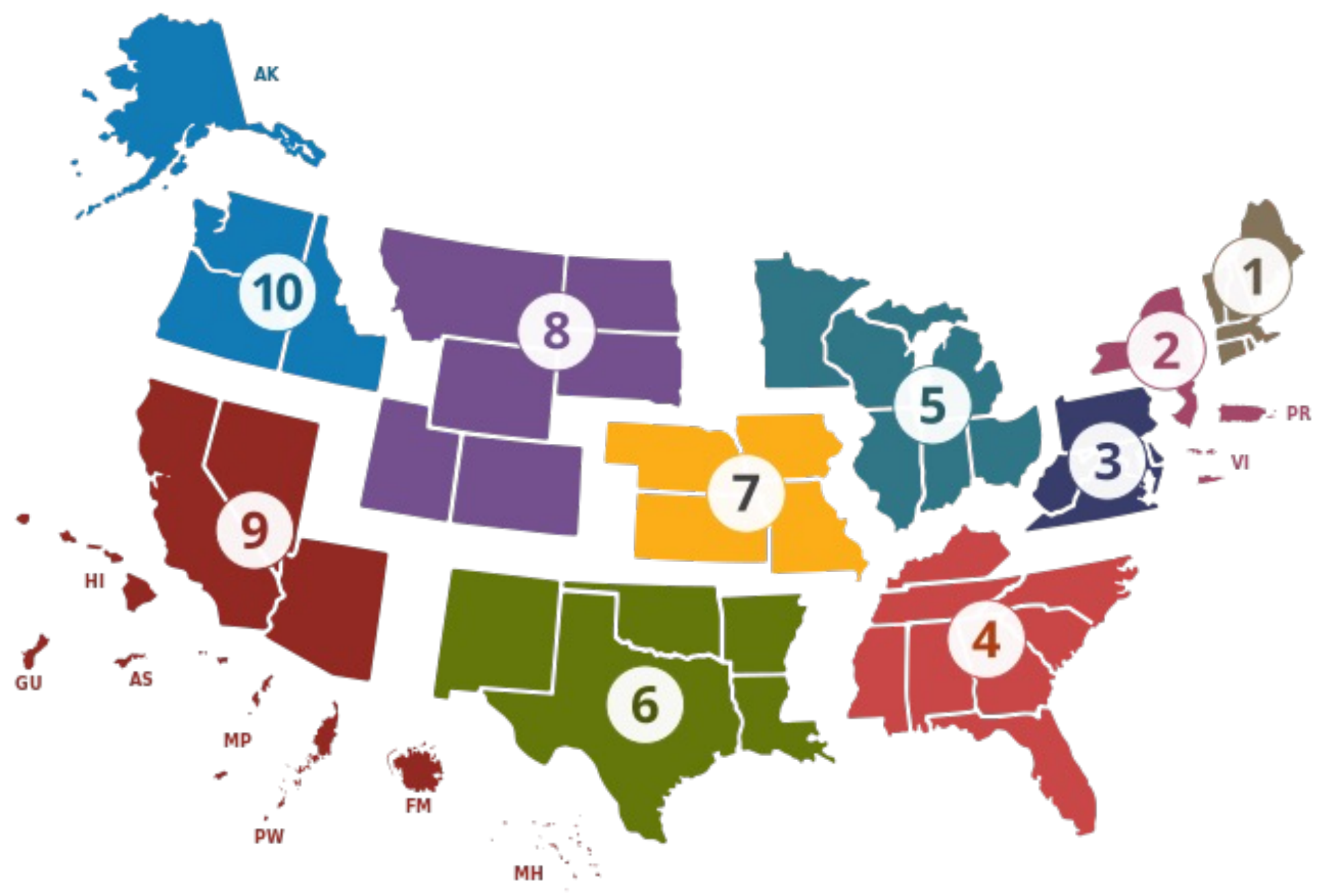
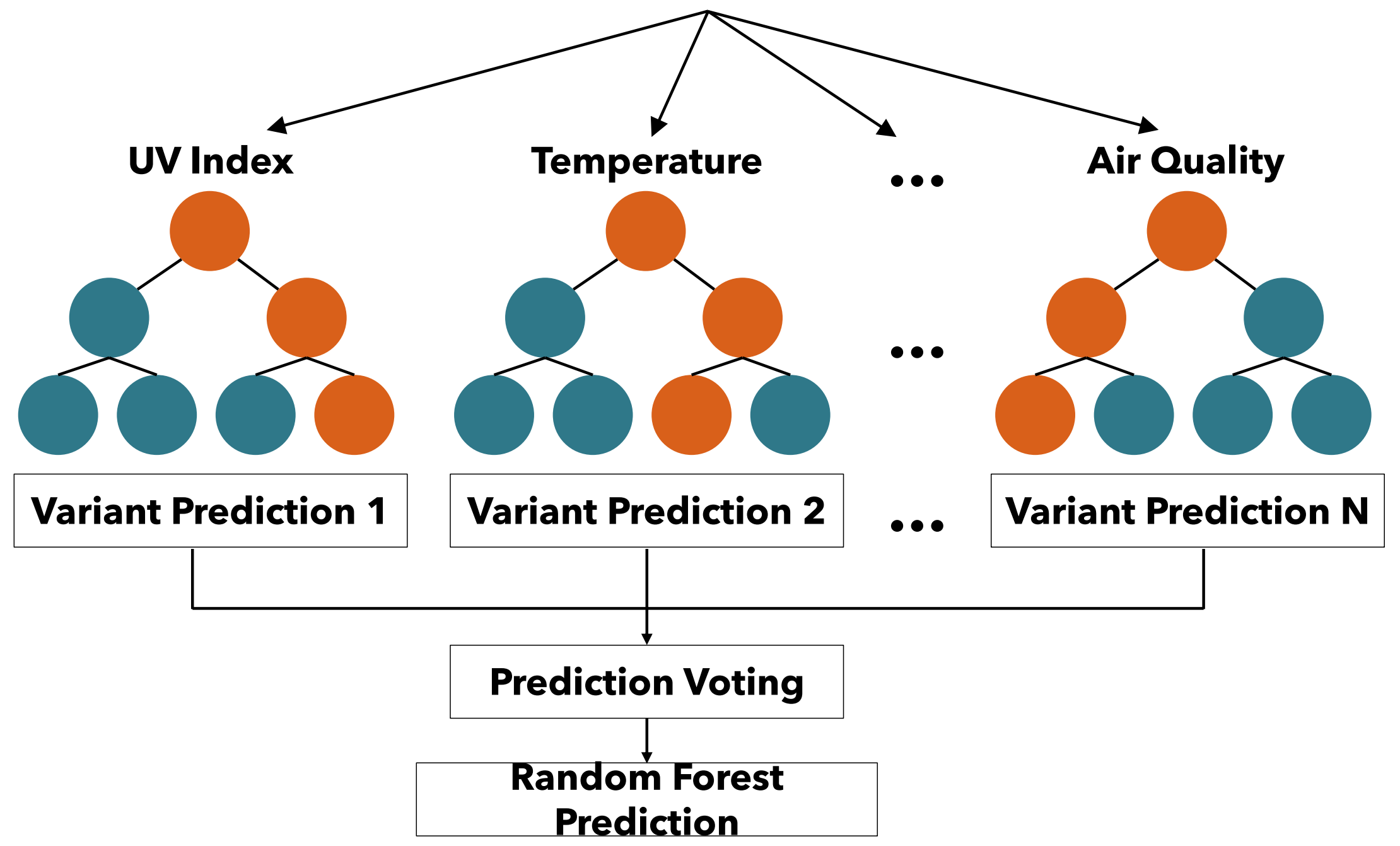


Figure 2. Random Forest Regression Model Structure



RESULTS

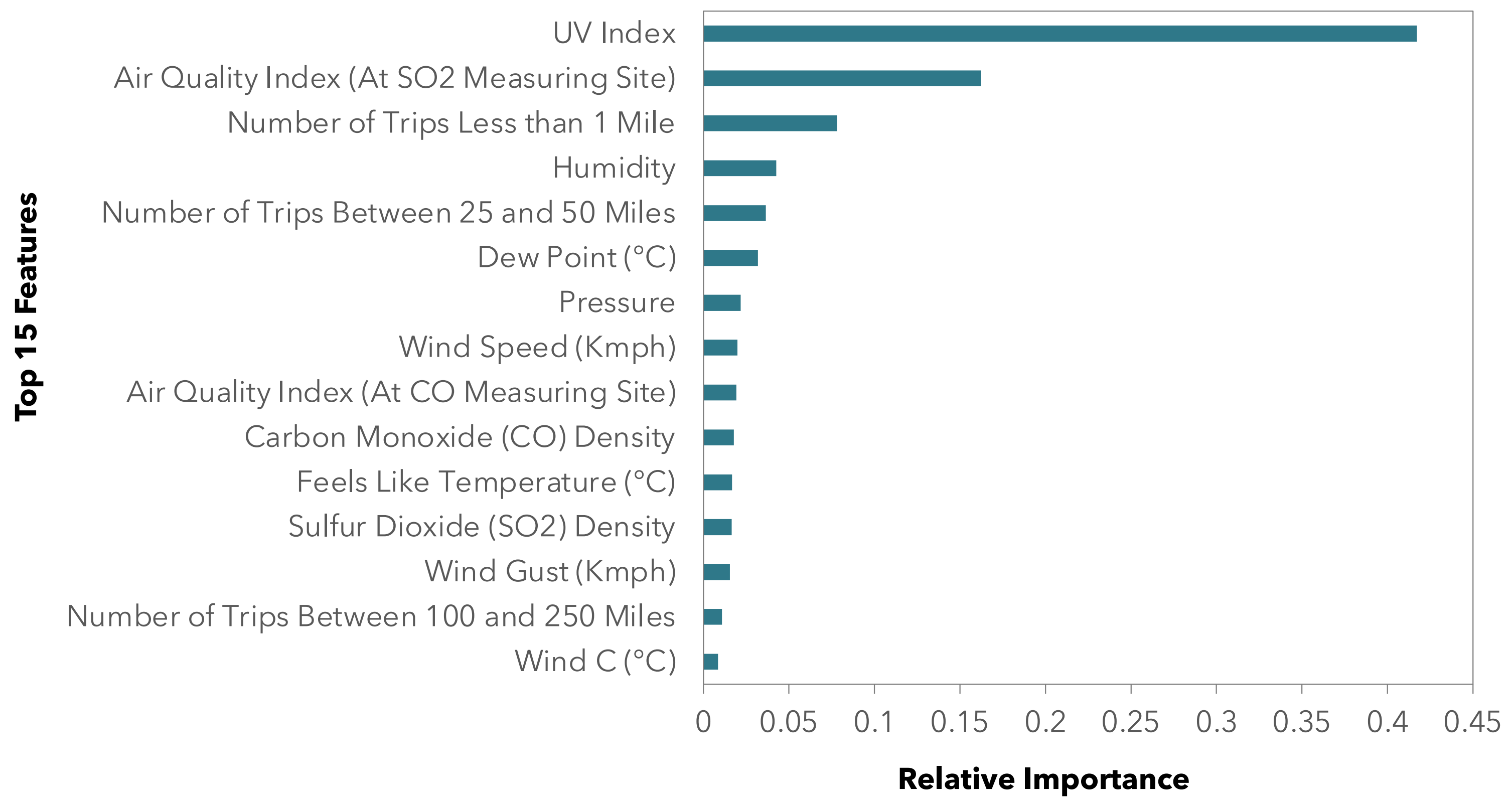
- The model showed high predictive accuracy, with R² values of 0.89 for Delta, 0.93 for BA.5, 0.94 for Alpha, and 0.92 for XBB.1.5, significantly surpassing the 0.72 R² value for a mixed-variant baseline **(Table 1)**
- Delta variant spread correlated strongly with ozone density, Alpha with temperature and air quality, XBB.1.5 with land area and income, and BA.5 with sun hours and UV index **(Figure 3)**
- These results suggest a complex interplay between environmental factors and variant spread. It appears that each variant has its favorable environment; for example, Delta may be more sensitive to ozone density but less sensitive to the temperature, and BA.5 may not be as sensitive to UV index as the other variants **(Figure 4)**

Table 1. Accuracy Metrics of the HHS Region for 4 Variants

Variant Name	Data Length	MSE	RMSE	MAE	R-Square
B.1.1.7 (Alpha)	4398	0.006641	0.081491	0.033258	0.936054
B.1.617.2 (Delta)	33357	0.021188	0.145561	0.031952	0.885554
BA.5 (Omicron subvariant)	8032	0.010736	0.103613	0.025812	0.925882
XBB.1.5 (Omicron subvariant)	210	0.006165	0.078517	0.040888	0.920208

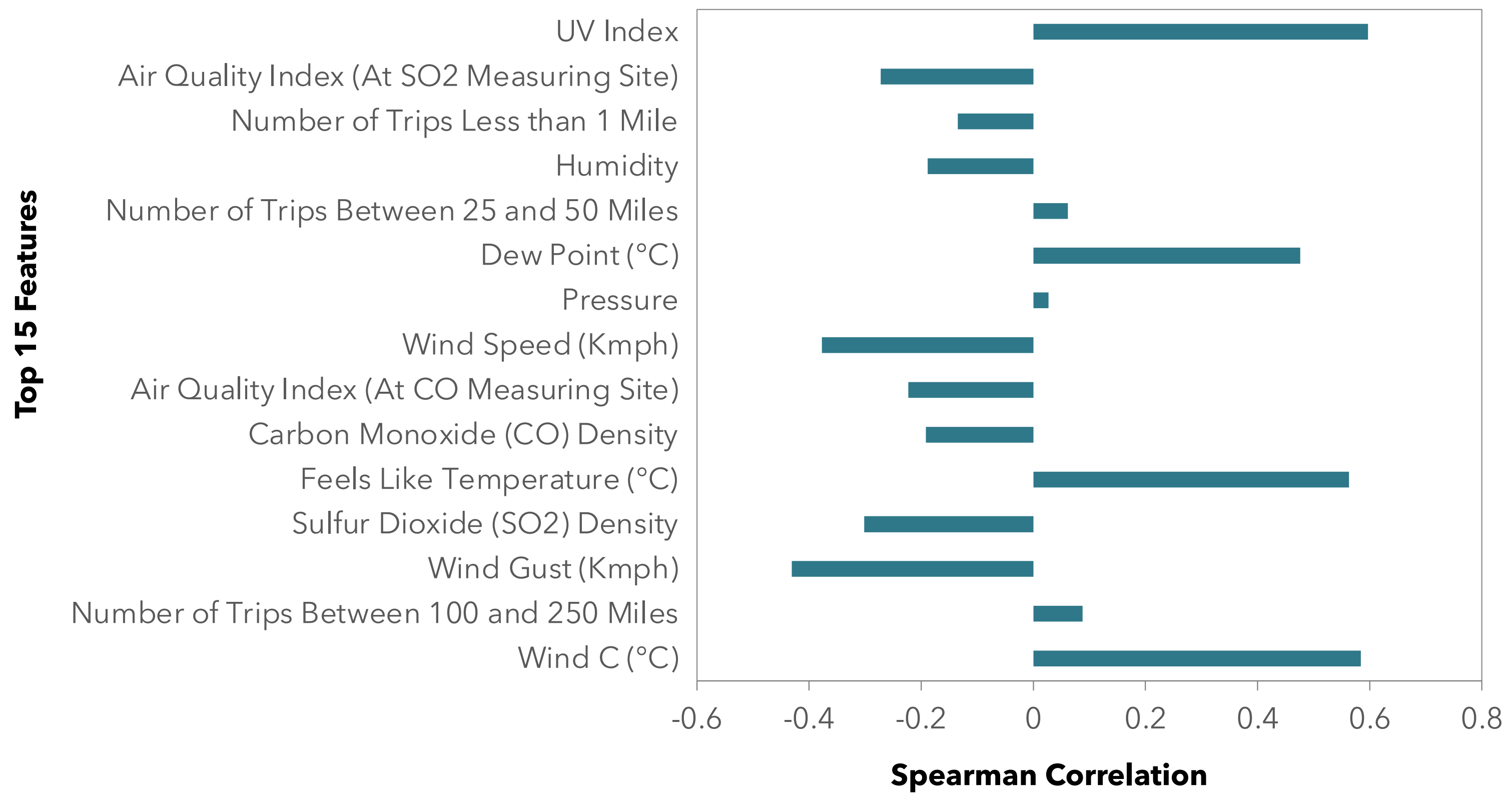
MAE, mean absolute error; MSE, mean squared error; RMSE, root mean squared error

Figure 3. The relative importance for the top 15 features that affects the share of BA.5



Note: Number of trips refers to the number of trips taken by residents, ozone (AQI) refers to the AQI at the ozone measuring site, and ozone (mean value) refers to ozone density.
Abbreviations: AQI; air quality index, CO; carbon monoxide, NO2; nitrogen dioxide

Figure 4. The spearman correlation for the top 15 features that affects the share of BA.5



Note: Number of trips refers to the number of trips taken by residents, ozone (AQI) refers to the AQI at the ozone measuring site, and ozone (mean value) refers to ozone density.
Abbreviations: AQI; air quality index, CO; carbon monoxide, NO2; nitrogen dioxide

CONCLUSIONS

- This research fills a crucial gap in understanding the regional dynamics of COVID-19 variant distribution
- By providing detailed insights into the geographical prevalence of specific variants, our study demonstrates the value of machine learning techniques for future targeted public health strategies and policies

REFERENCES

1. World Health Organization 2023 data.who.int, WHO Coronavirus (COVID-19) dashboard > Cases [Dashboard]. https://data.who.int/dashboards/covid19/cases
2. U.S. Department of Health and Human Services. (n.d.). Regional offices. Retrieved from: https://www.hhs.gov/about/agencies/iea/regional-offices/index.html