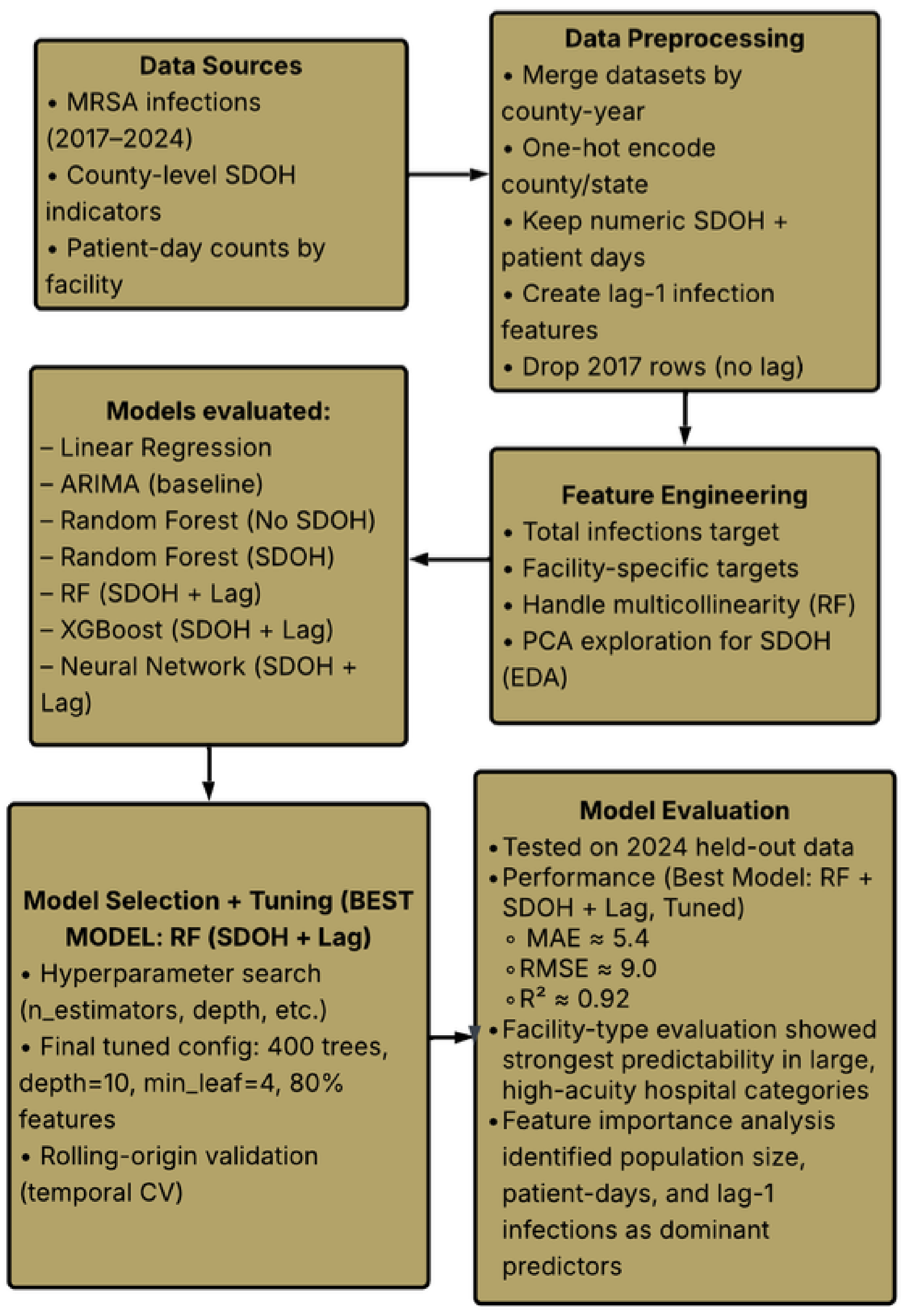
Social Determinants of Methicillin Resistant **Staphylococcus aureus** *Bloodstream Infection Patterns in California*

Hannah Bower, Alan Wang, AJ Subudhi
Georgia Institute of Technology

INTRODUCTION

MRSA bloodstream infections remain a major antimicrobial-resistant threat in the United States[1], yet the extent to which local socioeconomic and structural factors influence infection burden is less understood[2]. This project integrates eight years of **California MRSA surveillance (2017–2024)**[4] with **county-level social determinants of health (SDOH)** indicators to examine how community conditions shape geographic and temporal trends in MRSA infection rates across California counties. We link population size, healthcare utilization, housing instability, income inequality, and environmental metrics with both total MRSA infections and hospital-type specific patterns. By combining epidemiologic context with machine-learning models[3], **we aim to clarify which non-clinical features meaningfully drive MRSA variation and whether SDOH improve prediction beyond known healthcare exposure and population-scale factors**.

METHODS



Data Sources
- MRSA infections (2017–2024)
- County-level SDOH indicators
- Patient-day counts by facility

Data Preprocessing
- Merge datasets by county-year
- One-hot encode county/state
- Keep numeric SDOH + patient days
- Create lag-1 infection features
- Drop 2017 rows (no lag)

Feature Engineering
- Total infections target
- Facility-specific targets
- Handle multicollinearity (RF)
- PCA exploration for SDOH (EDA)

Models evaluated:
– Linear Regression
– ARIMA (baseline)
– Random Forest (No SDOH)
– Random Forest (SDOH)
– RF (SDOH + Lag)
– XGBoost (SDOH + Lag)
– Neural Network (SDOH + Lag)

Model Selection + Tuning (BEST MODEL: RF (SDOH + Lag)
- Hyperparameter search (n_estimators, depth, etc.)
- Final tuned config: 400 trees, depth=10, min_leaf=4, 80% features
- Rolling-origin validation (temporal CV)

Model Evaluation
- Tested on 2024 held-out data
- Performance (Best Model: RF + SDOH + Lag, Tuned)
  - MAE ≈ 5.4
  - RMSE ≈ 9.0
  - R^2 ≈ 0.92
- Facility-type evaluation showed strongest predictability in large, high-acuity hospital categories
- Feature importance analysis identified population size, patient-days, and lag-1 infections as dominant predictors

CONCLUSION

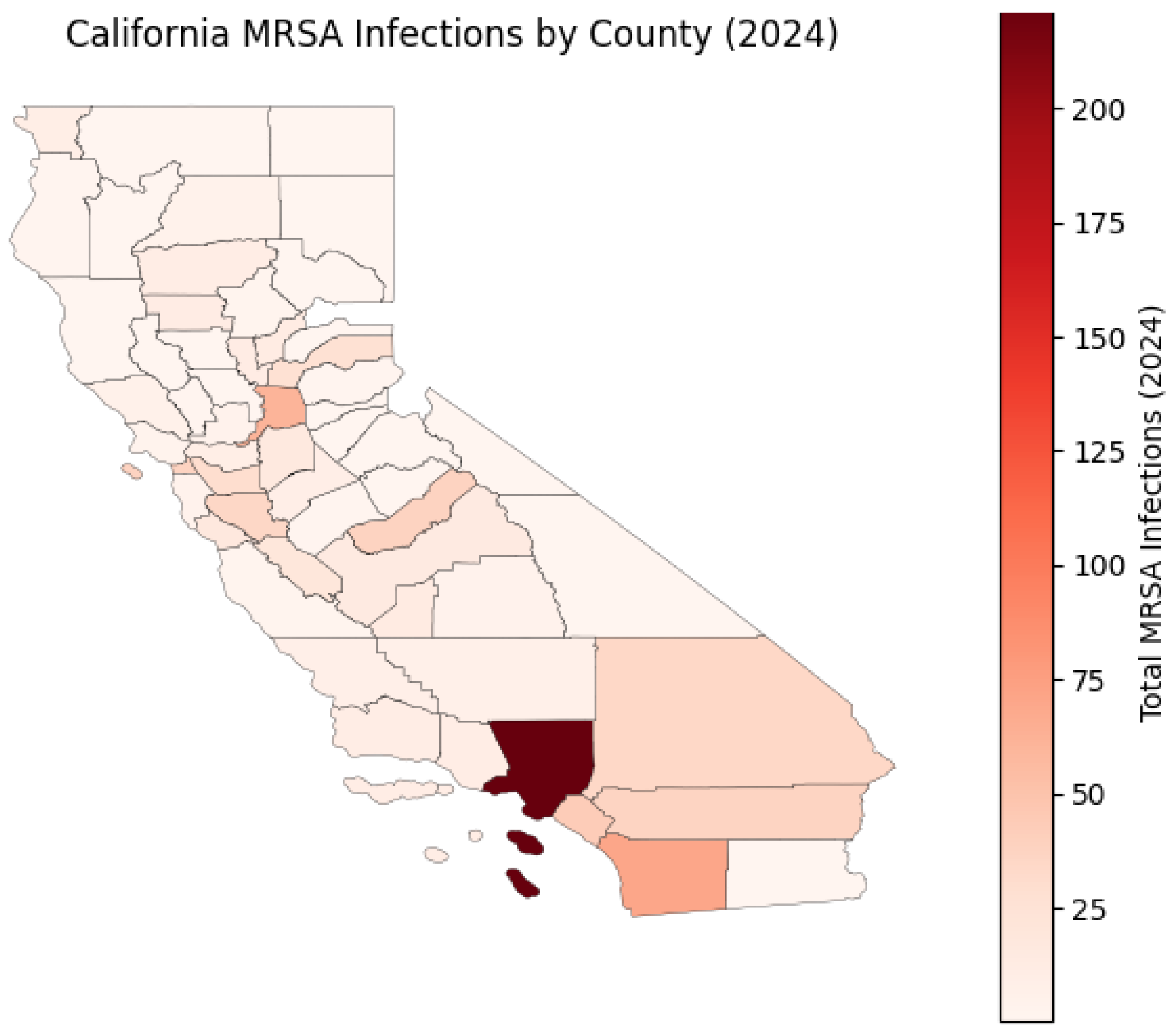
Our findings show that MRSA risk in California counties is driven primarily by **population scale, intensity of healthcare utilization, and the persistence of prior-year infections.**[4] Tree-based models, especially the tuned Random Forest with SDOH + lag features, effectively **capture these nonlinear and temporal patterns** and consistently outperform linear, boosting, and deep learning approaches.[3] Although SDOH indicators are often discussed in MRSA epidemiology,[2] most provided little additional predictive value once population and patient-day variables were included. Only a few structural measures, such as housing instability and income inequality, had meaningful signal. These results suggest that MRSA forecasting and prevention efforts should **prioritize lag-adjusted surveillance, high-volume hospital environments, and structural crowding indicators over broad SDOH feature sets.**

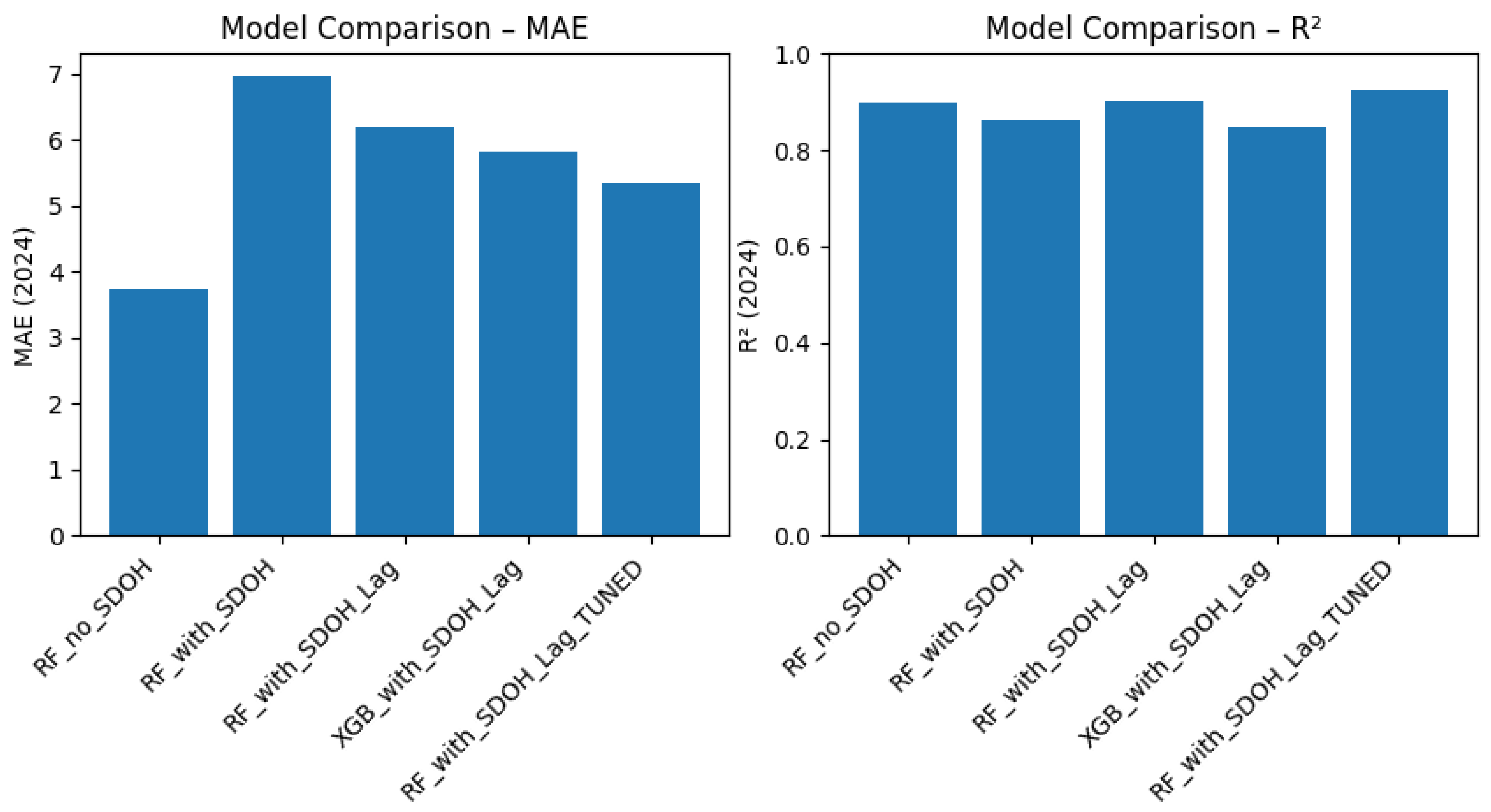
RESULTS

Overall, the map highlights a strong **urban rural divide** in MRSA epidemiology:

- **Urban counties:** persistent, high-volume transmission
- **Rural counties:** sporadic, low-volume, difficult to predict
- **Population size and healthcare exposure** emerge visually as the dominant drivers of MRSA burden

This geographic pattern is consistent with our model's feature importance results, which show that population and patient-day exposure are the primary predictors of county-level MRSA infections.


California MRSA Infections by County (2024)
Total MRSA Infections (2024)
200
175
150
125
100
75
50
25



The tuned Random forest model trained on both SDOH variables and lagged infected counts **(RF_with_SDOH_Lag_TUNED) was the best-performing model in our analysis.** With an RMSE ≈ 0.902 and an R^2 ≈ 0.924, this model outperformed linear, boosting, deep learning, univariate time series, and all other random forest approaches.

Hospital Type	MAE	RMSE	R^2
Major Teaching	2.5473	4.5277	0.9303
Long-Term Acute Care	0.6065	1.7703	0.9029
Community Large	0.8549	2.1003	0.8666
Pediatric	0.1229	0.4164	0.8366
Community Small	0.5199	0.8727	0.8166
Rehabilitation Unit	0.0646	0.1640	0.6889
Community Medium	0.7624	2.0356	0.5057
Critical Access	0.0835	0.2586	0.0465
Free-Standing Rehabilitation Hospital	0.0026	0.0155	0.0000

Using the RF_with_SDOH_Lag_Tuned model, we evaluated predictive accuracy across all nine hospital categories. **Our model performs best in Major Teaching hospitals and Long-Term Acute Care facilities**, the hospital types exhibiting the two highest R^2 values. These settings treat large numbers of patients requiring intensive medical care and long stays, creating steady, repeatable opportunities for MRSA transmission.

The tuned Random Forest showed that MRSA risk is driven primarily by **population scale** (labor force, population), **healthcare exposure** (patient-days in major hospital types), and **previous-year infections**, which together explain most of the model's predictive power. Only a few SDOH variables contributed meaningfully, reinforcing that **county size, hospital utilization, and temporal persistence** are the strongest and most reliable predictors of MRSA burden.

Rank	Feature	Importance	Variable Meaning
1	Labor Force	0.183	Total county labor force (SDOH)
2	Population	0.1334	County population estimate
3	Patient-days: Major Teaching	0.1132	Bed-days in Major Teaching hospitals
4	Patient-days: Community Medium	0.0578	Bed-days in medium community hospitals
5	Population Rank	0.0561	County population percentile
6	Lag-1: Major Teaching infections	0.0487	Previous year's MRSA in Major Teaching hospitals
7	Lag-1: Community Large infections	0.0482	Previous year's MRSA in Community Large hospitals
8	Patient-days: Pediatric	0.0476	Bed-days in Pediatric hospitals
9	Patient-days: Rehabilitation Unit	0.0469	Bed-days in Rehabilitation Units
10	Patient-days: Community Large	0.0427	Bed-days in Community Large hospitals

[1] Nikolaos Andreatos, Fadi Shehadeh, Elina Eleftheria Pliakos, and Eleftherios Mylonakis. 2018. The impact of antibiotic prescription rates on the incidence of MRSA bloodstream infections: a county-level, US-wide analysis. International journal of antimicrobial agents 52, 2 (2018), 195–200.
[2] Sarah Blackmon, Esther E Avendano, Sweta Balaji, Samson Alemu Argaw, Rebecca A Morin, Nanguneri Nirmala, Shira Doron, and Maya L Nadimpalli. 2025. Neighborhood-level income and MRSA infection risk in the USA: systematic review and meta-analysis. BMC Public Health 25, 1 (2025), 1074.
[3] Leo Breiman. 2001. Random Forests. 45, 1 (2001), 5–32. doi:10.1023/A: 1010933404324
[4] California Department of Public Health, Healthcare-Associated Infections Program. 2025. Methicillin-resistant Staphylococcus aureus (MRSA) bloodstream infections (BSI) in California hospitals. https://data.chhs.ca.gov/dataset/methicillin-resistant-staphylococcus-aureus-mrsa-bloodstream-infections-bsi-in-california-hospitals.

Created with BioRender Poster Builder