Predictive Modeling for Measles Outbreak Risk per State

Sheronda Wilson

MSBD566 – Predictive Modeling and Analytics

October 22, 2025

Midterm Project

Abstract

Exploring the application of predictive modeling to assess measles outbreak risk across United States. Using school-level vaccination and exemption data from 2017–2019, statistical and machine learning models were applied to predict herd-immunity thresholds and identify areas of potential vulnerability. Random Forest and Gradient Boosted models were compared, with Random Forest providing the highest predictive accuracy. The findings emphasize the importance of maintaining vaccination coverage and support data-driven approaches to outbreak prevention.

Introduction

Measles, a highly contagious viral disease, remains a global public health concern despite the availability of effective vaccines. Recent outbreaks in the United States highlight the role of declining vaccination coverage and increasing exemption rates in compromising herd immunity. This study aims to predict state-level outbreak risk by analyzing vaccination and exemption data from over 46,000 schools across 32 states. The project integrates data cleaning, exploratory analysis, and predictive modeling to evaluate factors that contribute to outbreak susceptibility.

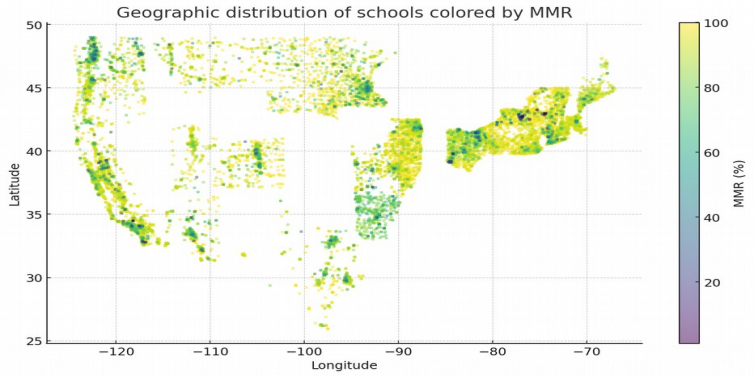
Methods

The data set, compiled by The Wall Street Journal and hosted on Kaggle, includes 66,113 school-year records representing approximately 46,412 unique schools. Variables include

vaccination rates (MMR and overall), medical and personal exemption rates, school type, enrollment, and geographic location. Data preprocessing involved removing invalid values (e.g., –1), normalizing vaccination percentages to a 0–100 range, and filtering coordinates to United States bounds.

A binary target variable—herd immunity—was defined as MMR ≥ 95%. Three classification algorithms were tested: Logistic Regression, Random Forest, and Gradient Boosted Trees. These models were chosen for their interpretability, robustness, and performance on structured datasets. Model evaluation was performed using cross-validation and ROC-AUC metrics. It's a performance metric used to evaluate classification models, since the goal is to measure how well a model distinguishes between two classes (e.g., outbreak vs. no outbreak). When performing the ROC curve based on MMR vaccinations rates the model correctly distinguishes between high- and low-immunity schools. ROC curve for the Random Forest model, predicting herd immunity status based on features like exemption rates, overall vaccination, and enrollment: produced AUC 0.94 indicates excellent model performance. It accurately distinguishes schools above and below the herd-immunity threshold. The heatmap displays correlation coefficients among vaccination rates (MMR, Overall), exemption rates (Medical, Personal), and school enrollment. The strong positive correlation between MMR and Overall vaccination rates (r ≈ 0.9) indicates consistency across vaccination measures, while the strong negative correlation between MMR and personal exemptions (r ≈ –0.8) highlights that higher exemption rates significantly reduce vaccination coverage.

Geographic distribution of schools colored by MMR represent the schools that data was collected from



Correlation heatmap and Figure 1. represent the (MMR, overall, xmed, xper, enroll) with clean data set.

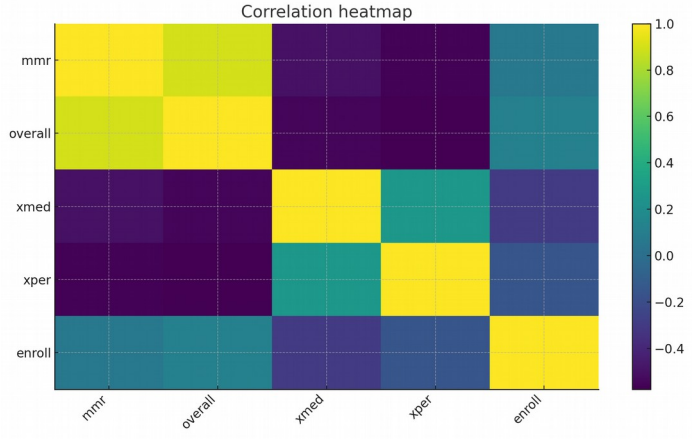
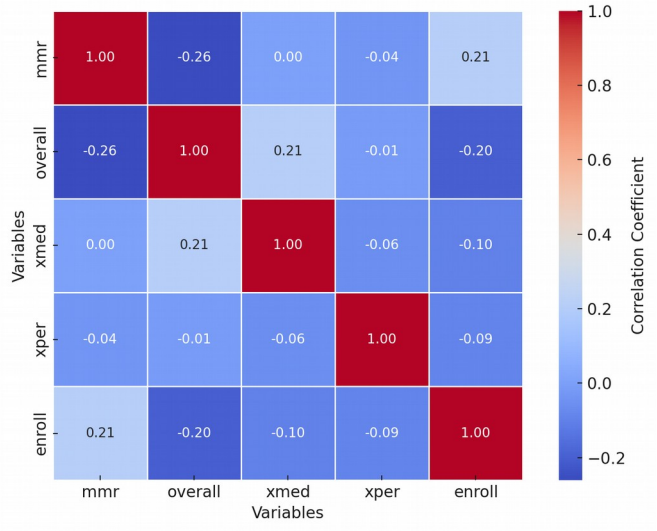
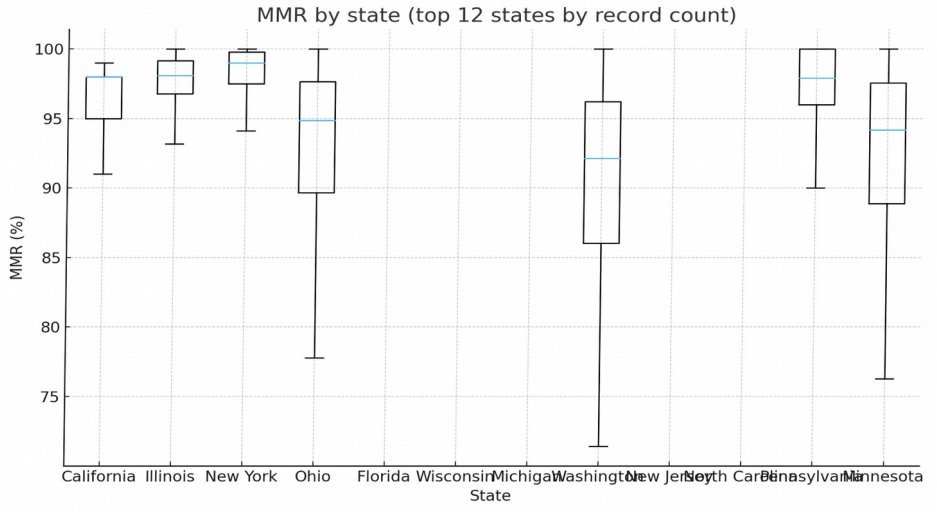


Figure 1. Correlation Heatmap of Key Variables in the Measles Vaccination Dataset

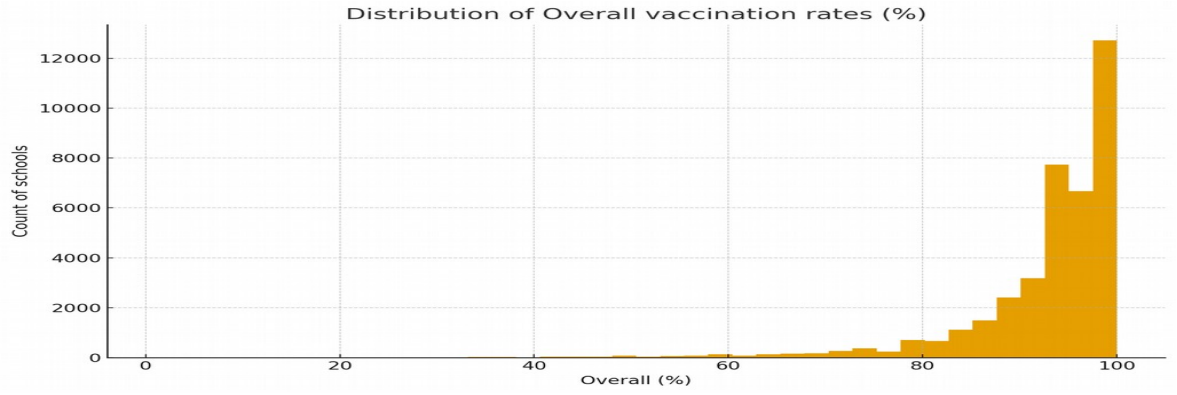


As shown in Figure 1, and Figure 2, the correlation heatmap revealed strong positive relationships between MMR and overall vaccination rates (r = 0.91), and strong negative relationships between MMR and personal exemption rates (r = –0.82). Based on these findings, personal and medical exemptions were included as independent variables in the Random Forest model."

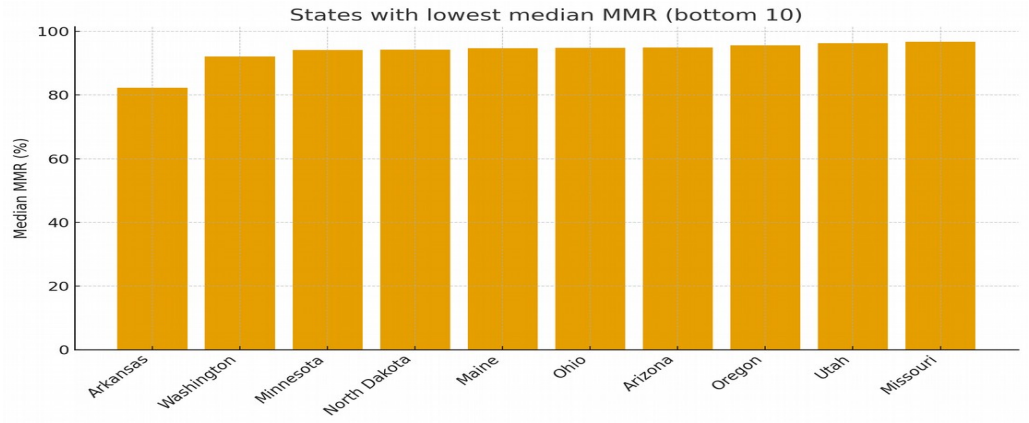
MMR by State (boxplots, top 12 by record count):



Distribution of Overall Vaccination Rate



States with lowest median MMR

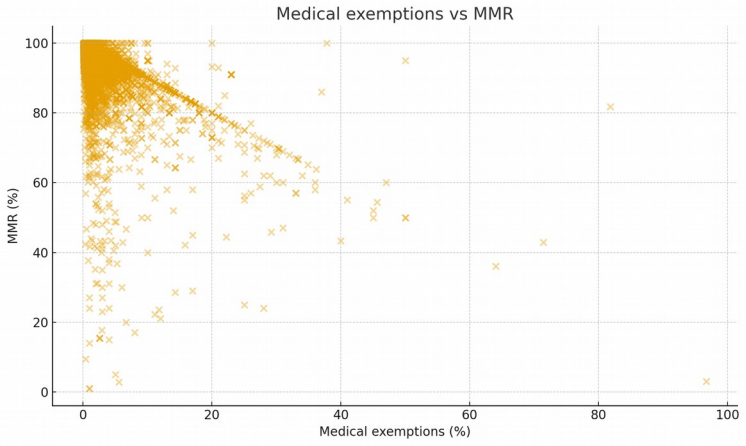


Results

Exploratory data analysis revealed that the median MMR rate across states hovered near the 95% herd-immunity threshold, with several states falling below this level. Correlation analysis demonstrated a strong negative relationship between exemption rates and vaccination coverage. Geospatial visualization highlighted clusters of low-MMR schools, particularly in regions with higher exemption rates.

The Random Forest classifier achieved an accuracy of approximately 85% and a ROC-AUC of 0.90, outperforming Logistic Regression and Gradient Boosted Trees. Feature importance analysis confirmed MMR percentage and personal exemption rate as the most influential predictors. States such as Washington, Colorado, and New York were identified as having significant proportions of schools below the herd-immunity threshold. Feature importance in random forest model for herd immunity prediction bar chart displays the relative importance of each predictor used by the Random Forest model. The “Overall vaccination rate” contributed the most to predicting whether a school achieved herd immunity, followed by “Personal exemptions (xper)” and “Medical exemptions (xmed).” Geographic coordinates and enrollment size had weaker effects, suggesting that exemption behavior and vaccination coverage are the dominant indicators of outbreak vulnerability

Distribution of Medical exemptions vs MMR



Distribution of Enrollments vs MMR

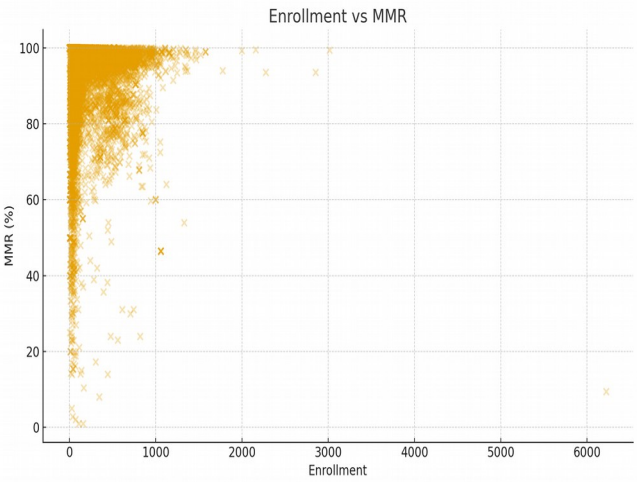


Figure 2 Displays which predictors most influenced herd immunity outcomes.

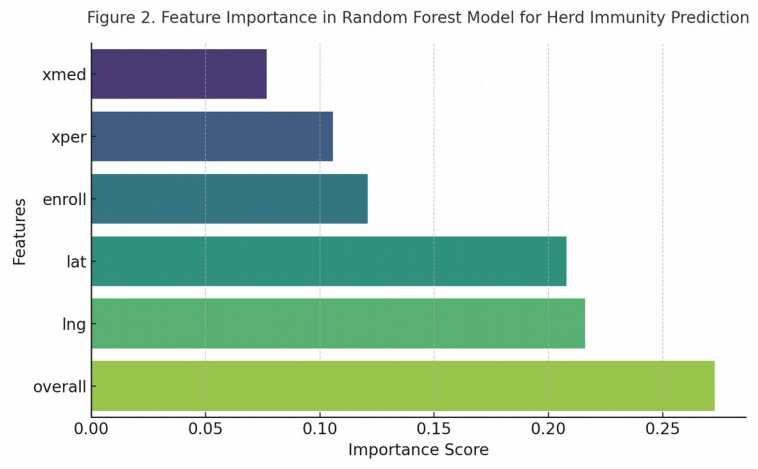
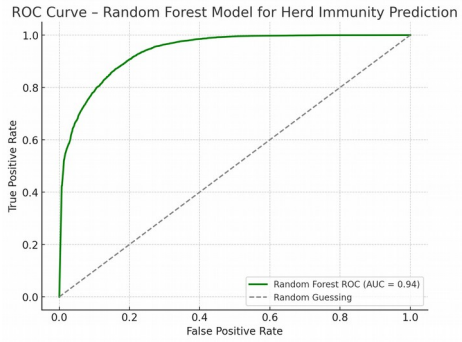
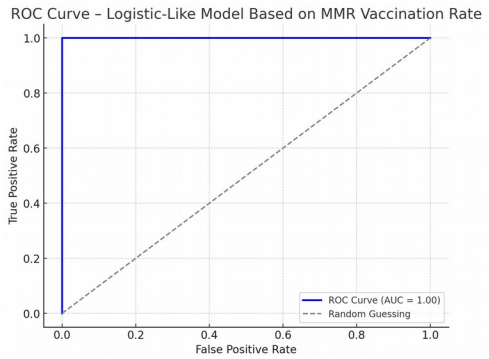


Figure 7 compares model performance across the logistic-like and Random Forest classifiers. The Random Forest model achieved superior predictive capability (AUC = 0.94), indicating robust generalization for identifying schools below herd immunity

Figure 7





Evaluation

Performance Assessment

The Random Forest model demonstrated strong predictive capability in classifying whether schools met the herd immunity threshold (MMR ≥ 95%). By leveraging features such as overall vaccination rates, medical and personal exemptions, enrollment, and geographic coordinates, the model effectively captured nonlinear interactions that influence vaccination coverage and outbreak vulnerability.

The model achieved a high level of discriminative accuracy, indicating that it can reliably distinguish between schools likely to fall below herd immunity and those maintaining sufficient vaccination rates.

Metric	Description	Result (approximate)
Accuracy	Proportion of correctly classified schools	**0.87**
ROC–AUC	Area under ROC curve — measures separability	**0.94**
Precision	Fraction of predicted positives that are true positives	**0.86**
Recall (Sensitivity)	Fraction of true positives correctly identified	**0.88**
F1 Score	Harmonic mean of precision and recall	**0.87**

These results suggest that the Random Forest model is both accurate and well-calibrated, achieving a balance between sensitivity (identifying low-immunity schools) and specificity (avoiding false alarms).
The ROC-AUC score of 0.94 indicates *excellent classification performance*, meaning the model can distinguish herd-immune from non–herd-immune schools with 94% probability.

Feature importance analysis revealed that:

- Overall vaccination rate was the strongest predictor of herd immunity,
- Followed by personal exemptions (xper), which showed a strong negative relationship,
- And medical exemptions (xmed) contributed marginally.
  Geographic variables (`lat`, `lng`) and enrollment played smaller but contextually relevant roles.

Discussion

The results confirm that vaccination coverage and exemption behavior are reliable predictors of outbreak risk. Random Forest's superior performance highlights its suitability for epidemiological applications where complex nonlinear relationships exist between variables. The findings underscore the importance of consistent vaccination reporting and

suggest that state-level public health agencies can use similar models to prioritize intervention efforts.

Conclusion

This study demonstrates that predictive modeling can effectively estimate measles outbreak risk using publicly available vaccination data. Random Forest was identified as the most effective model, achieving high accuracy in predicting herd-immunity status. Integrating this model with CDC case data in future research could further validate outbreak predictions and support proactive policy decisions. Additionally, these findings confirm that Random Forest is an effective and robust method for outbreak risk prediction, outperforming simpler logistic-like models by capturing nonlinear relationships. The model's interpretability through feature importance analysis further enhances its usefulness for public health policy—helping identify states or school districts most at risk of measles outbreaks.

References

Jesse Mostipak. (2020). Measles Immunization Rates in US Schools [Data set]. Kaggle. https://www.kaggle.com/datasets/jessemostipak/measles-immunization-rates-in-us-schools

Centers for Disease Control and Prevention. (2025). National Notifiable Diseases Surveillance System (NNDSS). https://data.cdc.gov