Predicting Antibiotic Resistance in Salmonella enteritidis

Introduction

Antibiotic resistance is a pressing public health issue with widespread economic consequences. As bacterial pathogens like *Salmonella enteritidis* develop resistance to common antibiotics, treating infections becomes increasingly challenging, leading to higher healthcare costs and greater risks for affected populations. *Salmonella enteritidis*, a significant cause of foodborne illness globally, has exhibited growing resistance to antibiotics, complicating treatment and control efforts.

This project aims to develop a predictive model that forecasts antibiotic resistance trends in *Salmonella enteritidis* by analyzing isolate data. Leveraging machine learning techniques, the model utilizes antibiotic susceptibility testing (AST) results, along with specimen metadata such as the year of collection and geographic region, to identify emerging resistance patterns. These predictions are intended to inform public health strategies, optimize resource allocation, and ultimately improve food safety measures.

Data Overview

The dataset used in this project is sourced from the CDC National Antimicrobial Resistance Monitoring System (NARMS), which provides comprehensive surveillance data on antimicrobial resistance in foodborne pathogens. This dataset includes detailed information on antibiotic susceptibility testing and resistance determinants, making it well-suited for predictive modeling of resistance trends. The dataset consists of various features, including:

- Specimen ID: A unique identifier for each isolate.
- Antibiotic Conclusion Columns: Indicating susceptibility (S), intermediate (I), or resistance (R) for various antibiotics.
- Year and Region: The year and geographic region where the specimen was collected.
- Age Group: The age group of the individual from whom the specimen was collected.

Data Wrangling

The data wrangling process involved several key steps to ensure the quality and relevance of the dataset:

- Duplicate Examination: The Specimen ID was checked for duplicate entries to ensure that each isolate is unique.
- Column Removal: Irrelevant columns were removed, along with any that contained more than 90% missing data.

- Imputation of Missing Values: Missing values in the Age Group and Specimen Source columns were imputed using the most common values from those columns.
- Data Year Selection: The dataset was filtered to include only the years from 2003 to 2023 to focus on the last 20 years of data.
- Removal of Antibiotic Conclusion Columns: Columns containing a majority of values marked as X (indicating no breakpoints or epidemiological cutoff values determined) were removed from the dataset.
- Calculation of Antibiotic Resistance Level (AR Level): The Antibiotic Resistance Level (AR Level) was calculated for each specimen by summing all counts of resistant (R) values across the conclusion columns.
- The cleaned DataFrame was saved as `salmonella_isolate_data_cleaned.csv` for further processing.

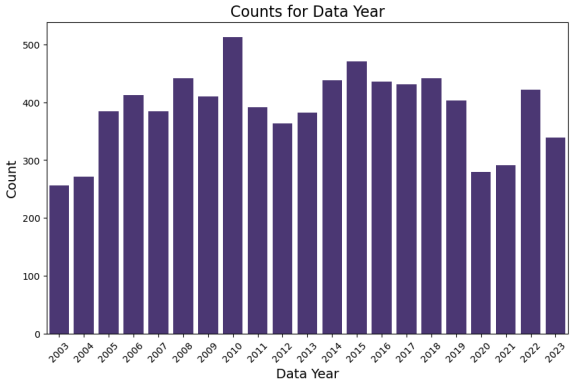
Salmonella Antibiotic Resistance Percentage

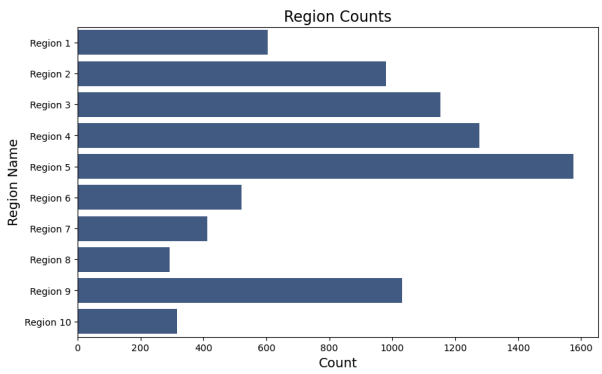
- Antibiotic resistance data was processed by grouping it by year and region to calculate the percentage of resistant strains for each antibiotic. The columns were renamed for clarity, formatted for readability, and the percentages were aggregated into a total resistance percentage. This aggregated data was saved as `salmonella_antibiotic_resistance_percentage.csv` for further processing.

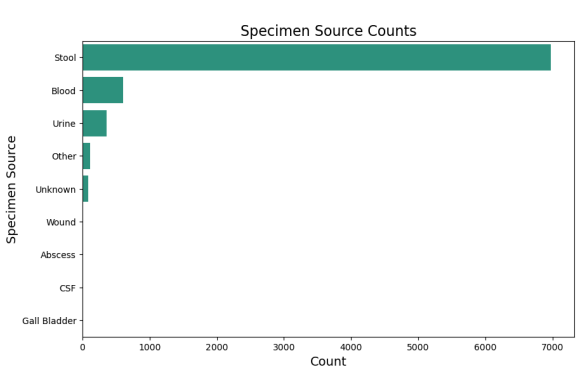
Exploratory Data Analysis (EDA)

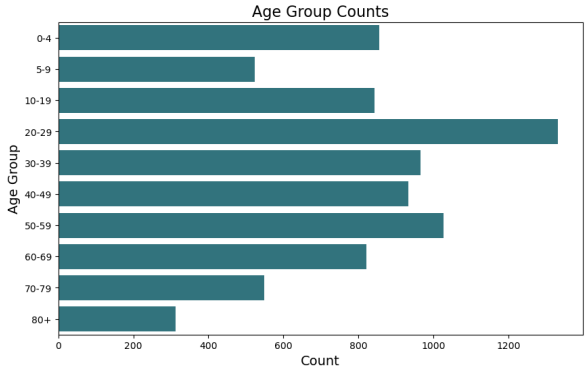
Sample Collection Trends

- Variations in sample collection by year and region suggest shifts in infection rates or surveillance efforts.
- The dominance of stool specimens supports the clinical practice of testing for gastrointestinal pathogens.
- The age group distribution indicates a higher incidence of infections in younger populations, highlighting the need for targeted public health interventions.



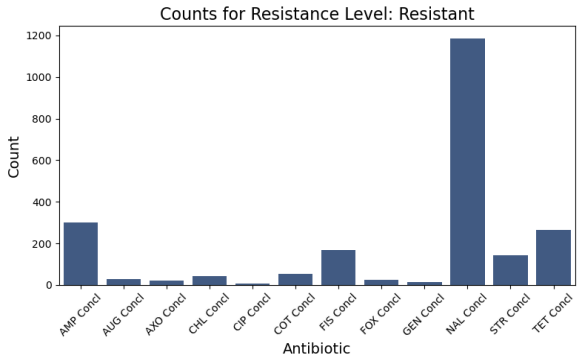


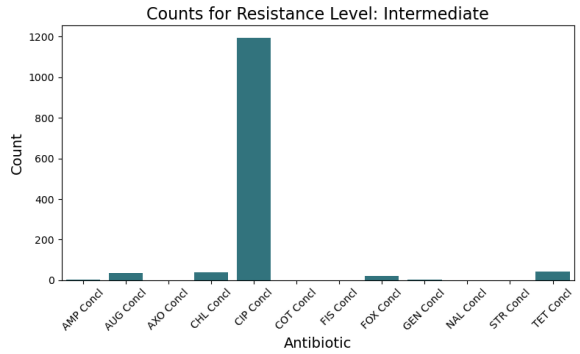




Antibiotic Susceptibility and Resistance

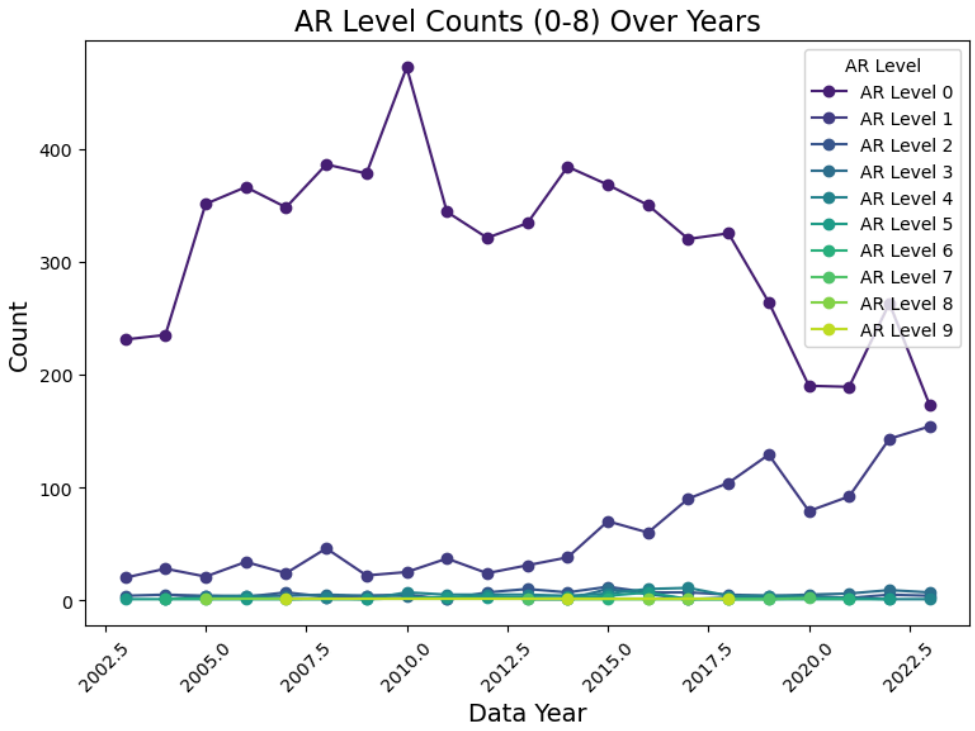
- The prevalence of susceptible strains underscores the effectiveness of many antibiotics, while emerging resistance is notable for specific antibiotics, particularly Ampicillin (AMP), Fosfomycin (FIS), Nalidixic Acid (NAL), Streptomycin (STR), and Tetracycline (TET). Intermediate resistance levels are also observed for Ciprofloxacin (CIP).





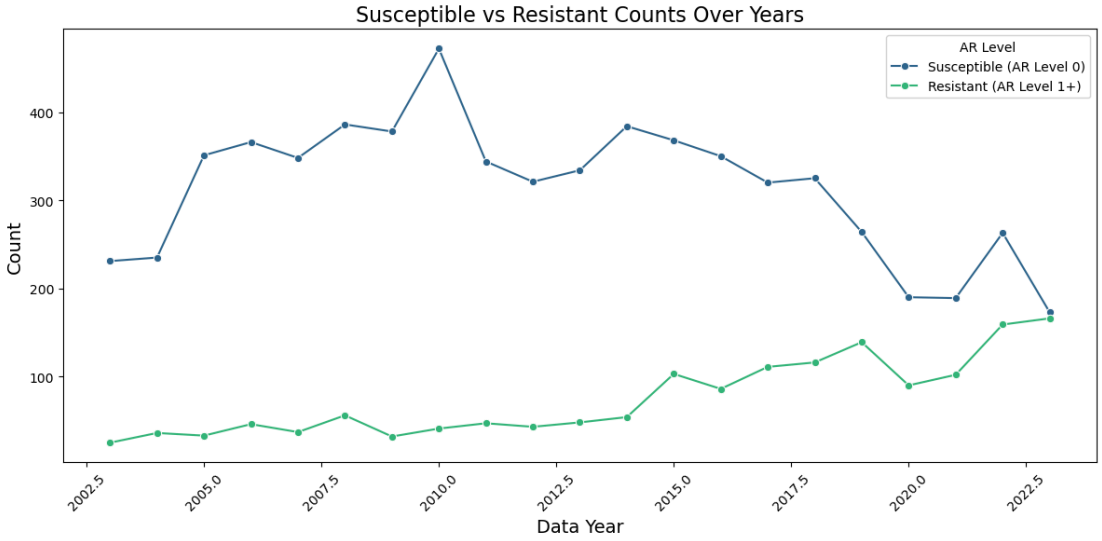
AR Level Counts (0-8) Over Years

- The data shows consistently high counts for AR Level 0 (no resistance), indicating significant susceptibility among isolates.
- There is a peak in AR Level 1 around 2011, followed by fluctuations in levels 2 to 4, suggesting the emergence of some resistance.
- Higher resistance levels (5-8) remain low, indicating severe resistance is relatively uncommon.



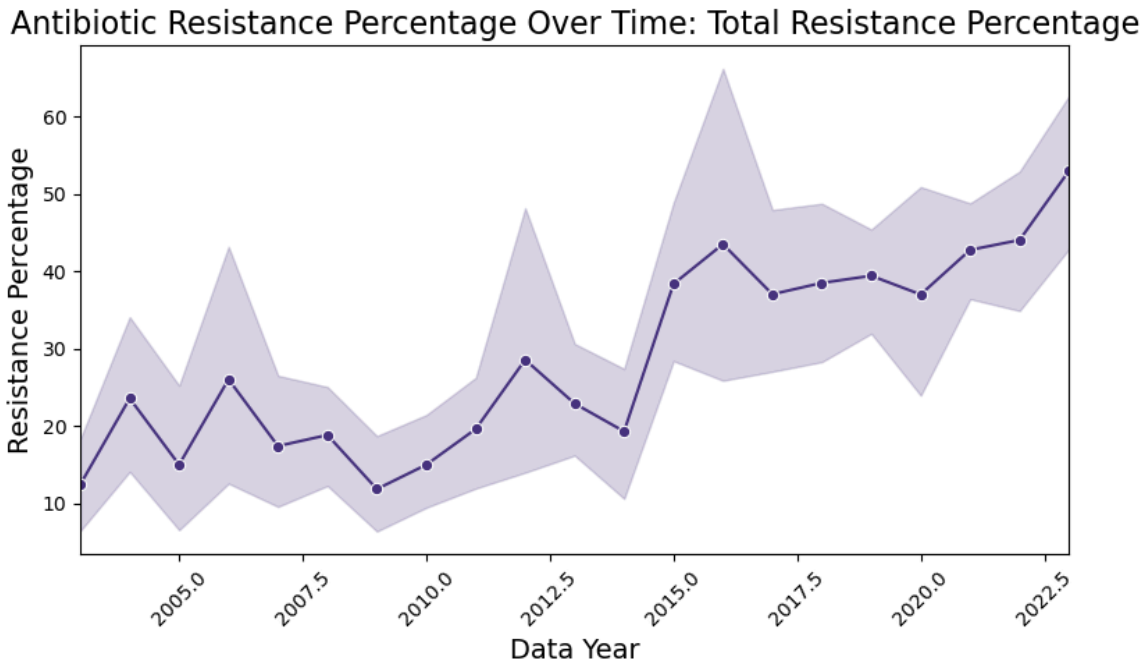
Susceptible vs. Resistant Counts Over Years

- The counts of susceptible isolates (AR Level 0) consistently exceed those of resistant isolates (AR Level 1+).
- However, a slight upward trend in resistant counts is observed from 2015 onward, raising concerns about increasing antibiotic resistance.



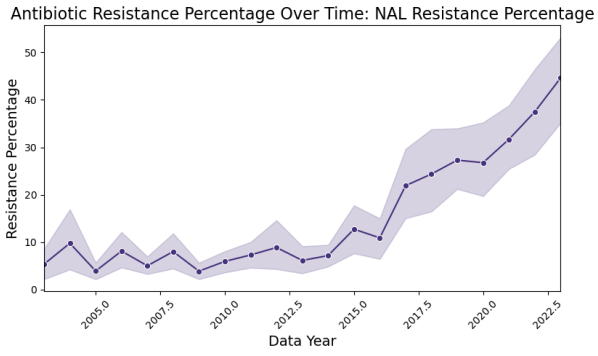
Overall Resistance Trends

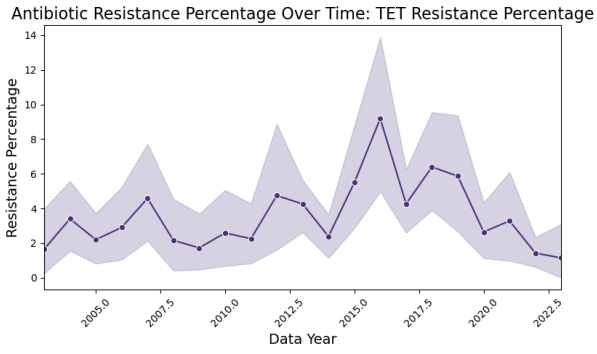
- The total antibiotic resistance percentage shows a clear upward trend, indicating a growing challenge in managing infections caused by *Salmonella enteritidis*. This suggests declining antibiotic effectiveness, necessitating public health interventions and revised treatment strategies.

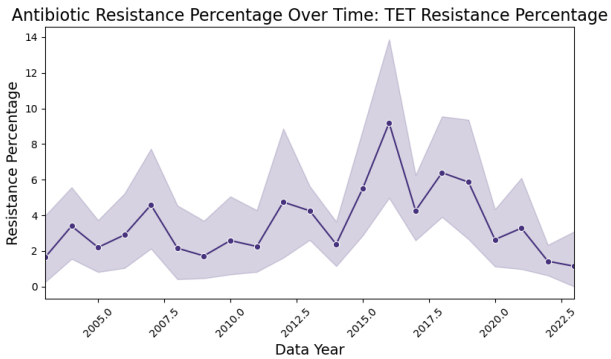


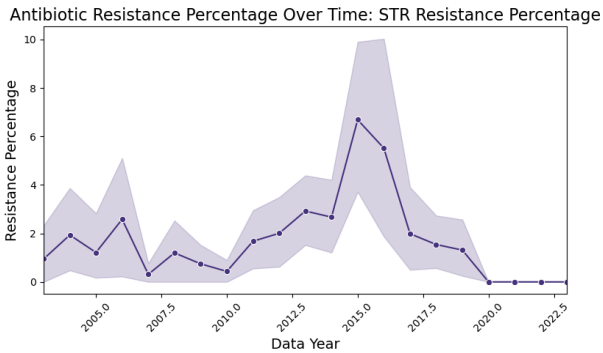
Variability Among Antibiotics

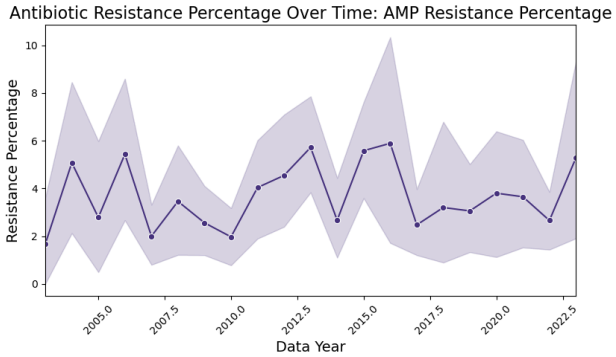
- Different antibiotics display distinct resistance patterns. Nalidixic Acid (NAL) shows the highest resistance percentages, followed by Tetracycline (TET) and Streptomycin (STR), which also exhibit increasing resistance.
- Other antibiotics such as Ampicillin (AMP), Ciprofloxacin (CIP), and Chloramphenicol (CHL) show sporadic resistance levels, indicating fluctuating effectiveness.

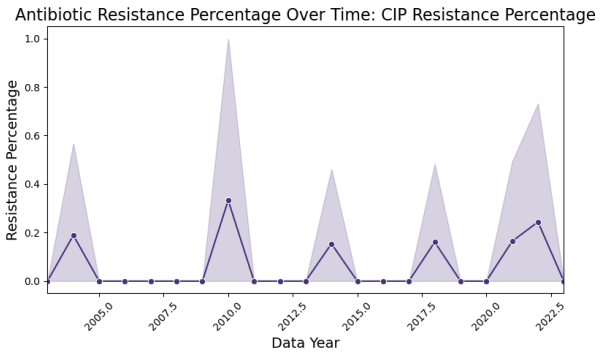












Comparison of Susceptible vs. Resistant

- The comparison reveals that while a majority of isolates remain susceptible, the rising trends in resistant counts for certain antibiotics are concerning. The narrowing gap between susceptible and resistant strains emphasizes the need for proactive measures to prevent further resistance development.

Consistent Patterns Across Antibiotics

- Many antibiotics show similar patterns of increasing resistance over time, highlighting a broader issue within the context of antibiotic use and management. The rising resistance percentages across multiple antibiotics suggest that the problem may not be isolated to specific drugs but reflects systemic challenges in combating antibiotic resistance.

Modeling

The exploratory data analysis (EDA) provided valuable insights into the patterns and trends associated with antibiotic resistance levels in *Salmonella enteritidis*. By examining various factors such as sample collection trends, demographic distributions, and the prevalence of antibiotic resistance, we gained a deeper understanding of the underlying issues contributing to this public health challenge.

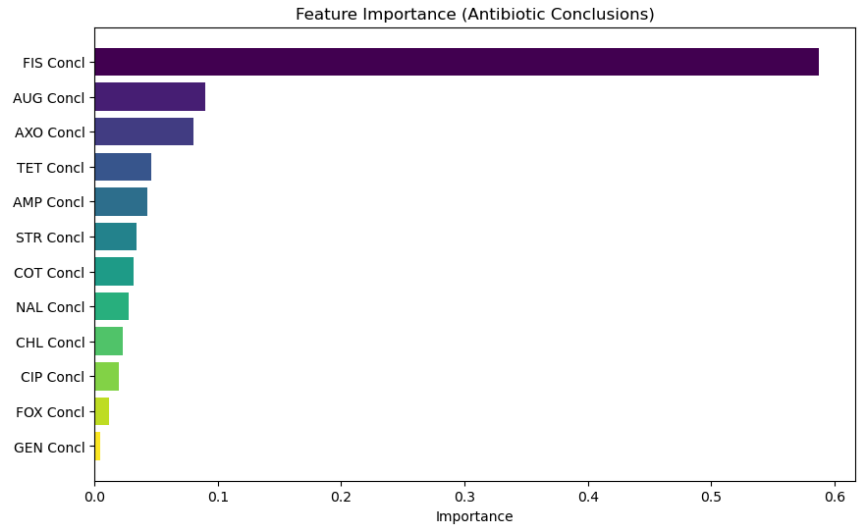
Building upon these insights, we proceeded to develop predictive models aimed at forecasting antibiotic resistance levels. The goal was to leverage the significant relationships identified during the EDA to create robust models that can inform future public health interventions and strategies.

Modeling Approaches

Random Forest Regressor

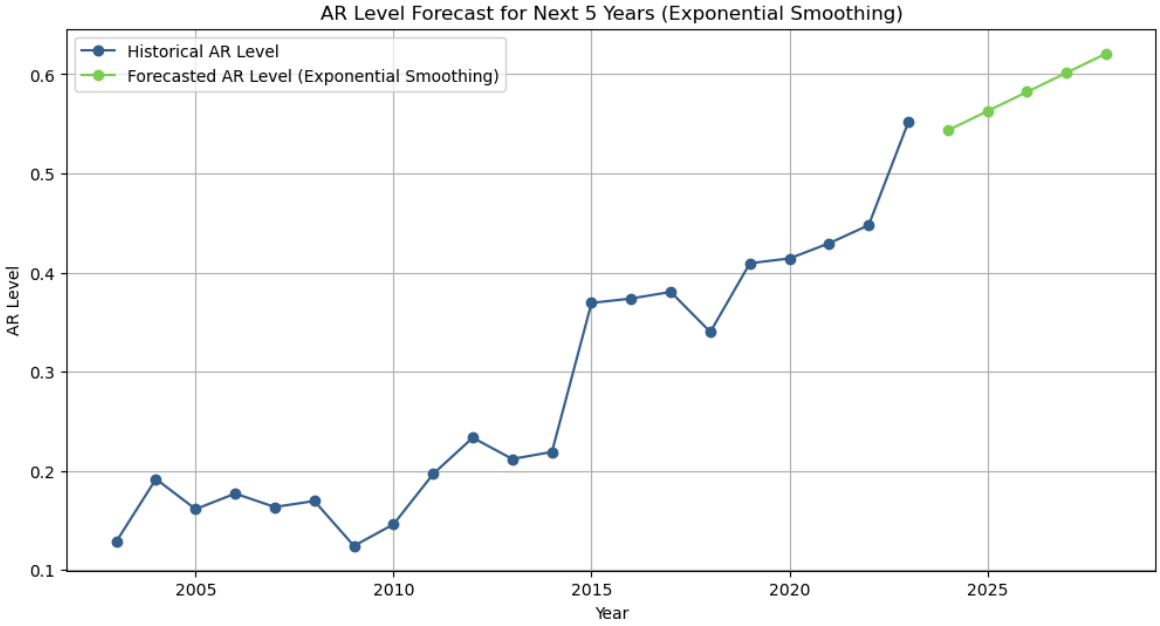
- The Random Forest model was trained on a subset of the data to predict AR Levels
- The relevant features and target variable were defined as follows:
    - X: The feature set includes the antibiotic conclusion columns
    - y: The target variable represents the AR Levels.
    - The dataset was split into training and testing sets, with 80% of the data used for training and 20% reserved for testing the model's performance

- Feature Importance Analysis
    - The analysis of feature importance highlighted the following influential features in predicting antibiotic resistance levels:
        - FIS Concl: 0.587
        - AUG Concl: 0.090
        - AXO Concl: 0.080
        - TET Concl: 0.046
        - AMP Concl: 0.043



Exponential Smoothing/Holt's Model

- This model was utilized to forecast AR Levels for the next five years, providing insights into potential future trends in antibiotic resistance. The forecasts indicated an upward trend in resistance levels, with the following key error metrics:



Modeling Results

The results from the modeling phase indicated that both the Random Forest Regressor and Exponential Smoothing models provided valuable insights into predicting antibiotic resistance levels. The Random Forest model demonstrated strong predictive capabilities, achieving a Mean Squared Error (MSE) of 0.0313 and an R-squared value of 0.9781. Feature importance analysis identified FIS Concl as a crucial factor influencing predictions.

Furthermore, the Exponential Smoothing/Holt's method forecasts indicated a projected increase in antibiotic resistance levels over the next five years. This trend underscores the importance of continuous monitoring and intervention efforts to effectively address rising resistance patterns.

Together, these models emphasize the complexity of antibiotic resistance dynamics and the need for a multifaceted approach to understanding and managing this public health issue.

Modeling Metrics

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Mean Absolute Percentage Error (MAPE)
Random Forest	0.0313	0.0384	0.1770	1.17%
Exponential Smoothing	0.0042	0.0538	0.0649	11.31%
Holt's Model	0.0042	0.0538	0.0649	11.31%

These metrics indicate that while the Random Forest model performs well with a low MSE and MAE, the Exponential Smoothing and Holt's models also show promise, particularly in forecasting trends, although they have a higher MAPE.

Conclusions

The comprehensive analysis of antibiotic resistance levels in *Salmonella enteritidis* revealed significant trends and patterns through both exploratory data analysis and modeling:

- Exploratory Data Analysis Insights: EDA highlighted variations in sample collection by year and region, suggesting potential shifts in infection rates or surveillance efforts. The dominance of stool specimens indicated common clinical practices for testing gastrointestinal pathogens. Additionally, the age group distribution revealed a higher incidence of infections in younger populations, necessitating targeted public health interventions. The prevalence of susceptible strains underscored the effectiveness of many antibiotics, while emerging resistance to specific antibiotics, particularly AMP, FIS, NAL, STR, and TET, raised concerns about future treatment options.

- Modeling Insights: The modeling phase demonstrated strong predictive capabilities, with the Random Forest Regressor achieving an R-squared value of 0.98 and a low Mean Squared Error (MSE) of 0.03. The feature importance analysis identified FIS Concl as a crucial factor influencing predictions. Furthermore, the Exponential Smoothing forecasts indicated a projected increase in antibiotic resistance levels over the next five years, emphasizing the urgent need for continuous monitoring and intervention to combat rising resistance patterns.

Overall, the integration of EDA and modeling results emphasizes the growing challenge of managing antibiotic resistance in *Salmonella enteritidis* and highlights the importance of ongoing surveillance and targeted strategies to mitigate this public health threat.

Future Directions

Incorporate Additional Features: Integrate more data attributes, such as geographical and demographic information related to Salmonella enteritidis cases. Understanding these factors, along with data on food outbreaks and salmonella-related hospitalizations, can enhance the model's predictive power and provide valuable insights into how environmental and social determinants impact resistance levels. This holistic approach can help identify at-risk populations and locations, allowing for more effective public health interventions.

Model Validation: Continuously validate the model against new data to ensure robustness and adaptability to changing resistance patterns. This is particularly important as antibiotic resistance evolves over time.

Explore Advanced Algorithms: Consider employing machine learning techniques that capture complex relationships within the data, such as Gradient Boosting or Neural Networks. These methods may provide better predictions for Salmonella enteritidis resistance.

Long-Term Trend Analysis: Investigate trends beyond 2028 using advanced forecasting methods to support public health planning. Monitoring antibiotic resistance trends in Salmonella enteritidis is crucial for developing targeted intervention strategies and informing policymakers on effective responses to outbreaks.