Understanding COVID-19 Vulnerabilities:
A Malaysian Data Analysis

01 CRISP-DM

02 Data Pre-processing

03 Exploratory Data Analysis

04 Modeling

05 Result Evaluation & Interpretation

01

CRISP-DM

Business Understanding

Background

The COVID-19 pandemic has brought unprecedented challenges to Malaysia, affecting the healthcare system, economy, and society at large. The vast amounts of data generated during the pandemic provide an opportunity to analyze the factors influencing case severity, vaccination effectiveness, and fatality rates. Understanding these factors is crucial for designing targeted health interventions and enabling informed policymaking. Despite significant vaccination efforts, questions remain about their effectiveness in reducing case severity and mortality rates. This project adopts a data-driven approach to uncover critical insights, optimize healthcare resource allocation, and improve future pandemic response readiness.

Key Stakeholders

- **Primary**: Ministry of Health Malaysia (KKM).
- **Secondary**: Healthcare providers, policymakers, and public health researchers.

Research Questions

1. What factors significantly influence COVID-19 outcomes in Malaysia?
2. Which demographics and regions are most vulnerable?
3. How do vaccination rates impact case severity and recovery rates?

Business Understanding

Business Objectives

- Build predictive models to identify critical factors influencing COVID-19 outcomes.
- Guide targeted health interventions and resource allocation.
- Evaluate vaccination impacts on reducing case severity.

Success Criteria

- Provide actionable insights for equitable healthcare resource distribution.
- Deliver predictive tools for effective pandemic management.

Data Understanding

1. **Daily COVID-19 Cases by Age Group & State**: https://data.gov.my/data-catalogue/covid_cases_age
2. **Daily COVID-19 Imported, Recovered, Active, and Cluster cases by State**: https://data.gov.my/data-catalogue/covid_cases
3. **Daily COVID-19 Cases by Vaccination Status & State**:https://data.gov.my/data-catalogue/covid_cases_vaxstatus
4. **Daily COVID-19 Vaccine Registrations by State**:https://data.gov.my/data-catalogue/vaxreg_covid
5. **Transactional Records: Deaths due to COVID-19**: https://data.gov.my/data-catalogue/covid_deaths_linelist

Data Understanding

Key Findings from Exploratory Data Analysis (EDA)

- **Temporal Trends**: Case spikes in August 2021 and March 2022.
- **Geographical Trends**: Selangor had the highest mortality rates.
- **Demographics**: Age group 18–39 was most affected.
- **Vaccination Impact**: Shifted from unvaccinated to vaccinated individuals post-2022 due to coverage.

Data Quality

- Issues with consistency and completeness were resolved through data cleaning.

Data Preparation

- Addressed missing values through imputation and standardization.
- Engineered features like vaccination effectiveness, recovery rate, and case fatality rate.
- Integrated primary and supplementary datasets.
- Performed 80-20 train-test split.

Modeling

- Random Forest and XGBoost used for severity prediction.
- Hyperparameter tuning using GridSearchCV improved performance.

Evaluation

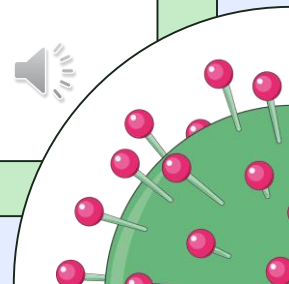
- Metrics: MAE, RMSE, R^2.
- Random Forest outperformed XGBoost slightly in predictive performance.

Deployment

- Integrate models into KKM systems for healthcare resource planning.
- Regular updates and monitoring established for continuous improvement.

02

Data Preprocessing



Data Preprocessing

Data Cleaning

- Impute missing value by filling ‘zero’ for death and vaccine registration record
- Remove outliers using IQR

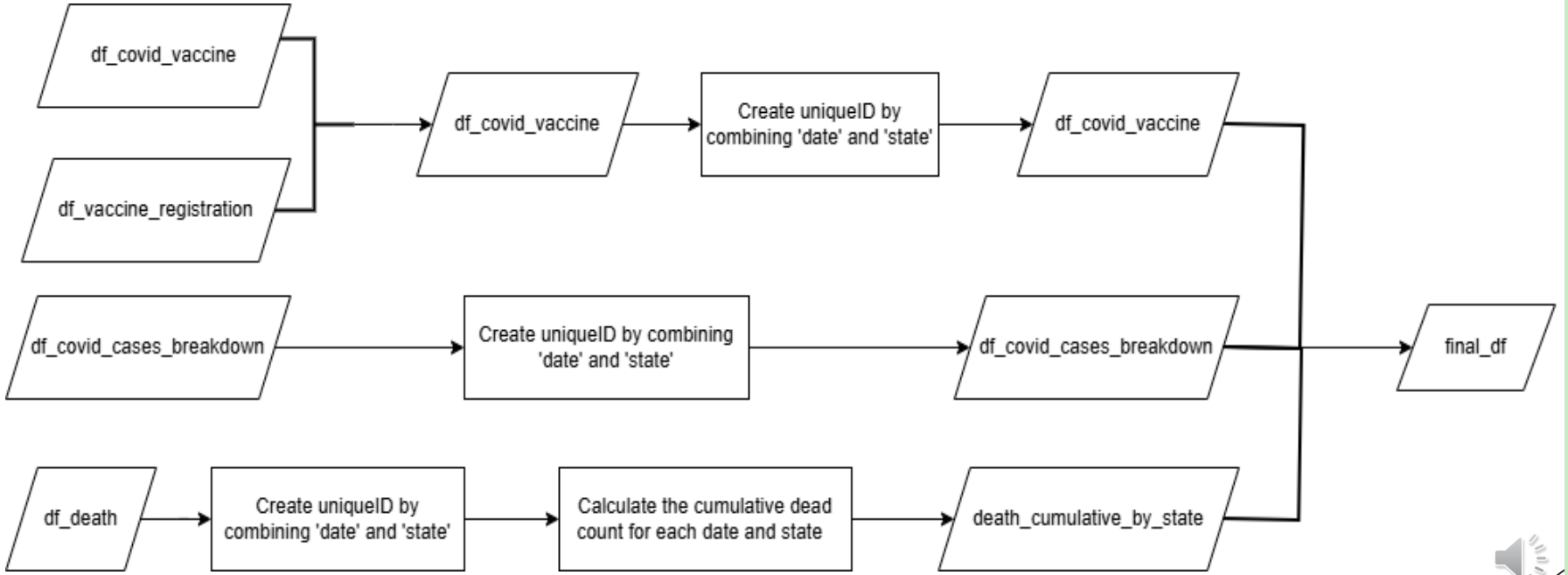
Data Transformation

- Pivot cases for each age group column
- Calculate durations of patients from testing positive to death

  days_to_death = date_death – date_tested
- Create ‘Gender’ column from ‘Male’ binary columns
- Assign age for each death case to respective age group

Data Preprocessing

Data Integration



Data Preprocessing

Feature Engineering

→ Create new variables such as:

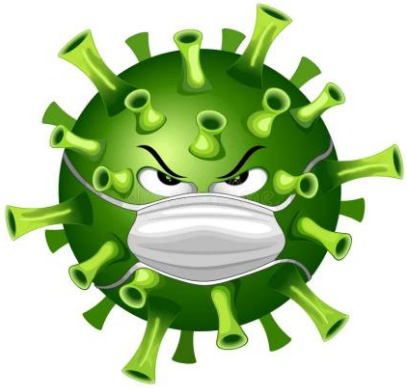
- **State_severity_index:** Combines deaths and active cases as a proportion of all cases.
- **Active_case_severity:** How severe active cases are compared to new cases.
- **Fatality_rate:** The proportion of deaths among all reported cases.
- **Vaccination_effectiveness:** Recovery rate versus unvaccinated rate.

Data Transformation

→ Features are normalized to values between 0 and 1 using **MinMaxScaler** to enable comparison of their relative importance.

Data Splitting

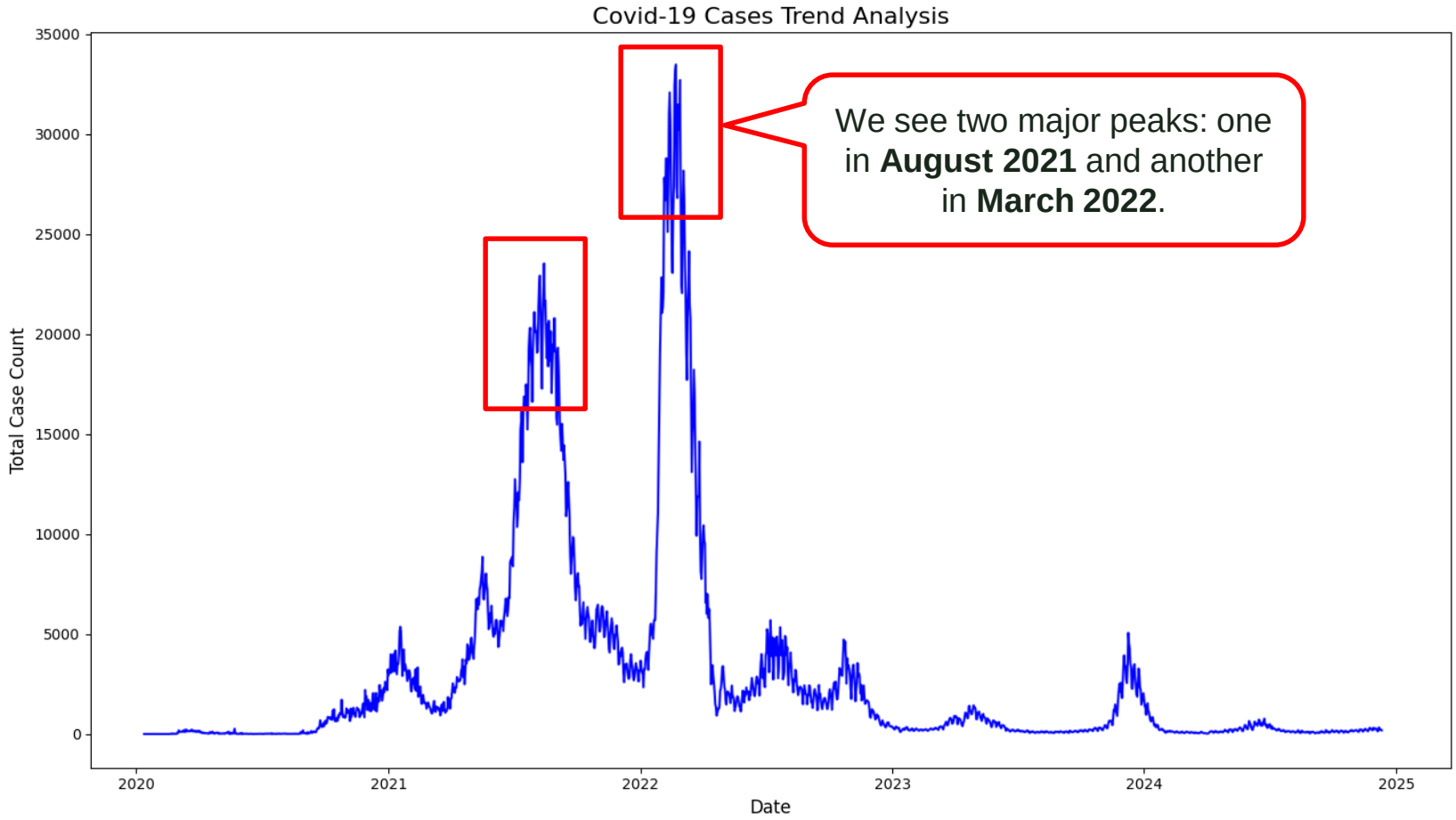
→ The cleaned dataset is split into a ratio of 80: 20 for training and testing purposes.



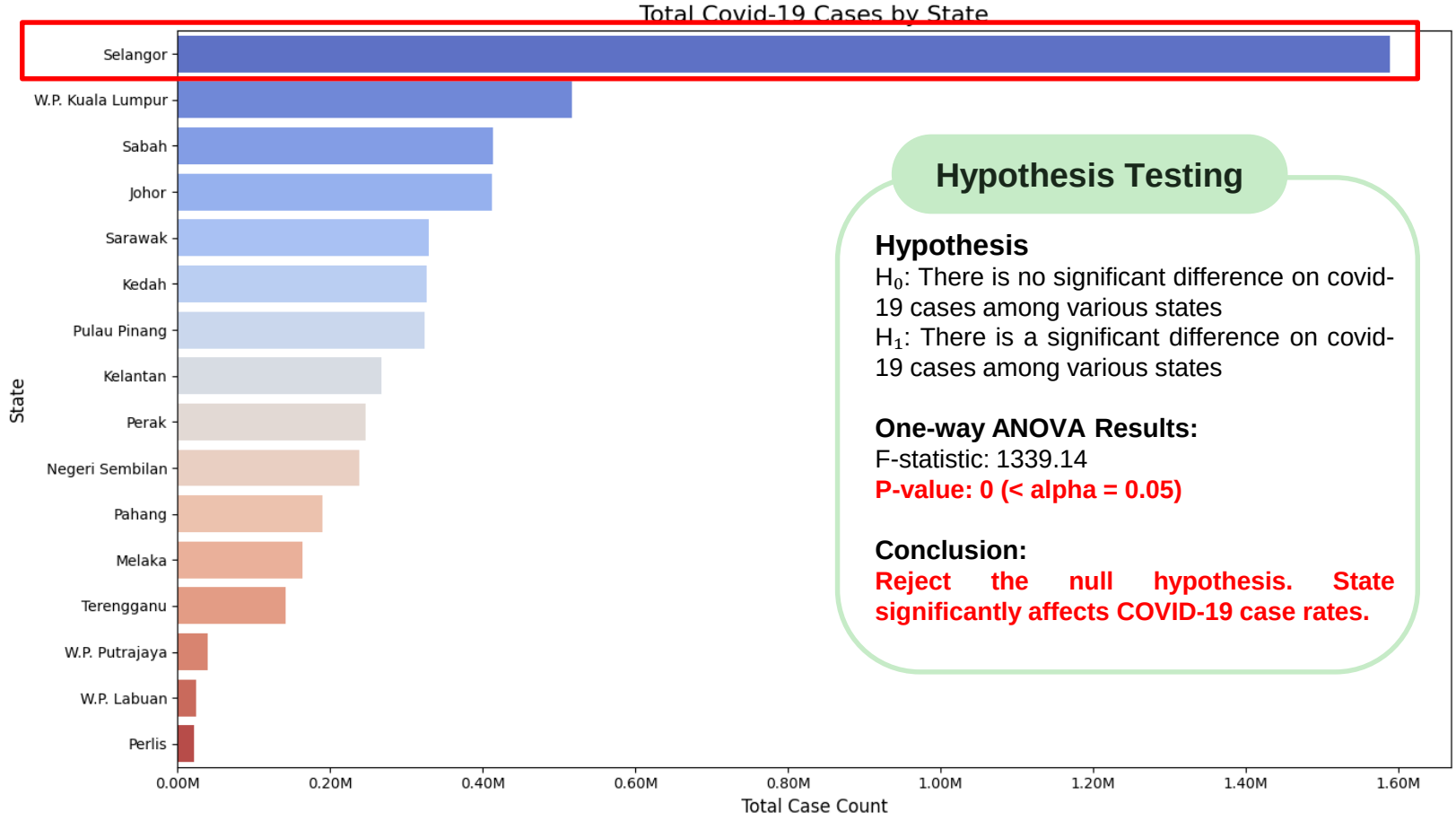
03 Exploratory Data Analysis

COVID-19 Outcomes Decoded: What the Data Says

Overall Covid-19 Cases Trend Analysis



Did Location Affect Covid-19 Cases?



Hypothesis Testing

Hypothesis

H_0: There is no significant difference on covid-19 cases among various states

H_1: There is a significant difference on covid-19 cases among various states

One-way ANOVA Results:

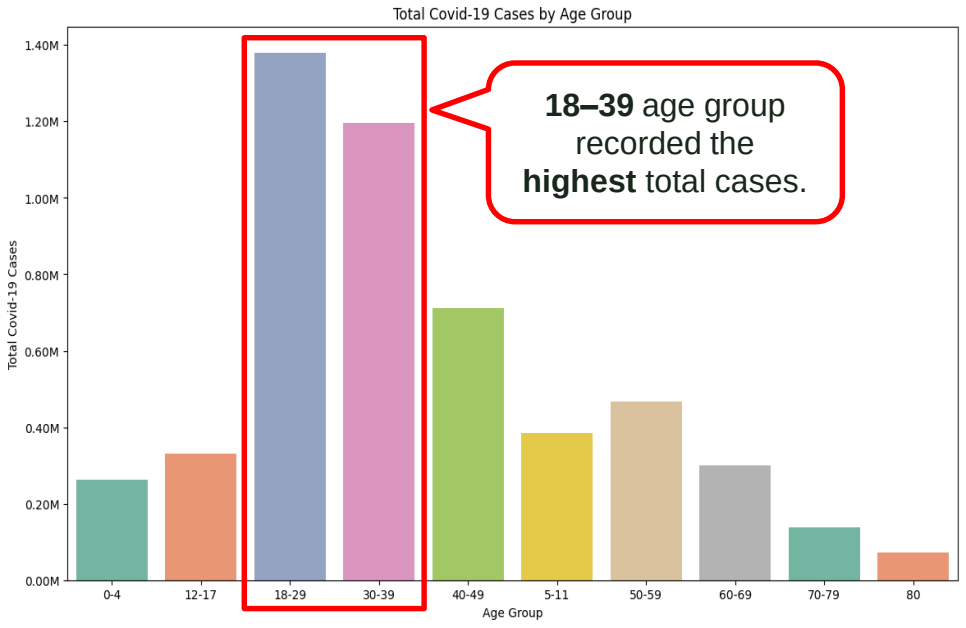
F-statistic: 1339.14

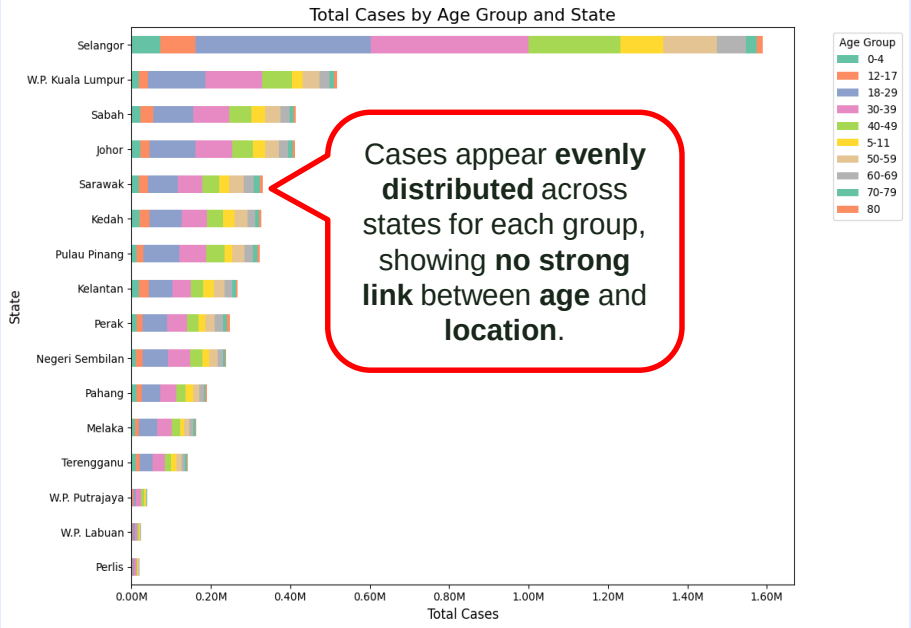
P-value: 0 (< alpha = 0.05)

Conclusion:

Reject the null hypothesis. State significantly affects COVID-19 case rates.

Which Groups are More Vulnerable to COVID-19?





Hypothesis Testing

Hypothesis

H_0: There is no significant difference on covid-19 cases among various age groups

H_1: There is a significant difference on covid-19 cases among various age groups

One-way ANOVA Results:

F-statistic: 1175.58

P-value: 0 (< alpha = 0.05)

Conclusion:

Reject the null hypothesis. Age group affect cases rate.

Which Groups are More Vulnerable to COVID-19?

Tukey HSD (Honestly Significant Difference) Results

1

group1	group2	meandiff	p-adj	lower	upper	reject
0-4	12-17	2.36	0.01	0.36	4.37	TRUE
0-4	18-29	38.92	0.00	36.91	40.92	TRUE
0-4	30-39	32.49	0.00	30.48	34.49	TRUE
0-4	40-49	15.68	0.00	13.68	17.69	TRUE
0-4	5-11	4.27	0.00	2.27	6.28	TRUE
0-4	50-59	7.15	0.00	5.15	9.16	TRUE
0-4	60-69	1.27	0.60	-0.74	3.27	FALSE
0-4	70-79	-4.38	0.00	-6.38	-2.37	TRUE
0-4	80	-6.64	0.00	-8.65	-4.64	TRUE
12-17	18-29	36.55	0.00	34.55	38.56	TRUE
12-17	30-39	30.12	0.00	28.12	32.13	TRUE
12-17	40-49	13.32	0.00	11.32	15.33	TRUE
12-17	5-11	1.91	0.08	-0.10	3.91	FALSE
12-17	50-59	4.79	0.00	2.78	6.79	TRUE
12-17	60-69	-1.09	0.78	-3.10	0.91	FALSE
12-17	70-79	-6.74	0.00	-8.75	-4.73	TRUE
12-17	80	-9.01	0.00	-11.01	-7.00	TRUE
18-29	30-39	-6.43	0.00	-8.43	-4.42	TRUE
18-29	40-49	-23.23	0.00	-25.24	-21.23	TRUE
18-29	5-11	-34.65	0.00	-36.65	-32.64	TRUE
18-29	50-59	-31.77	0.00	-33.77	-29.76	TRUE
18-29	60-69	-37.65	0.00	-39.65	-35.64	TRUE
18-29	70-79	-43.29	0.00	-45.30	-41.29	TRUE
18-29	80	-45.56	0.00	-47.56	-43.55	TRUE

2

2

1

group1	group2	meandiff	p-adj	lower	upper	reject
30-39	40-49	-16.80	0.00	-18.81	-14.80	TRUE
30-39	5-11	-28.22	0.00	-30.22	-26.21	TRUE
30-39	50-59	-25.34	0.00	-27.34	-23.33	TRUE
30-39	60-69	-31.22	0.00	-33.22	-29.21	TRUE
30-39	70-79	-36.86	0.00	-38.87	-34.86	TRUE
30-39	80	-39.13	0.00	-41.14	-37.12	TRUE
40-49	5-11	-11.41	0.00	-13.42	-9.41	TRUE
40-49	50-59	-8.53	0.00	-10.54	-6.53	TRUE
40-49	60-69	-14.42	0.00	-16.42	-12.41	TRUE
40-49	70-79	-20.06	0.00	-22.07	-18.06	TRUE
40-49	80	-22.33	0.00	-24.33	-20.32	TRUE
5-11	50-59	2.88	0.00	0.87	4.89	TRUE
5-11	60-69	-3.00	0.00	-5.01	-1.00	TRUE
5-11	70-79	-8.65	0.00	-10.65	-6.64	TRUE
5-11	80	-10.91	0.00	-12.92	-8.91	TRUE
50-59	60-69	-5.88	0.00	-7.89	-3.88	TRUE
50-59	70-79	-11.53	0.00	-13.53	-9.52	TRUE
50-59	80	-13.79	0.00	-15.80	-11.79	TRUE
			0.00	-7.65	-3.64	TRUE

Key Observations

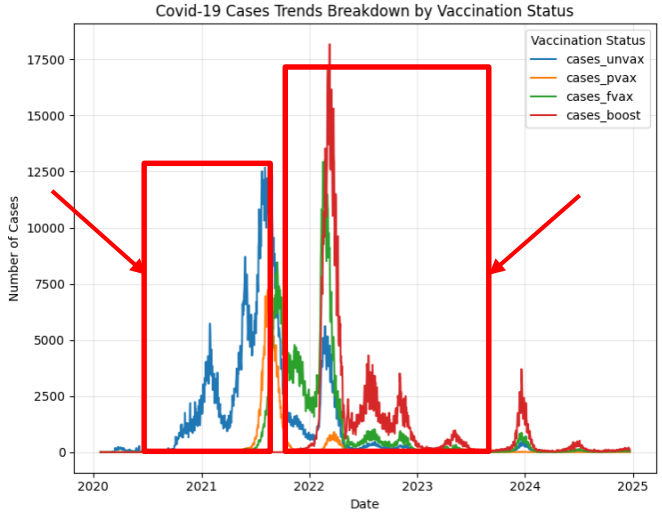
Significant Differences:

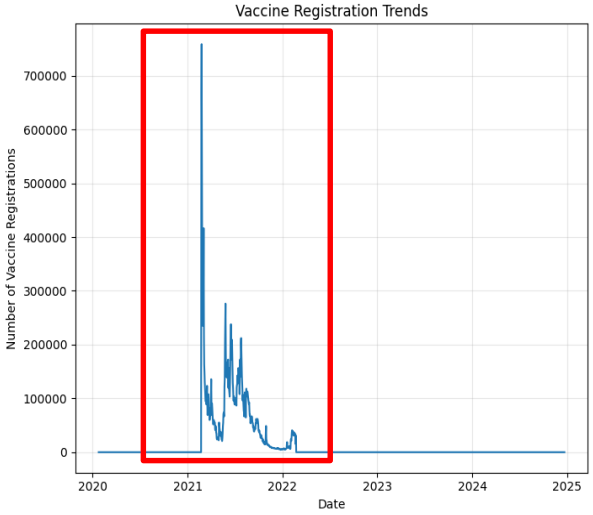
1. Most age group comparison have **p-value < 0.05**, which means they are **statistically significant.**

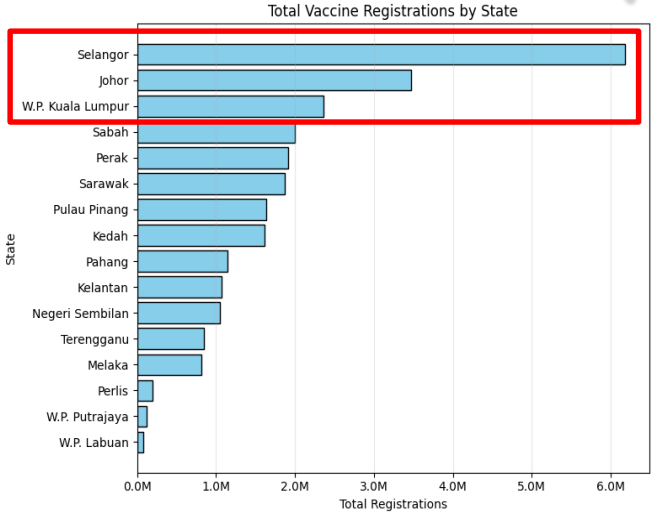
Non-significant differences:

2. 0-4 vs. 60-69, 12-17 vs. 5-11, 12-17 vs. 60-69

Can Vaccination Truly Change the Course of a Pandemic?







Hypothesis Testing

One-way ANOVA

H_0: There is no significant difference of vaccinations to covid-19 cases

H_1: There is a significant difference of vaccinations to covid-19 cases

Results:

F-statistic: 358.1

P-value: 1.67e-231 (< alpha 0.05)

Conclusion:

Reject the null hypothesis. There is a significant difference in COVID-19 cases among various vaccination status.

Note:
cases_unvax = unvaccinated
cases_pvax = partially vaccinated
cases_fvax = fully vaccinated
cases_boost = boosted

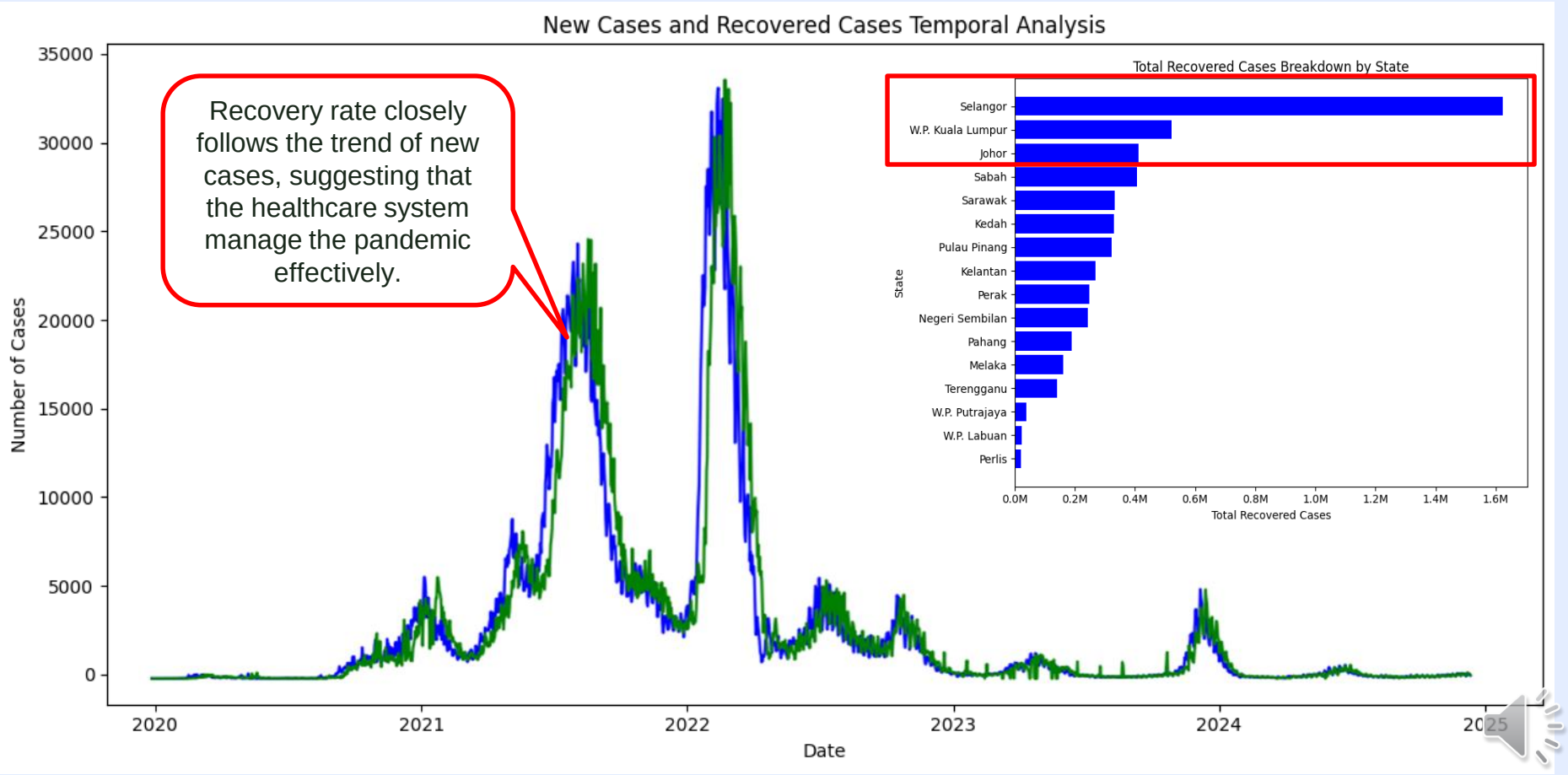
Tukey HSD (Honestly Significant Difference)

Multiple Comparison of Means - Tukey HSD, FWER=0.05

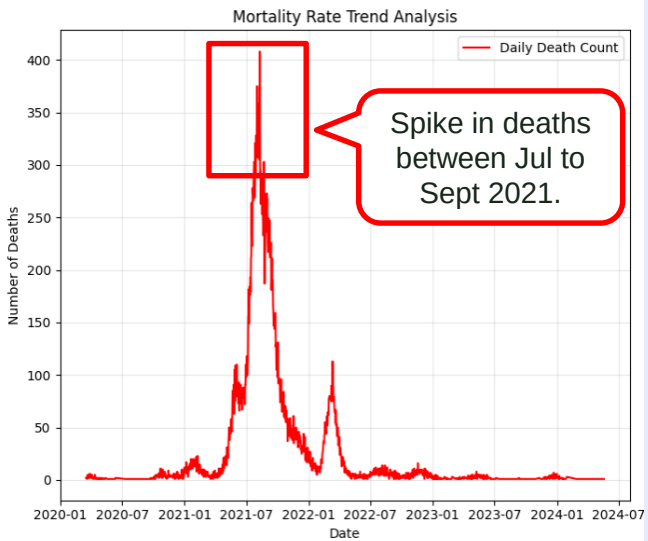
group1	group2	meandiff	p-adj	lower	upper	reject
cases_boost	cases_fvax	-0.0696	1.0	-4.5234	4.3842	False
cases_boost	cases_pvax	-34.391	0.0	-38.8448	-29.9372	True
cases_boost	cases_unvax	21.7295	0.0	17.2757	26.1833	True
cases_fvax	cases_pvax	-34.3214	0.0	-38.7752	-29.8676	True
cases_fvax	cases_unvax	21.7991	0.0	17.3453	26.2529	True
cases_pvax	cases_unvax	56.1205	0.0	51.6667	60.5743	True

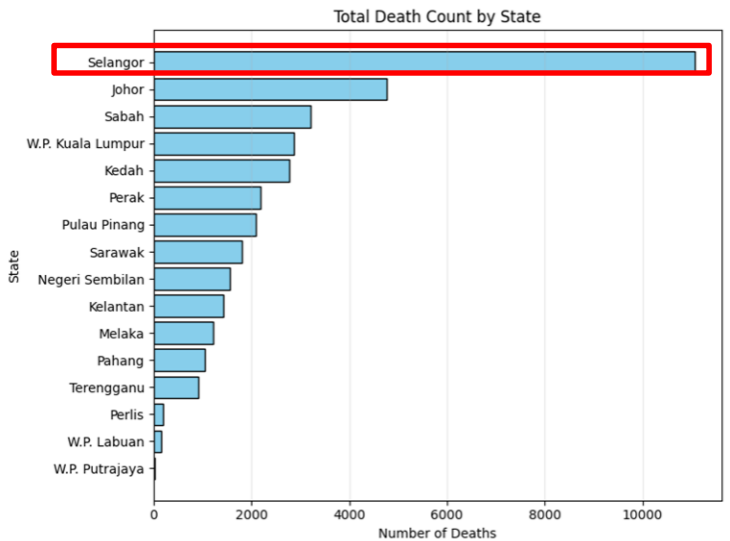
Conclusion: Getting vaccinated plays a key role in managing the pandemic.

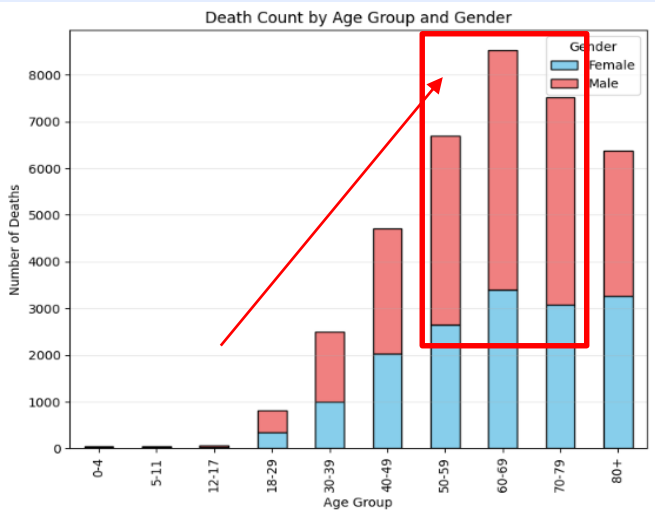
Did Most COVID-19 Cases End in Recovery?

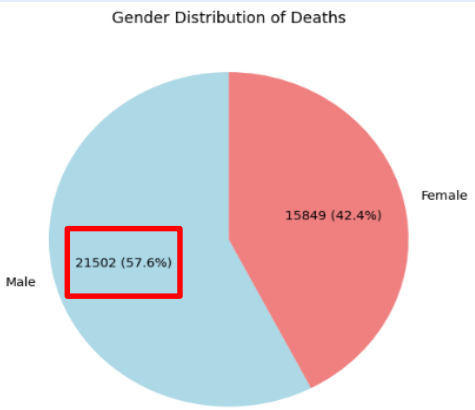


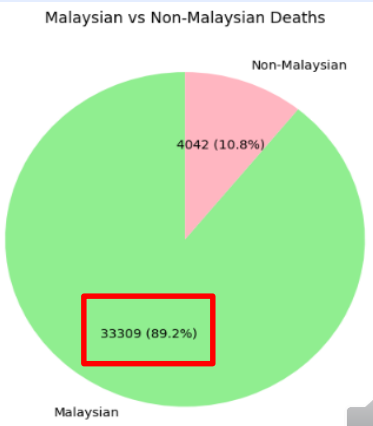
Covid-19 Mortality Rate in Malaysia



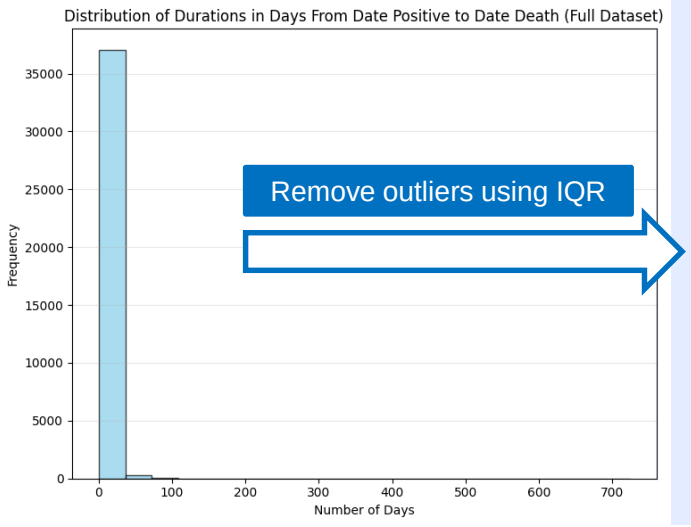




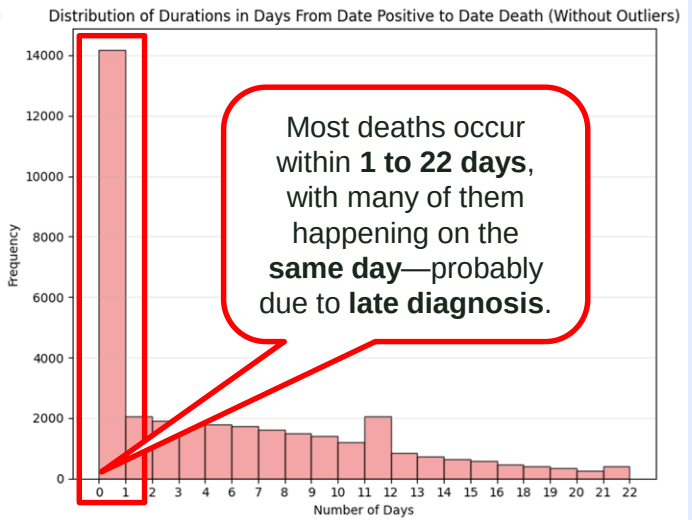




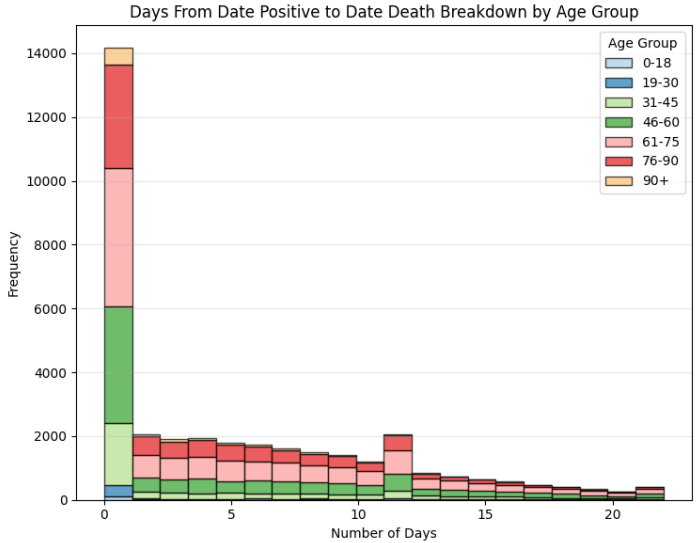
Have you ever wondered how long it takes for patients to go from testing positive to death?

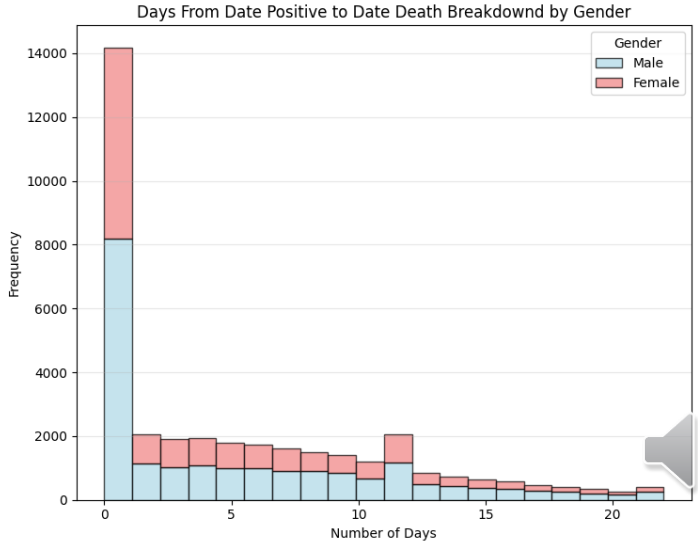


Remove outliers using IQR



Most deaths occur within **1 to 22 days**, with many of them happening on the **same day**—probably due to **late diagnosis**.





04

Modeling

Model Selection

Objective

- Identify **significant factors** influencing Covid-19 outcomes
- Predict **severity** or **resources needs**

Candidate Models

1. **Random Forest**
2. **XGBoost**

Reasons of model selection

- Excel in **feature-rich datasets** and provide **interpretability** (feature importance ranking)
- Support **regression** & can predict values in any **continuous range**

Model Training

Input of Prediction

→ Raw features:

- Location (states), new cases, recovered cases, active cases, cluster related cases, unvaccinated cases, partially vaccinated cases, fully vaccinated cases, death cases, boosted cases

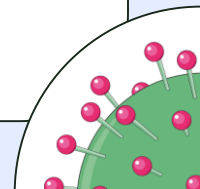
→ Engineered features:

- Vaccination effectiveness, state severity index, active case severity, case fatality rate

Output of Prediction

→ **Severity Prediction:**

- Predict the **risk score** of the specific region (states in Malaysia) for better resource allocation planning.



Risk Score

Proposed Risk Score Formula

$$\boldsymbol{Risk\ Score} = (w1 \times State\ Severity\ Index) + (w2 \times Active\ Cases\ Severity) + (w3 \times Fatality\ Rate) - (w4 \times Vaccination\ Effectiveness)$$

Weights are assigned as below:

- **w1 = 0.3**: Higher state severity index indicates higher risk for specific regions.
- **w2 = 0.3**: Active cases severity reflects higher case numbers.
- **w3 = 0.3**: Fatality rate shows disease severity.
- **w4 = 0.1**: Vaccination reduces risk, so it is subtracted, it has the lowest weightage as human will still have Covid-19 even though they are vaccinated.

Risk Score

Normalize Risk Score

Before:

- Min. Risk Score = -6.28212
- Max. Risk Score = 85.11177

After:

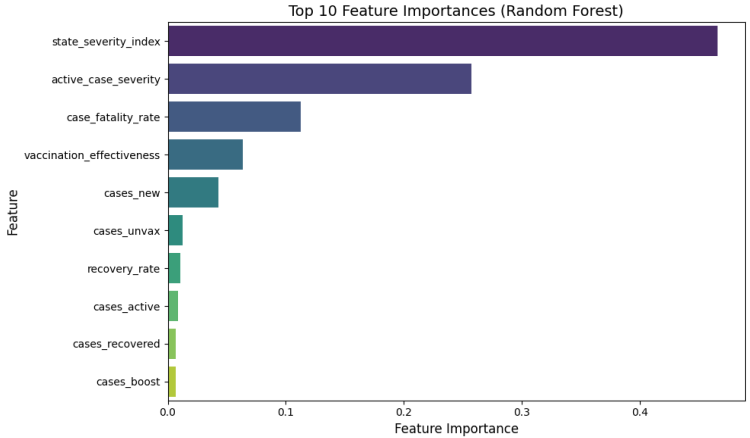
- Min. Risk Score = 0.0
- Max. Risk Score = 100.0

05

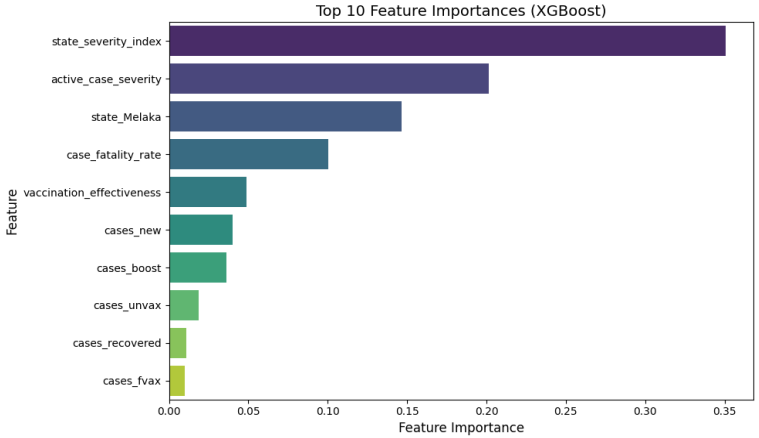
Results Evaluation & Interpretation

Feature Importance

Random Forest



XGBoost



Model Hyperparameter Tuning

GridSearchCV

108	3	324
Param combinations	Folds	Total fits per model

Random Forest

Best Parameters

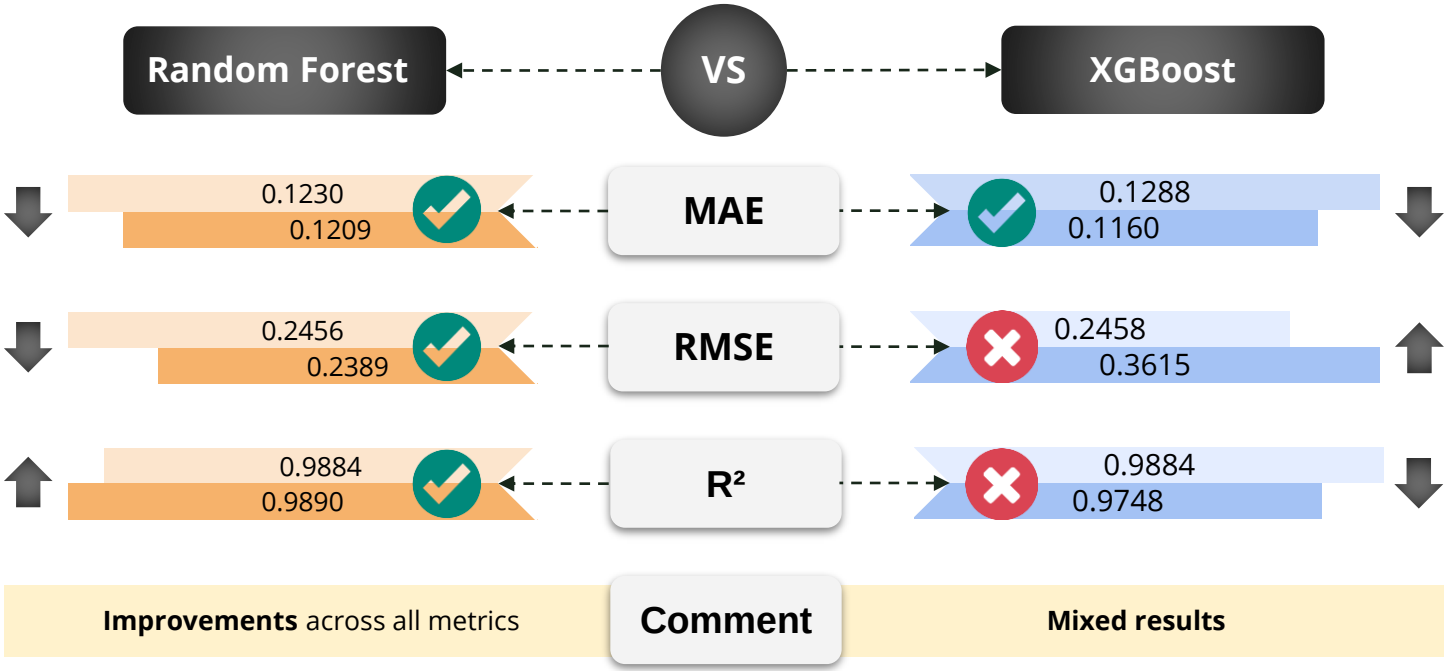
- Number of Trees (n_estimators): 100
- Max Depth (max_depth): 20
- Min Samples per Leaf (min_samples_leaf): 1
- Min Samples per Split (min_samples_split): 2

XGBoost

Best Parameters

- Number of Trees (n_estimators): 200
- Max Depth (max_depth): 5
- Learning Rate (learning_rate): 0.1
- Fraction of samples (subsample): 1.0
- Fraction of columns (colsample_bytree): 1.0

Model Performance Comparison



Random Forest is recommended as it demonstrates more **stable and consistent performance** across all metrics.

Thanks!