Milestone 3: Evaluation, Interpretation, Tool Development, and Presentation

- ## Member:

Yuexi Lyu (67468519)

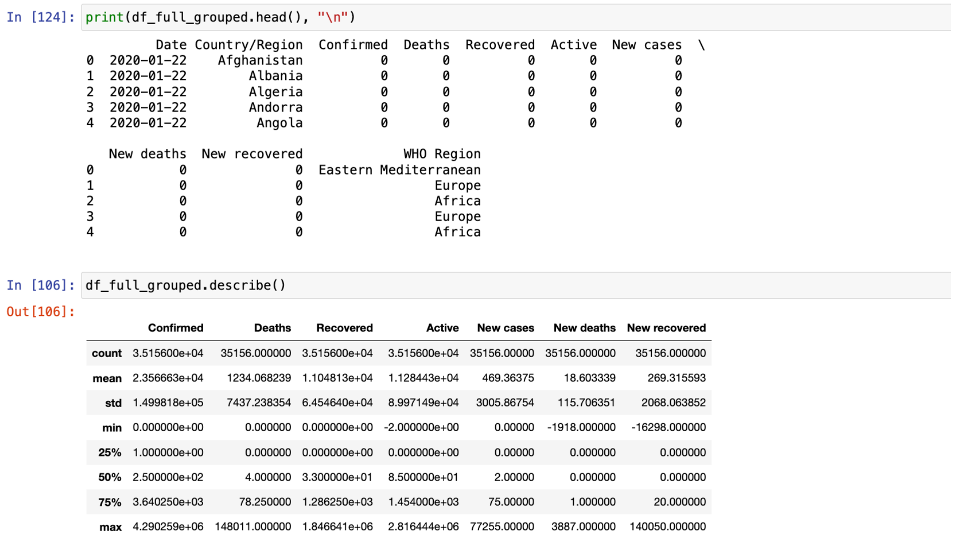
- ## Objective of the project

The main goal of this project is to analyze global COVID-19 epidemic data and build a machine learning model to predict patient recovery rates. In addition, an interactive web dashboard built using Streamlit was developed to visualize the global epidemic data and forecast COVID-19 recovery-related indicators worldwide.

- ## Data to be used.

The following datasets have been utilized for this analysis:

Full Grouped Dataset(full_grouped.csv): Time-series data of confirmed cases, recoveries, deaths, and active cases worldwide.

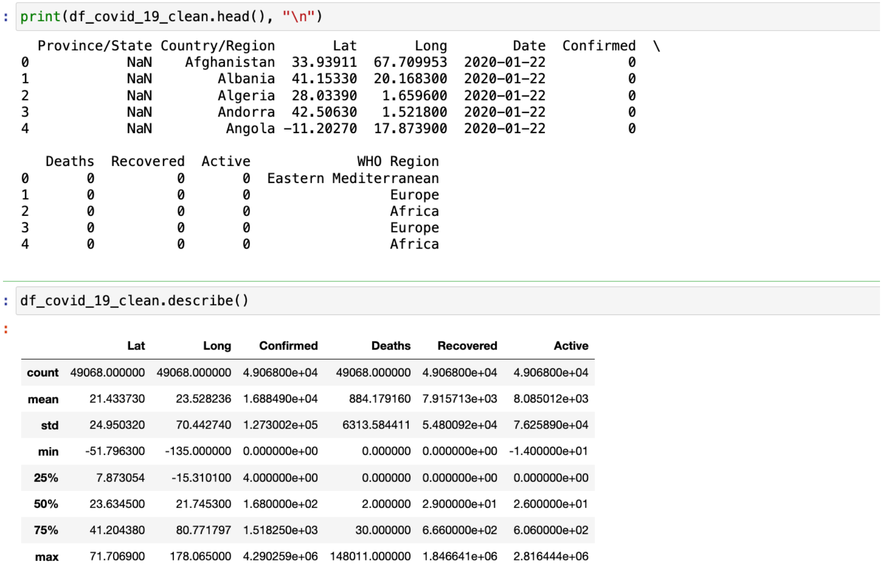


```
In [124]: print(df_full_grouped.head(), "\n")

                Date Country/Region  Confirmed  Deaths  Recovered  Active  New cases  \
          0  2020-01-22    Afghanistan          0       0          0       0          0
          1  2020-01-22        Albania          0       0          0       0          0
          2  2020-01-22        Algeria          0       0          0       0          0
          3  2020-01-22        Andorra          0       0          0       0          0
          4  2020-01-22         Angola          0       0          0       0          0

             New deaths  New recovered             WHO Region
          0           0              0  Eastern Mediterranean
          1           0              0                 Europe
          2           0              0                 Africa
          3           0              0                 Europe
          4           0              0                 Africa
```

```
In [106]: df_full_grouped.describe()
Out[106]:
```

	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered
count	3.515600e+04	35156.000000	3.515600e+04	3.515600e+04	35156.00000	35156.000000	35156.000000
mean	2.356663e+04	1234.068239	1.104813e+04	1.128443e+04	469.36375	18.603339	269.315593
std	1.499818e+05	7437.238354	6.454640e+04	8.997149e+04	3005.86754	115.706351	2068.063852
min	0.000000e+00	0.000000	0.000000e+00	-2.000000e+00	0.00000	-1918.000000	-16298.000000
25%	1.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.00000	0.000000	0.000000
50%	2.500000e+02	4.000000	3.300000e+01	8.500000e+01	2.00000	0.000000	0.000000
75%	3.640250e+03	78.250000	1.286250e+03	1.454000e+03	75.00000	1.000000	20.000000
max	4.290259e+06	148011.000000	1.846641e+06	2.816444e+06	77255.00000	3887.000000	140050.000000

COVID-19 Clean Complete Dataset(covid_19_clean_complete.csv): Comprehensive dataset including country-level statistics with additional geographic information (latitude, longitude).



```
: print(df_covid_19_clean.head(), "\n")
```

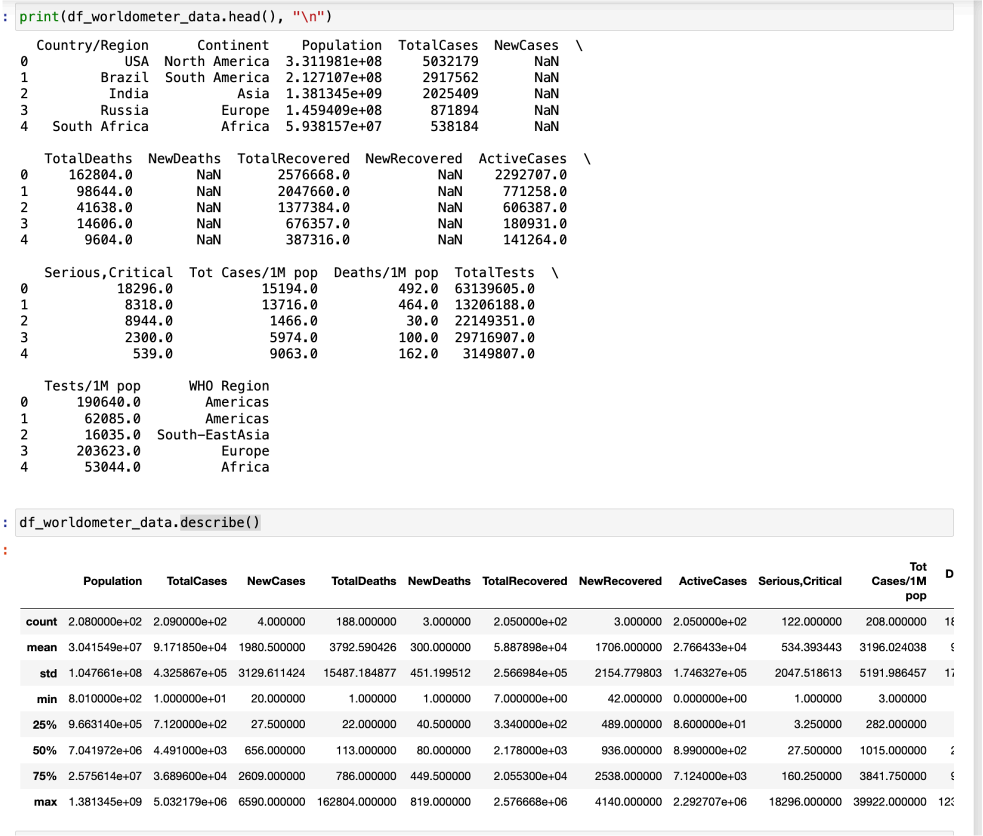
```
  Province/State Country/Region       Lat       Long        Date  Confirmed  \
0            NaN    Afghanistan  33.93911  67.709953  2020-01-22          0   
1            NaN        Albania  41.15330  20.168300  2020-01-22          0   
2            NaN        Algeria  28.03390   1.659600  2020-01-22          0   
3            NaN        Andorra  42.50630   1.521800  2020-01-22          0   
4            NaN         Angola -11.20270  17.873900  2020-01-22          0   

   Deaths  Recovered  Active             WHO Region  
0       0          0       0  Eastern Mediterranean  
1       0          0       0                 Europe  
2       0          0       0                 Africa  
3       0          0       0                 Europe  
4       0          0       0                 Africa  
```

```
: df_covid_19_clean.describe()
```

:

	Lat	Long	Confirmed	Deaths	Recovered	Active
count	49068.000000	49068.000000	4.906800e+04	49068.000000	4.906800e+04	4.906800e+04
mean	21.433730	23.528236	1.688490e+04	884.179160	7.915713e+03	8.085012e+03
std	24.950320	70.442740	1.273002e+05	6313.584411	5.480092e+04	7.625890e+04
min	-51.796300	-135.000000	0.000000e+00	0.000000	0.000000e+00	-1.400000e+01
25%	7.873054	-15.310100	4.000000e+00	0.000000	0.000000e+00	0.000000e+00
50%	23.634500	21.745300	1.680000e+02	2.000000	2.900000e+01	2.600000e+01
75%	41.204380	80.771797	1.518250e+03	30.000000	6.660000e+02	6.060000e+02
max	71.706900	178.065000	4.290259e+06	148011.000000	1.846641e+06	2.816444e+06

Worldometer(worldometer_data.csv) Dataset: A snapshot of country-level statistics such as total cases, deaths, recovered cases, active cases, and per capita statistics.



```
: print(df_worldometer_data.head(), "\n")
```

```
  Country/Region      Continent    Population  TotalCases  NewCases  \
0            USA  North America  3.311981e+08     5032179       NaN   
1         Brazil  South America  2.127107e+08     2917562       NaN   
2          India           Asia  1.381345e+09     2025409       NaN   
3         Russia         Europe  1.459409e+08      871894       NaN   
4   South Africa         Africa  5.938157e+07      538184       NaN   

   TotalDeaths  NewDeaths  TotalRecovered  NewRecovered  ActiveCases  \
0     162804.0        NaN       2576668.0           NaN    2292707.0   
1      98644.0        NaN       2047660.0           NaN     771258.0   
2      41638.0        NaN       1377384.0           NaN     606387.0   
3      14606.0        NaN        676357.0           NaN     180931.0   
4       9604.0        NaN        387316.0           NaN     141264.0   

   Serious,Critical  Tot Cases/1M pop  Deaths/1M pop  TotalTests  \
0           18296.0           15194.0          492.0  63139605.0   
1            8318.0           13716.0          464.0  13206188.0   
2            8944.0            1466.0           30.0  22149351.0   
3            2300.0            5974.0          100.0  29716907.0   
4             539.0            9063.0          162.0   3149807.0   

   Tests/1M pop       WHO Region  
0      190640.0         Americas  
1       62085.0         Americas  
2       16035.0  South-EastAsia  
3      203623.0           Europe  
4       53044.0           Africa  
```

```
: df_worldometer_data.describe()
```

:

	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	D
count	2.080000e+02	2.090000e+02	4.000000	188.000000	3.000000	2.050000e+02	3.000000	2.050000e+02	122.000000	208.000000	18
mean	3.041549e+07	9.171850e+04	1980.500000	3792.590426	300.000000	5.887898e+04	1706.000000	2.766433e+04	534.393443	3196.024038	9
std	1.047661e+08	4.325867e+05	3129.611424	15487.184877	451.199512	2.566984e+05	2154.779803	1.746327e+05	2047.518613	5191.986457	17
min	8.010000e+02	1.000000e+01	20.000000	1.000000	1.000000	7.000000e+00	42.000000	0.000000e+00	1.000000	3.000000	
25%	9.663140e+05	7.120000e+02	27.500000	22.000000	40.500000	3.340000e+02	489.000000	8.600000e+01	3.250000	282.000000	
50%	7.041972e+06	4.491000e+03	656.000000	113.000000	80.000000	2.178000e+03	936.000000	8.990000e+02	27.500000	1015.000000	2
75%	2.575614e+07	3.689600e+04	2609.000000	786.000000	449.500000	2.055300e+04	2538.000000	7.124000e+03	160.250000	3841.750000	9
max	1.381345e+09	5.032179e+06	6590.000000	162804.000000	819.000000	2.576668e+06	4140.000000	2.292707e+06	18296.000000	39922.000000	123

- ## Why Streamlit?

For this project, Streamlit was chosen because:
It allows for quick deployment of interactive dashboards directly from Python scripts, and for analysis directly from my data after data analysis and feature engineering.
It supports advanced data visualization using Plotly and Matplotlib.
It integrates seamlessly with popular data science tools such as Pandas, Scikit-learn, and NumPy.

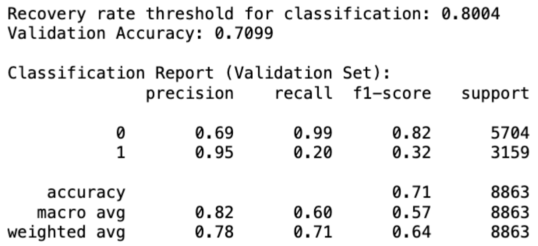
● Tech stack

Programming languages and libraries: Python, Pandas, NumPy, Matplotlib, Seaborn, Dash, Scipy, Scikit-learn, etc.
Data processing and storage: Jupyter Notebook, CSV.
Visualization and reporting: Matplotlib, Seaborn, Plotly, etc.
Machine learning and prediction: train_test_split, LinearRegression, PolynomialFeatures, MLPClassifier, LabelEncoder, StandardScaler, Model Evaluation Metrics and so on.
Tool display: streamlit.

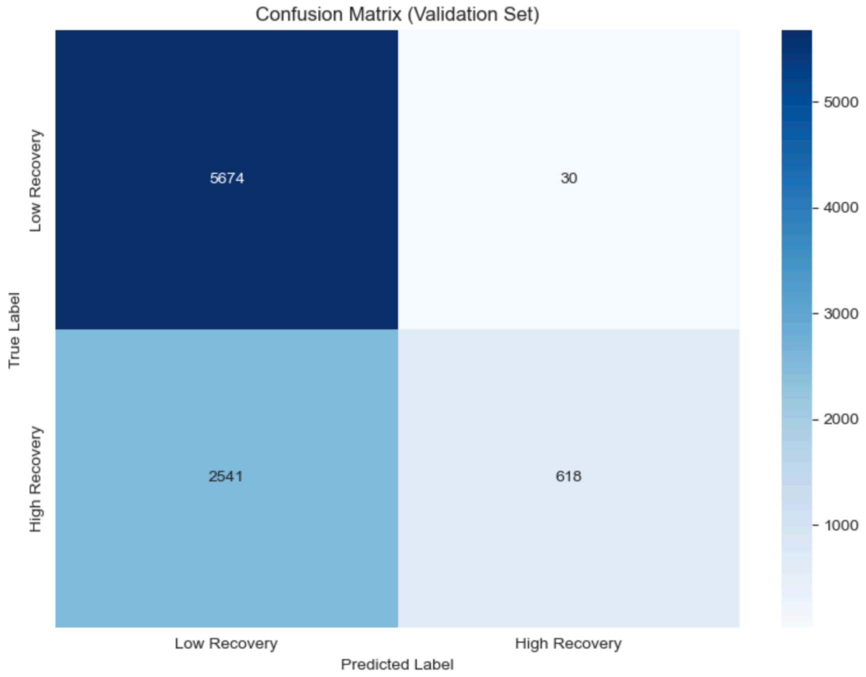
● Evaluate the performance of the model on the test set using the same metrics as the training set.

(1) Logistic Regression:

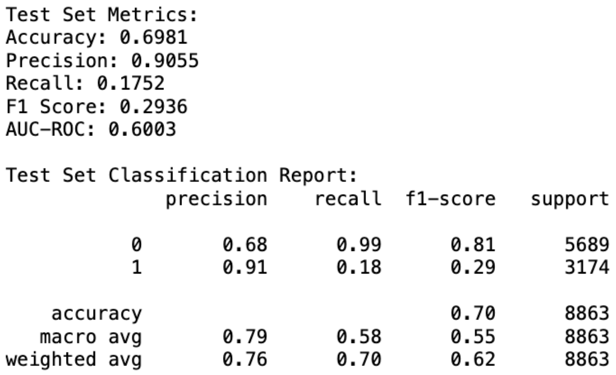
To predict the COVID-19 cure rate, I implemented a logistic regression model by converting the continuous recovery rate into a binary variable, using the median of the training set (0.8004) as the threshold: values above were labeled as high (1), and below as low (0). All features were normalized to ensure consistent scaling.
The model achieved 70.99% accuracy on the validation set. While it identified low recovery rates well (recall = 0.99), it struggled with high recovery rate cases (recall = 0.20), as shown in the confusion matrix:
True Negatives: 5674, true Positives: 618, False Negatives: 2541
This imbalance highlights the limitations of logistic regression in handling skewed classes in recovery rate prediction.



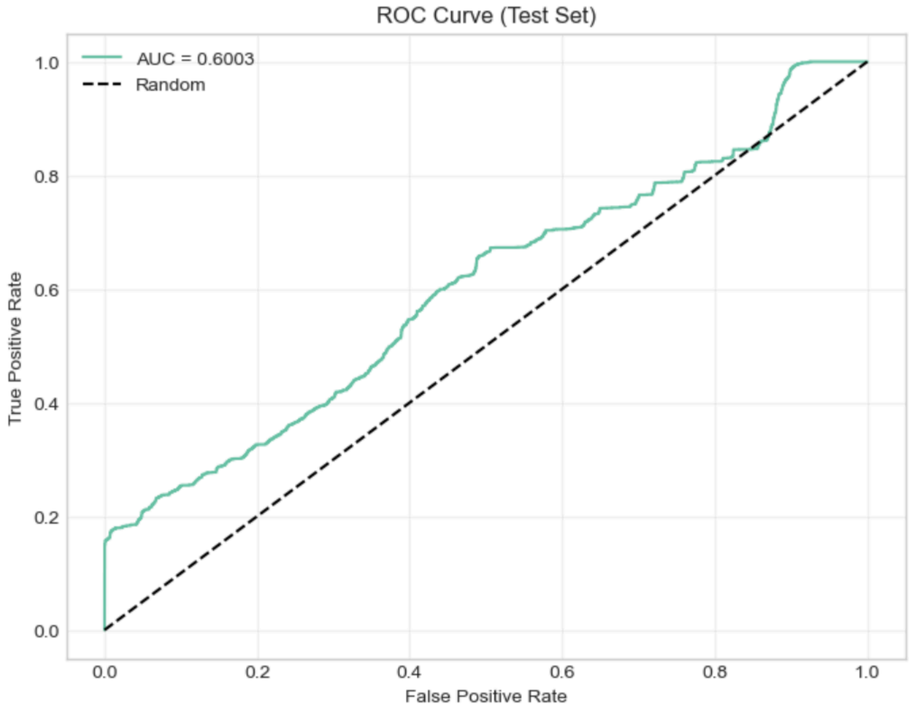
```
Recovery rate threshold for classification: 0.8004
Validation Accuracy: 0.7099

Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.69      0.99      0.82      5704
           1       0.95      0.20      0.32      3159

    accuracy                           0.71      8863
   macro avg       0.82      0.60      0.57      8863
weighted avg       0.78      0.71      0.64      8863
```



A comprehensive evaluation on the test set revealed clear limitations of the logistic regression model. Despite achieving 69.81% accuracy and a high precision of 90.55%, the model suffered from a low recall of 17.52%, leading to a low F1 score of 29.36%. This indicates it frequently missed high recovery rate cases. The AUC-ROC was only 0.6003, close to random guessing, as reflected by the ROC curve near the diagonal. The classification report confirmed this imbalance: while 99% of low recovery rate samples were correctly identified, only 18% of high recovery rate cases were detected. These results suggest that logistic regression is insufficient for capturing the complex patterns influencing COVID-19 recovery, and more advanced models or enhanced feature engineering may be required.



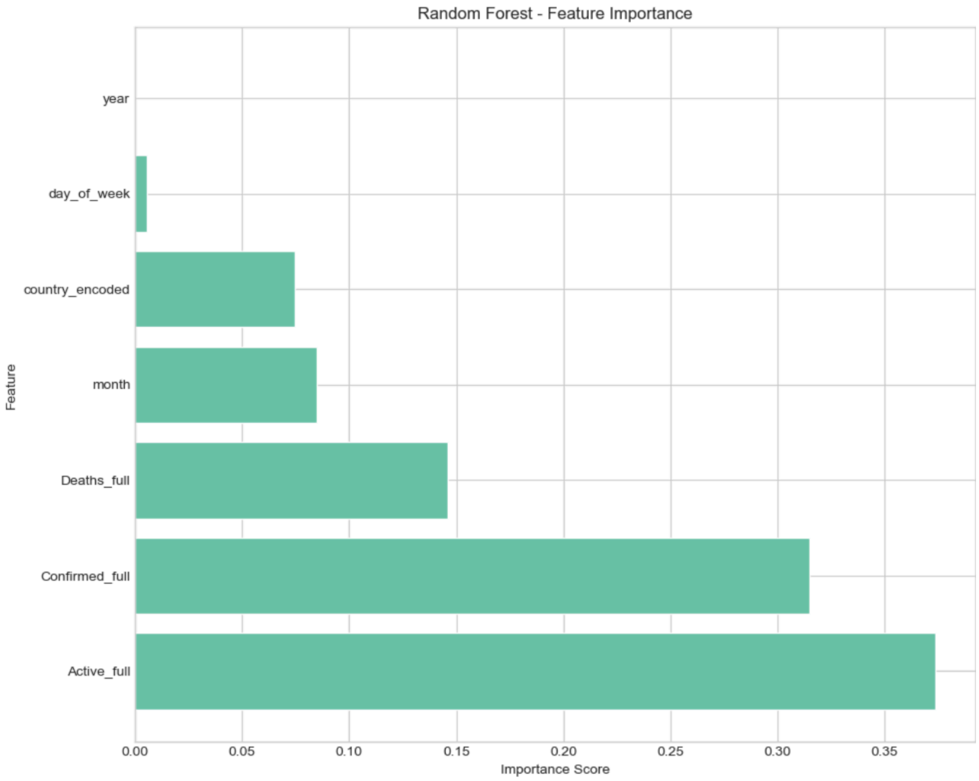
```
Test Set Metrics:
Accuracy: 0.6981
Precision: 0.9055
Recall: 0.1752
F1 Score: 0.2936
AUC-ROC: 0.6003

Test Set Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.99      0.81      5689
           1       0.91      0.18      0.29      3174

    accuracy                           0.70      8863
   macro avg       0.79      0.58      0.55      8863
weighted avg       0.76      0.70      0.62      8863
```



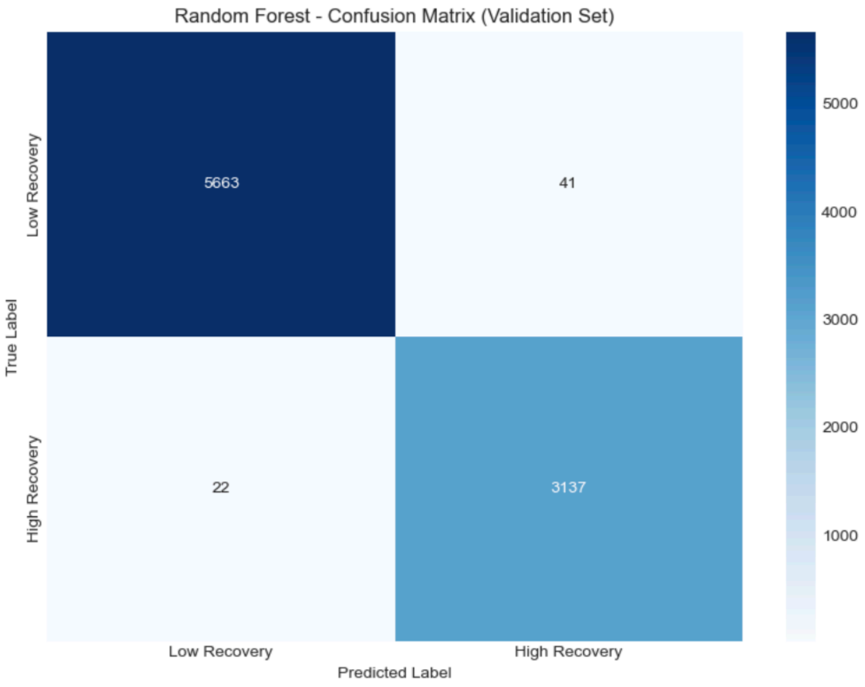
(2) Random Forest

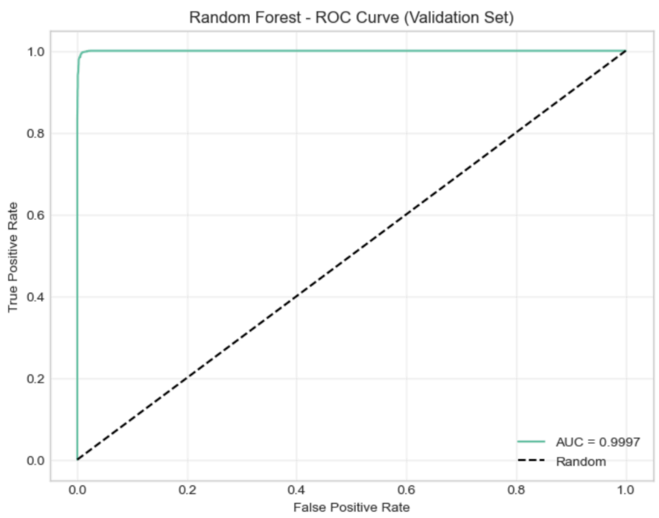
The random forest model was used to classify and predict the COVID-19 recovery rate. The model was trained using multiple features, including the number of active cases, cumulative confirmed cases, number of deaths, country code, and time-related variables. The results showed that the model performed exceptionally well on the validation set, achieving an accuracy of 99.29%. The confusion matrix indicated a very low classification error, while the ROC and PR curves further verified the model’s robustness in handling imbalanced data. Overall, the model demonstrated extremely high predictive accuracy and strong practical applicability.

```
Recovery rate threshold for classification: 0.8004
Out-of-bag score: 0.9946
```

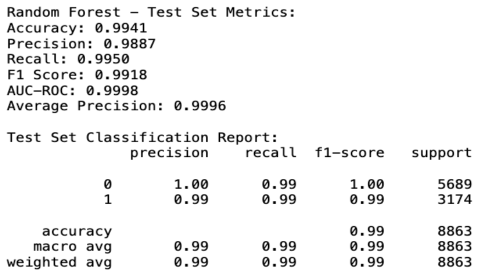


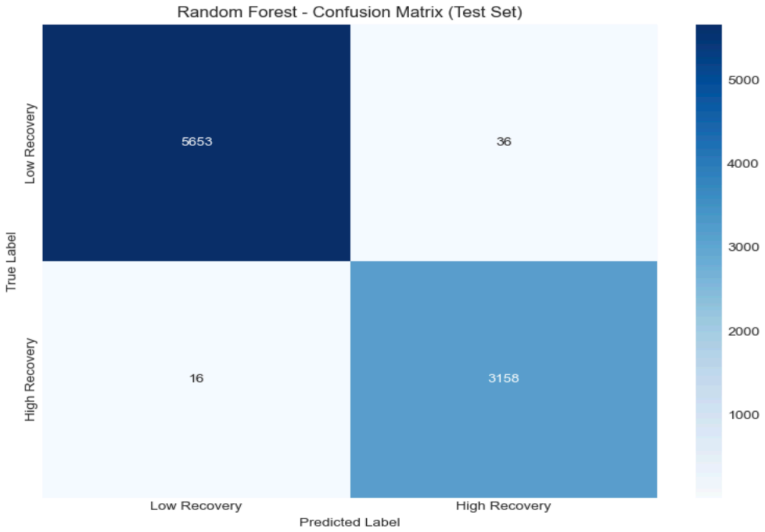
```
Validation Set Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      5704
           1       0.99      0.99      0.99      3159

    accuracy                           0.99      8863
   macro avg       0.99      0.99      0.99      8863
weighted avg       0.99      0.99      0.99      8863
```

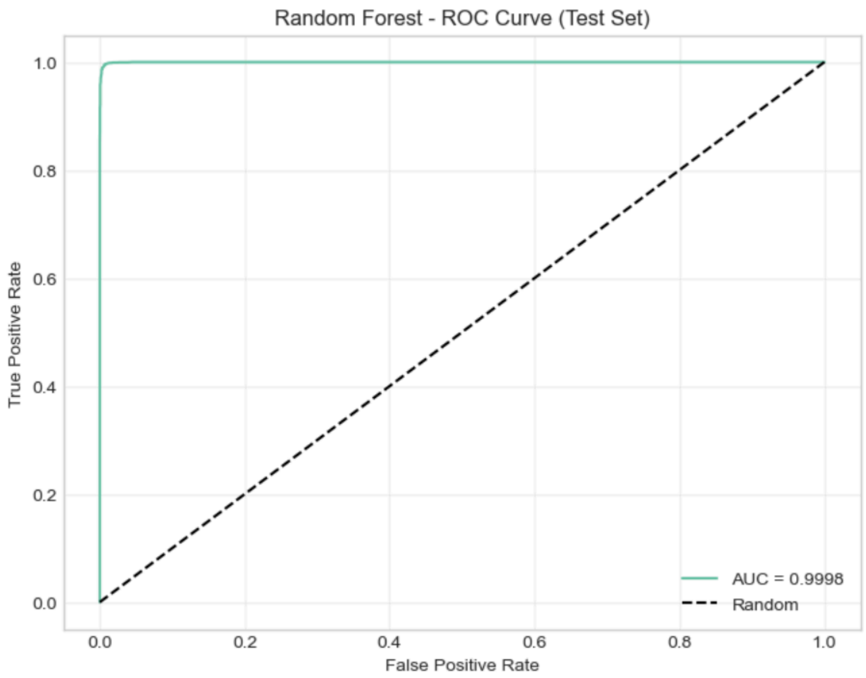




According to the results on the test set, the random forest model demonstrated outstanding performance in the COVID-19 recovery rate classification task. The overall performance was highly consistent with that on the validation set, indicating that the model had strong generalization ability. The confusion matrix further confirmed this conclusion: only 36 low-recovery samples were misclassified as high-recovery, and 16 high-recovery samples were misclassified as low-recovery, resulting in an extremely low error rate. In summary, the model not only fit the training and validation sets well, but also exhibited extremely robust performance on the test set.


```
Random Forest - Test Set Metrics:
Accuracy: 0.9941
Precision: 0.9887
Recall: 0.9950
F1 Score: 0.9918
AUC-ROC: 0.9998
Average Precision: 0.9996

Test Set Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5689
           1       0.99      0.99      0.99      3174

    accuracy                           0.99      8863
   macro avg       0.99      0.99      0.99      8863
weighted avg       0.99      0.99      0.99      8863
```





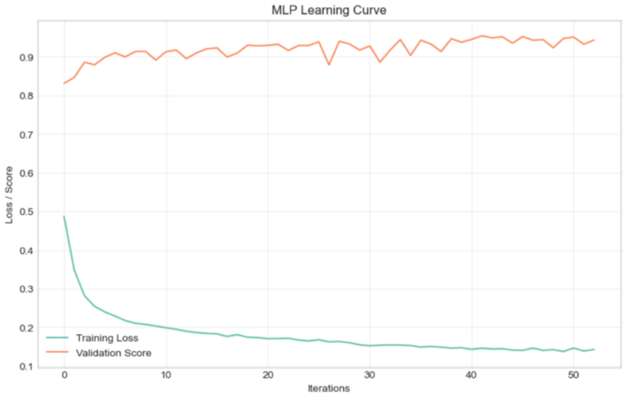
(3) Neural networks:

A multi-layer perceptron (MLPClassifier) neural network was used to classify and predict the COVID-19 recovery rate. The model architecture consists of three hidden layers with 64, 32, and 16 neurons, respectively. It uses the ReLU activation function and the Adam optimizer, and employs early stopping to prevent overfitting.

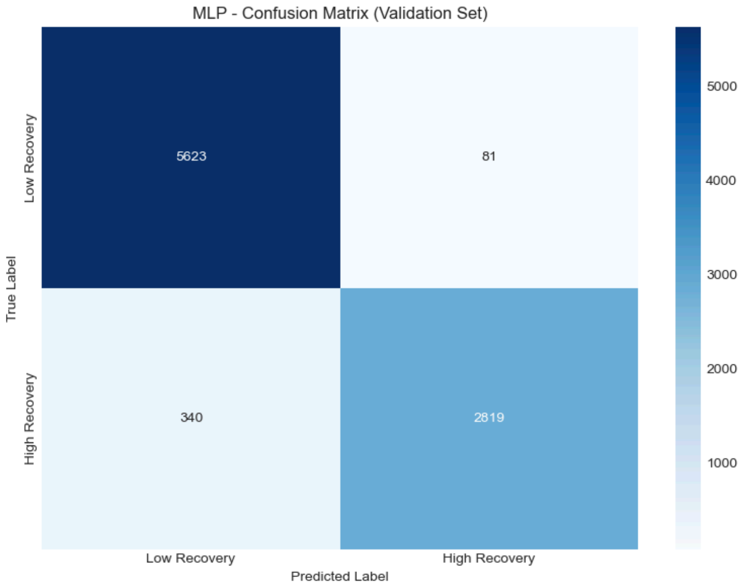
The learning curve shows that the model converged rapidly within the first 10 epochs: training loss steadily decreased and stabilized, while the validation score consistently remained above 0.9, indicating strong learning capability and training stability.

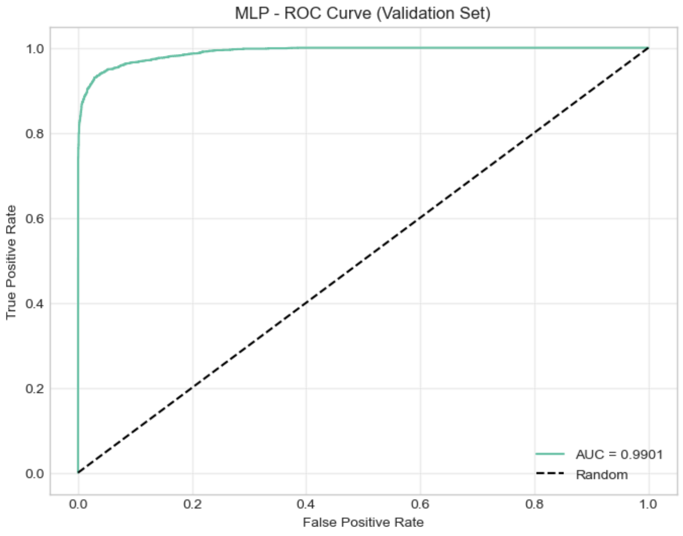
Evaluation on the validation set demonstrated that the MLP model achieved solid overall performance. However, the confusion matrix revealed a relatively high number of false negatives (340) for high recovery rate samples, which led to a lower recall in that category. In contrast, the classification performance for low recovery rate samples was significantly better.

```
Validation score did not improve more than tol=0.000100 for 10 consecutive epochs. Stopping.
```

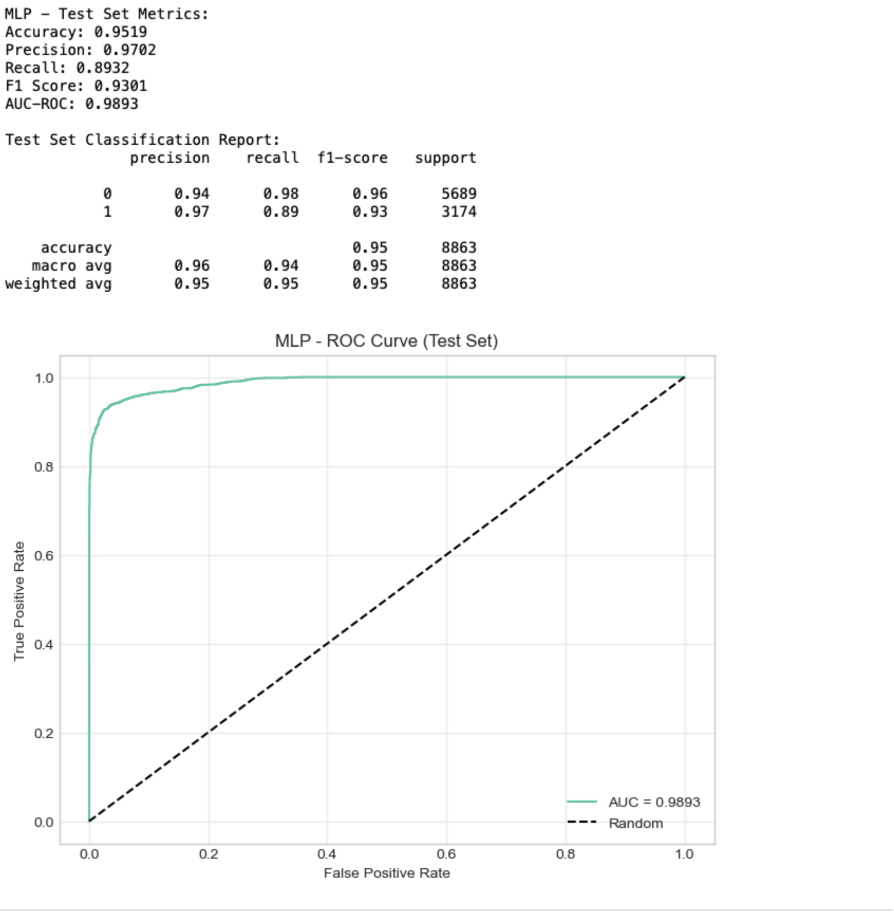


```
Neural Network (MLPClassifier) - Validation Set Metrics:
Accuracy: 0.9525
Precision: 0.9721
Recall: 0.8924
F1 Score: 0.9305
AUC-ROC: 0.9901

Validation Set Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      5704
           1       0.97      0.89      0.93      3159

    accuracy                           0.95      8863
   macro avg       0.96      0.94      0.95      8863
weighted avg       0.95      0.95      0.95      8863
```





On the test set, the MLP neural network model also demonstrated a relatively robust prediction ability. The overall accuracy was 95.19%, the precision reached 97.02%, the recall rate was 89.32%, the F1 score was 93.01%, and the AUC-ROC was 0.9893, which shows that the model has a strong recognition ability for the "high recovery rate" category. Overall, the MLP model maintained a high prediction performance during the test phase, especially showing strong fitting and generalization capabilities in multi-feature complex data.



```
MLP - Test Set Metrics:
Accuracy: 0.9519
Precision: 0.9702
Recall: 0.8932
F1 Score: 0.9301
AUC-ROC: 0.9893

Test Set Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      5689
           1       0.97      0.89      0.93      3174

    accuracy                           0.95      8863
   macro avg       0.96      0.94      0.95      8863
weighted avg       0.95      0.95      0.95      8863
```

● Machine Learning Summary:

Logistic regression performed poorly in predicting COVID-19 recovery rates due to its inherent assumption of a linear relationship between features and the target variable. This assumption limits its ability to capture the complex, nonlinear patterns often present in real-world epidemic data.

Random Forest was selected because several input features—such as Confirmed_full, Recovered_full, and Deaths_full—are either mathematically derived from or strongly correlated with the target label. Random Forest models excel at capturing decision boundaries and feature interactions, making them well-suited for detecting such deterministic relationships. However, this also raises the risk of data leakage, as features closely related to the label may artificially inflate the model's predictive performance.

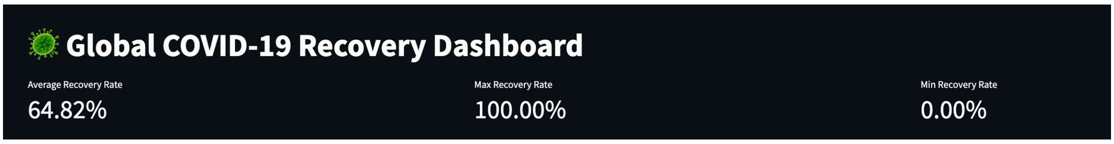
The neural network model (MLPClassifier) also achieved high accuracy, largely due to its ability to approximate complex nonlinear relationships in the data. When trained on well-preprocessed data that incorporates both temporal and country-level characteristics, the multi-layer architecture enables the model to effectively learn nuanced patterns associated with recovery trends.

● Potential biases or limitations in the model:

One concern with current models is bias caused by imbalanced data or uneven geographic distribution. If certain countries dominate the dataset or have more complete records, the model may not generalize well to underrepresented regions. Additionally, factors such as healthcare infrastructure, reporting standards, and government intervention, while critical, are not included in the current feature set, limiting the model's applicability in the real world.

● Streamlit app:

(1) KPI Summary Card
The average, maximum, and minimum COVID-19 recovery rates were calculated based on the most recent available data for each country. These metrics provide a comprehensive overview of the global recovery landscape and serve as key indicators in evaluating the effectiveness of pandemic response efforts across different regions.



🦠 Global COVID-19 Recovery Dashboard

Average Recovery Rate
64.82%

Max Recovery Rate
100.00%

Min Recovery Rate
0.00%

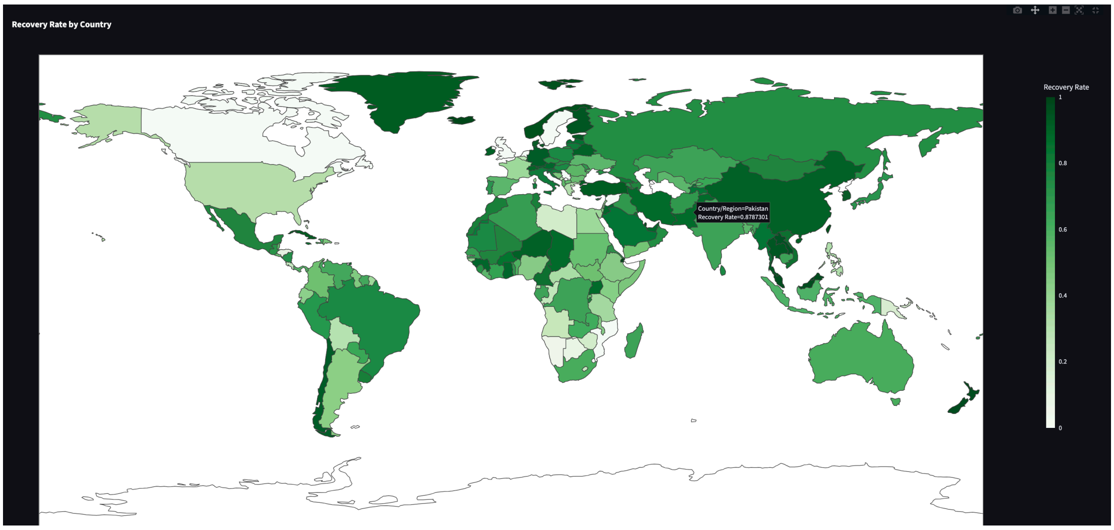
(2) Global Cure Rate Map

This map illustrates the number of COVID-19 tests conducted per million people, providing insight into global testing coverage.

The color depth represents testing intensity: darker blue indicates a higher number of tests (with values approaching 1 million tests per million people), while lighter blue or white signifies lower testing rates or missing data. The color bar legend on the right ranges from 0 to 1M, corresponding to the number of tests per million people.

The map includes an interactive feature: when the user hovers over a country, its name and testing details are displayed. For example:
Country/Region: Russia   Number of tests per million: 266,524.0

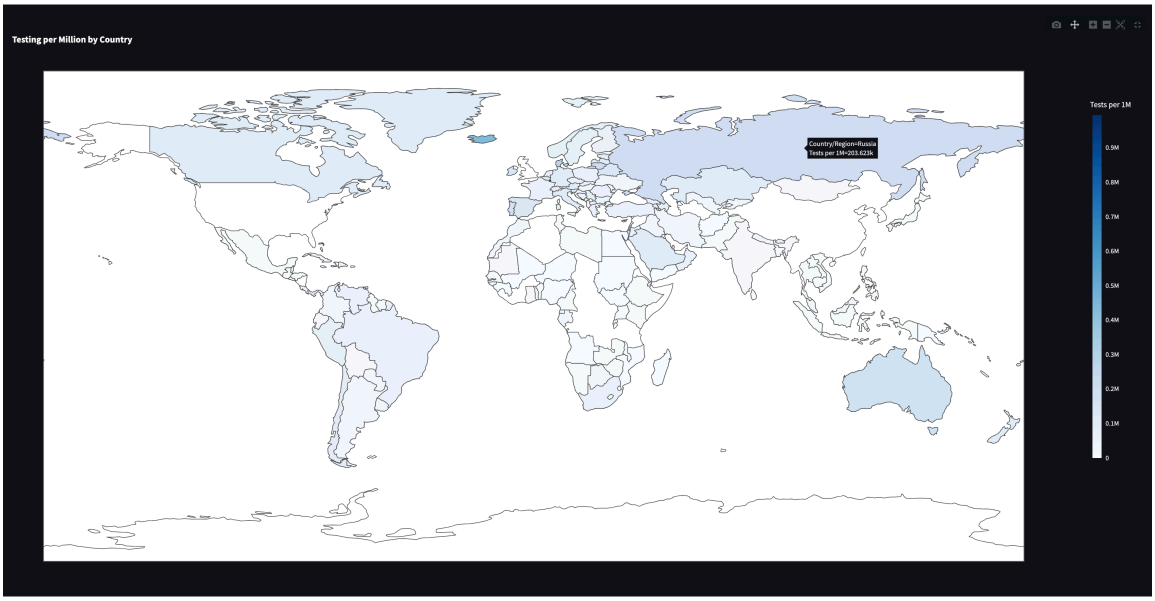
This visualization helps users quickly assess disparities in testing coverage across countries and regions.



(3) Testing Density Map
This map displays the number of COVID-19 tests conducted per million people, serving as an indicator of global testing coverage.

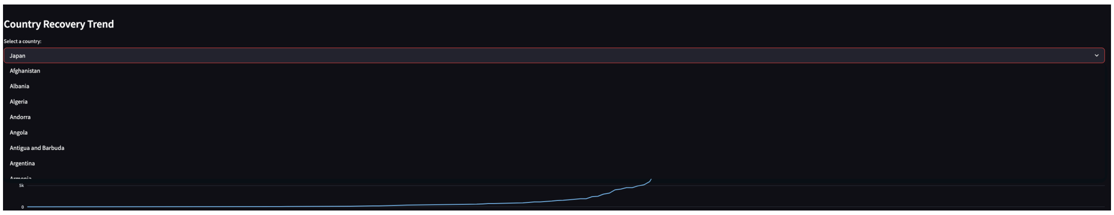
The color intensity reflects the testing density: darker blue indicates a higher number of tests (with values approaching 1 million tests per million people), while lighter blue or white represents lower testing rates or missing data. The colorbar legend on the right ranges from 0 to 1M, representing the number of tests per million population.

The map includes an interactive feature: when the user hovers over a country, its name and corresponding testing information are displayed. For example:
Country/Region: Russia   Number of tests per million: 266,524.0
This visualization enables a quick and intuitive comparison of testing coverage across countries, helping to identify regions with robust or insufficient testing practices.



(4) Country-level Cure Trends

Based on the model processing I did (we have seen that random forests work well before, so we used random forests) and the data after data cleaning and feature engineering, users can select a country/region and then use a slider to select the cure rate forecast for the next 0 to 30 days.



Country Recovery Trend

Select a country: Japan

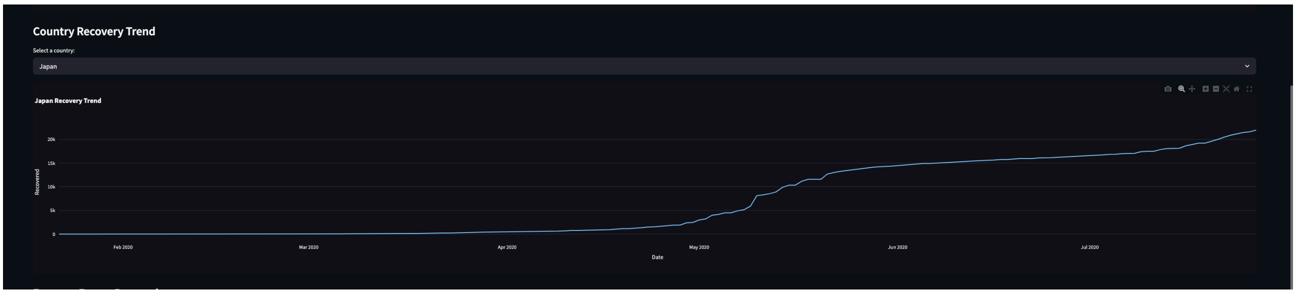
Taking Japan as an example, the line chart shows the cumulative number of people who have recovered in Japan since the outbreak in 2020. The timeline runs from January 2020 to August 2020, and the figure shows three characteristic stages:

Japan Recovery Trend (upper part)

Initial stability (January-March): The number of people who have recovered has increased slowly.

A sharp increase in April-May: It may be related to the first wave of control response to the epidemic, and the number of people who have recovered has risen rapidly.

After June, it has stabilized: The epidemic has eased and the curve has slowed down.



Country Recovery Trend

Select a country: Japan

Forecast Future Recoveries (lower part)

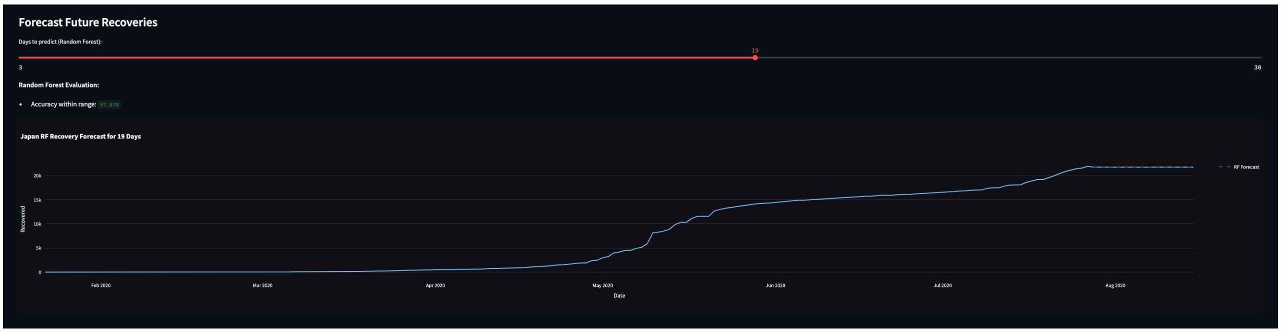
The user selects the number of forecast days through the slider, which is 19 days here.

The Random Forest Regressor model is used to predict the number of people who will recover in the future.

The following information is displayed:

Interval accuracy: 97.87%, indicating that the forecast effect is highly consistent with historical data.

The forecast curve is naturally connected with the real data, and the forecast value is displayed as a dotted line (RF Forecast), which is very clear in visualization. It is obvious that the model predicts that the number of recovered people will maintain a relatively stable upward trend in the next 19 days.



Forecast Future Recoveries

Days to predict (Random Forest): 19

Random Forest Evaluation:

- Accuracy within range: 97.87%

● Key Findings

Through this prediction dashboard, we found that the tree model can well learn the relationship between current features (such as date index) and the number of recovered people.

When the trend change is stable and controllable, and the trend of the number of recovered people is relatively gentle, it is suitable to use ensemble learning models such as random forests for prediction.

In addition, I found that high testing density is associated with high recovery tracking rate. Countries with higher testing density tend to have more complete and timely recovery data, resulting in higher and more accurate recovery rates. This can also be reflected in our streamlit dashboard.

● Future work

Several improvements could be considered in the future:
Integrate additional features: Future iterations could integrate more refined features such as age distribution, hospitalization rates, ICU capacity, vaccination rates, and public policy interventions to further improve model performance.
Add uncertainty estimates: Integrating prediction intervals or confidence intervals in forecast visualizations can help decision makers better assess risk and plan under uncertainty.
Real-Time Deployment and Updates: Deploying dashboards to the cloud and connecting them to real-time APIs allows for continuous monitoring and forecasting as new data becomes available.

● Conclusion

This project conducted a comprehensive analysis of global COVID-19 recovery trends through systematic data cleaning, feature engineering, and the application of multiple machine learning models. Among the models evaluated, random forests and neural networks demonstrated significantly superior performance, exhibiting strong predictive accuracy and generalization ability.

An interactive Streamlit dashboard was developed to support real-time data exploration, recovery forecasting, and cross-country visual comparisons. This tool enables users to intuitively grasp recovery trends and assess the global pandemic situation with greater clarity.

Moreover, the project identified several key insights—most notably, the positive correlation between testing density and the quality of recovery tracking. These findings underscore the importance of incorporating a broader set of epidemiological and policy-related features in future iterations to enhance model robustness and real-world applicability.