Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241126

Random Forest model-based risk prediction of COVID-19 regional infection

Yang Li

Faculty of Science, The University of Sydney, Camperdown NSW 2006, Australia

yali0621@uni.sydney.edu.au

Abstract. The current prevalence of the COVID-19 pandemic worldwide has posed numerous challenges and questions. To assist governments, medical institutions, and the public in making informed decisions and minimize the risk of further spread of COVID-19, this paper employs the Random Forest model to predict the infection risk within certain regions. The dataset utilized underwent data cleaning and feature engineering, allowing predictions to be made using publicly accessible data such as local basic climate conditions. After conducting performance comparisons with other common machine learning models, including Linear Regression and Decision Tree Regressor, it was found that the Random Forest Regressor model exhibited superior performance across all evaluation metrics, with all error values below 0.05. Notably, the MAE for the Random Forest model was only 0.001089. This strongly suggests that the Random Forest model outperforms the other models used in this task.

Keywords: COVID-19, Machine Learning, Random Forest, Coronavirus, Data Analytics.

1. Introduction

As COVID-19 spreads globally, predicting and analyzing regional infection risks has become an urgent issue. The dynamic nature of disease transmission poses challenges to traditional forecasting methods, necessitating the use of advanced machine learning techniques to enhance prediction accuracy and timeliness.

Previous research primarily focused on individual variations. For instance, studies by Mohammad Pourhomayoun and colleagues leveraged multiple machine learning models to address predictions related to patients' mortality risks [1-2]. Najada Firza and her team employed the Random Forest model to develop a predictive model based on lifestyle risks and health factors to estimate the severity of COVID-19 infections. Similarly, Lijing Jia and associates created high-performance interpretable machine learning models for predicting the risk of worsening conditions in COVID-19 patients, using extensive clinical data [3].

While most of these studies have made significant progress in in-depth research at the individual level, they mainly concentrate on studying the conditions of already infected individuals. There remains a relative scarcity of studies addressing group infection risk predictions. This paper introduces the application of the Random Forest model to forecast COVID-19 infection risks. It further explores how various features, such as population density, climatic conditions, and socio-economic indicators, can be utilized to predict the risk of COVID-19 infections.

The aim of this study is not merely to predict the infection risks of COVID-19. It also seeks to identify key factors influencing these risks, providing data support for public health decision-makers. This would assist them in making more informed decisions during epidemic prevention and control.

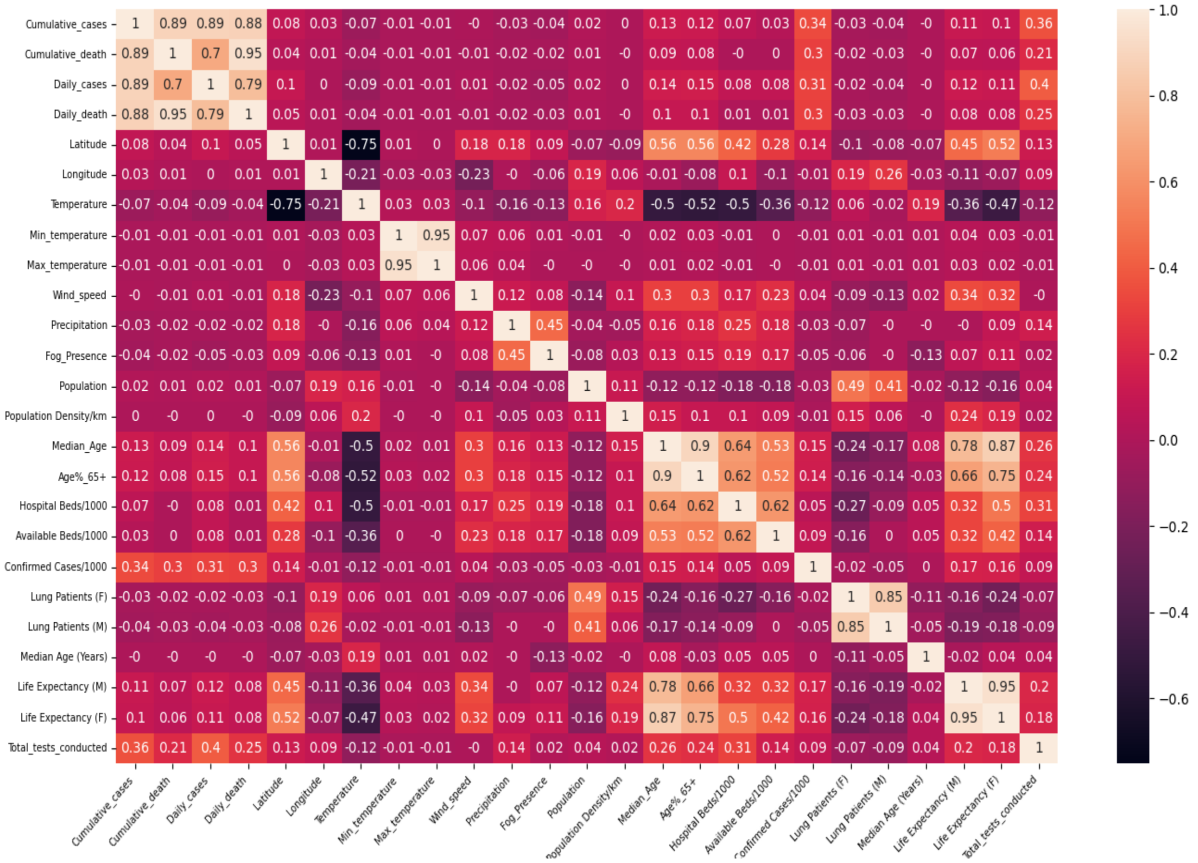
2. Prediction method based on random forest model

In this chapter, the focus shifts to the primary methodology of the study - a prediction approach hinged on the Random Forest model. The predictive capability and robustness of the Random Forest model, especially in scenarios grappling with multifaceted datasets like the one in question, render it an ideal candidate for our examination. This chapter unfolds in a structured manner, commencing with a description of the dataset and the inherent challenges associated with its use. This is followed by an in-depth account of the data cleaning processes undertaken, which is crucial to ensure the reliability and accuracy of any subsequent analyses.

Thereafter, we delve into the specifics of the Random Forest model, elucidating its mechanics, importance in the realm of machine learning, and its unique capacity to address the study's aims. We discuss its methodological underpinnings, its ability to cater to the complexities of predicting COVID-19 risk factors, and its intrinsic robustness against potential outliers and uncertainties often encountered in public health data. By the chapter's conclusion, readers will possess a clear understanding of why the Random Forest model was employed, how it operates, and its potential advantages in forecasting COVID-19 infection risks.

2.1. Dataset

The dataset used in this study was sourced from Kaggle, specifically the "CoVCSD - Covid-19 Countries Statistical Dataset." [4]. The dataset encompasses 6,486 observations and 29 features. The heat-map of this dataset is presented in Figure 1.



	Cumulative_cases	Cumulative_death	Daily_cases	Daily_death	Latitude	Longitude	Temperature	Min_temperature	Max_temperature	Wind_speed	Precipitation	Fog_Presence	Population	Population Density/km	Median_Age	Age%_65+	Hospital Beds/1000	Available Beds/1000	Confirmed Cases/1000	Lung Patients (F)	Lung Patients (M)	Median Age (Years)	Life Expectancy (M)	Life Expectancy (F)	Total_tests_conducted
Cumulative_cases	1	0.89	0.89	0.88	0.08	0.03	-0.07	-0.01	-0.01	-0	-0.03	-0.04	0.02	0	0.13	0.12	0.07	0.03	0.34	-0.03	-0.04	-0	0.11	0.1	0.36
Cumulative_death	0.89	1	0.7	0.95	0.04	0.01	-0.04	-0.01	-0.01	-0.01	-0.02	-0.02	0.01	-0	0.09	0.08	-0	0	0.3	-0.02	-0.03	-0	0.07	0.06	0.21
Daily_cases	0.89	0.7	1	0.79	0.1	0	-0.09	-0.01	-0.01	0.01	-0.02	-0.05	0.02	0	0.14	0.15	0.08	0.08	0.31	-0.02	-0.04	-0	0.12	0.11	0.4
Daily_death	0.88	0.95	0.79	1	0.05	0.01	-0.04	-0.01	-0.01	-0.01	-0.02	-0.03	0.01	-0	0.1	0.1	0.01	0.01	0.3	-0.03	-0.03	-0	0.08	0.08	0.25
Latitude	0.08	0.04	0.1	0.05	1	0.01	-0.75	0.01	0	0.18	0.18	0.09	-0.07	-0.09	0.56	0.56	0.42	0.28	0.14	-0.1	-0.08	-0.07	0.45	0.52	0.13
Longitude	0.03	0.01	0	0.01	0.01	1	-0.21	-0.03	-0.03	-0.23	-0	-0.06	0.19	0.06	-0.01	-0.08	0.1	-0.1	-0.01	0.19	0.26	-0.03	-0.11	-0.07	0.09
Temperature	-0.07	-0.04	-0.09	-0.04	-0.75	-0.21	1	0.03	0.03	-0.1	-0.16	-0.13	0.16	0.2	-0.5	-0.52	-0.5	-0.36	-0.12	0.06	-0.02	0.19	-0.36	-0.47	-0.12
Min_temperature	-0.01	-0.01	-0.01	-0.01	0.01	-0.03	0.03	1	0.95	0.07	0.06	0.01	-0.01	-0	0.02	0.03	-0.01	0	-0.01	0.01	-0.01	0.01	0.04	0.03	-0.01
Max_temperature	-0.01	-0.01	-0.01	-0.01	0	-0.03	0.03	0.95	1	0.06	0.04	-0	-0	-0	0.01	0.02	-0.01	-0	-0.01	0.01	-0.01	0.01	0.03	0.02	-0.01
Wind_speed	-0	-0.01	0.01	-0.01	0.18	-0.23	-0.1	0.07	0.06	1	0.12	0.08	-0.14	0.1	0.3	0.3	0.17	0.23	0.04	-0.09	-0.13	0.02	0.34	0.32	-0
Precipitation	-0.03	-0.02	-0.02	-0.02	0.18	-0	-0.16	0.06	0.04	0.12	1	0.45	-0.04	-0.05	0.16	0.18	0.25	0.18	-0.03	-0.07	-0	-0	-0	0.09	0.14
Fog_Presence	-0.04	-0.02	-0.05	-0.03	0.09	-0.06	-0.13	0.01	-0	0.08	0.45	1	-0.08	0.03	0.13	0.15	0.19	0.17	-0.05	-0.06	-0	-0.13	0.07	0.11	0.02
Population	0.02	0.01	0.02	0.01	-0.07	0.19	0.16	-0.01	-0	-0.14	-0.04	-0.08	1	0.11	-0.12	-0.12	-0.18	-0.18	-0.03	0.49	0.41	-0.02	-0.12	-0.16	0.04
Population Density/km	0	-0	0	-0	-0.09	0.06	0.2	-0	-0	0.1	-0.05	0.03	0.11	1	0.15	0.1	0.1	0.09	-0.01	0.15	0.06	-0	0.24	0.19	0.02
Median_Age	0.13	0.09	0.14	0.1	0.56	-0.01	-0.5	0.02	0.01	0.3	0.16	0.13	-0.12	0.15	1	0.9	0.64	0.53	0.15	-0.24	-0.17	0.08	0.78	0.87	0.26
Age%_65+	0.12	0.08	0.15	0.1	0.56	-0.08	-0.52	0.03	0.02	0.3	0.18	0.15	-0.12	0.1	0.9	1	0.62	0.52	0.14	-0.16	-0.14	-0.03	0.66	0.75	0.24
Hospital Beds/1000	0.07	-0	0.08	0.01	0.42	0.1	-0.5	-0.01	-0.01	0.17	0.25	0.19	-0.18	0.1	0.64	0.62	1	0.62	0.05	-0.27	-0.09	0.05	0.32	0.5	0.31
Available Beds/1000	0.03	0	0.08	0.01	0.28	-0.1	-0.36	0	-0	0.23	0.18	0.17	-0.18	0.09	0.53	0.52	0.62	1	0.09	-0.16	0	0.05	0.32	0.42	0.14
Confirmed Cases/1000	0.34	0.3	0.31	0.3	0.14	-0.01	-0.12	-0.01	-0.01	0.04	-0.03	-0.05	-0.03	-0.01	0.15	0.14	0.05	0.09	1	-0.02	-0.05	0	0.17	0.16	0.09
Lung Patients (F)	-0.03	-0.02	-0.02	-0.03	-0.1	0.19	0.06	0.01	0.01	-0.09	-0.07	-0.06	0.49	0.15	-0.24	-0.16	-0.27	-0.16	-0.02	1	0.85	-0.11	-0.16	-0.24	-0.07
Lung Patients (M)	-0.04	-0.03	-0.04	-0.03	-0.08	0.26	-0.02	-0.01	-0.01	-0.13	-0	-0	0.41	0.06	-0.17	-0.14	-0.09	0	-0.05	0.85	1	-0.05	-0.19	-0.18	-0.09
Median Age (Years)	-0	-0	-0	-0	-0.07	-0.03	0.19	0.01	0.01	0.02	-0	-0.13	-0.02	-0	0.08	-0.03	0.05	0.05	0	-0.11	-0.05	1	-0.02	0.04	0.04
Life Expectancy (M)	0.11	0.07	0.12	0.08	0.45	-0.11	-0.36	0.04	0.03	0.34	-0	0.07	-0.12	0.24	0.78	0.66	0.32	0.32	0.17	-0.16	-0.19	-0.02	1	0.95	0.2
Life Expectancy (F)	0.1	0.06	0.11	0.08	0.52	-0.07	-0.47	0.03	0.02	0.32	0.09	0.11	-0.16	0.19	0.87	0.75	0.5	0.42	0.16	-0.24	-0.18	0.04	0.95	1	0.18
Total_tests_conducted	0.36	0.21	0.4	0.25	0.13	0.09	-0.12	-0.01	-0.01	-0	0.14	0.02	0.04	0.02	0.26	0.24	0.31	0.14	0.09	-0.07	-0.09	0.04	0.2	0.18	1

Figure 1. Heat-map of the dataset.

This dataset is organized by countries/regions. The test regions are primarily located between 0-65 degrees north latitude, 50-100 degrees west longitude, and 0-100 degrees east longitude. The temperature range lies between 0-30 degrees Celsius, and the wind speed varies from 0-20m/s. Approximately 32.6% of the individuals in the dataset are aged above 65. Histograms of some of the features can be observed in figure 2.

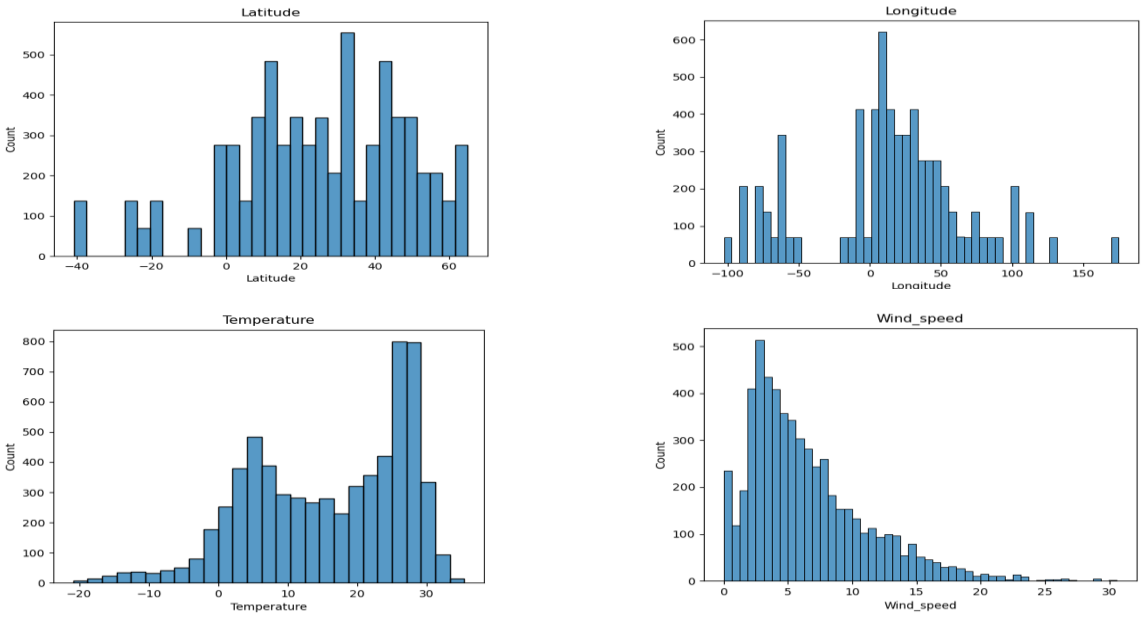


Figure 2. Histograms of selected features.

From the above, it’s evident that the dataset offers a rich sample variety with minimal age bias. However, one notable limitation is a significant bias in the regional sampling. The features contained in the dataset are presented as Table 1.

Table 1. All features of the dataset.

Serial No.	Features	Data types
1	Country	Character
2	Cumulative_cases	Discrete
3	Cumulative_death	Discrete
4	Daily_cases	Discrete
5	Daily_death	Discrete
6	Latitude	Discrete
7	Longitude	Discrete
8	Temperature	Discrete
9	Min_temperature	Discrete
10	Max_temperature	Discrete
11	Wind_speed	Discrete
12	Precipitation	Binary
13	Fog_Presence	Binary
14	Population	Discrete
15	Population Density/km	Discrete
16	Median_Age	Discrete

Table 1. (continued).

17	Sex Ratio	Discrete
18	Age%_65+	Discrete
19	Hospital Beds/1000	Discrete
20	Available Beds/1000	Discrete
21	Confirmed Cases/1000	Discrete
22	Lung Patients (F)	Discrete
23	Lung Patients (M)	Discrete
24	Life Expectancy (M)	Discrete
25	Life Expectancy (F)	Discrete
26	Total_tests_conducted	Discrete
27	Out_Travels (mill.)	Discrete
28	In_travels(mill.)	Discrete
29	Domestic_Travels (mill.)	Discrete

2.2. Data Cleaning

Given the significant data scarcity, with over 90% of missing values, the columns ‘Sex Ratio’, ‘Out_Travels (mill.)’, ‘In_travels(mill.)’, and ‘Domestic_Travels (mill.)’ were initially removed. Building upon this, 197 duplicate rows were further eliminated. To mitigate the risk of overfitting, certain extreme data points were discarded. Missing values were then imputed using the mean. Ultimately, 30% of the dataset was reserved for testing, while the remaining 70% was utilized for training.

2.3. Random Forest Model

The exploration of various machine learning algorithms led this study to give particular attention to the Random Forest, an ensemble learning method based on decision trees. The selection of this model was informed by its exemplary performance when dealing with similar problems. The ensuing sections delve deeper into its application and principles in the context of this research.

At its core, the working mechanism of the Random Forest involves constructing multiple decision trees. Training for each tree is based on randomly drawn data samples and feature subsets, as depicted in Figure 3.

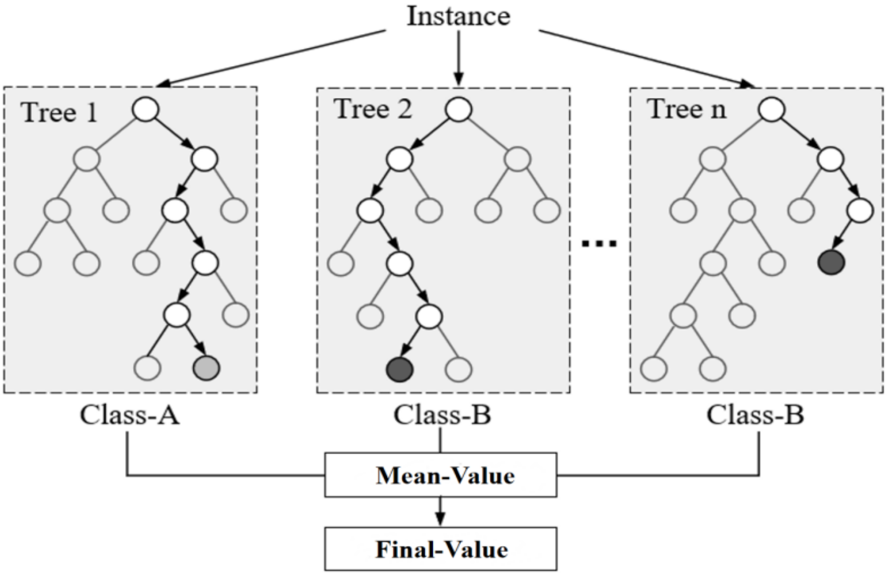


Figure 3. Random Forest Regressor Model.

In the context of this study, this signifies that different trees might concentrate on varying features, be it population density, climatic conditions, or any other factor relevant to the transmission of COVID-19.

The uniqueness of each tree ensures a diverse model, a feature of paramount importance considering the dataset with its 6486 instances and 29 features. This diversity not only enhances the model's generalization capability but also substantially mitigates the risk of over-fitting.

During prediction, each tree provides its evaluation independently, with the Random Forest consolidating these by averaging results across all trees:

$$\bar{y} = \frac{1}{N_{trees}} \sum_{K=1}^{N_{trees}} \bar{y}_k(x) \tag{1}$$

Where $\bar{y}_k$ is the prediction of the k_{th} tree and K represents the total number of trees in the forest.

In the realm of COVID-19 risk prediction, this methodology allows for a comprehensive risk estimation by taking into account a variety of influencing factors and their interactions.

Yet, the capabilities of the Random Forest don't just end here. It also furnishes an assessment of feature importance, granting insights into which factors wield the most influence over the transmission of COVID-19.

Lastly, it was observed that the Random Forest exhibits high robustness against outliers and uncertainties in the data. Given the volatile and intricate nature of public health data, this aspect proves particularly valuable.

3. Experimental Results and Analysis

In this study, the problem addressed pertains to regression, specifically, the risk of COVID-19 infection (number of infections per thousand). Accordingly, three metrics, namely MAE, MSE, and RMSE, were chosen for evaluation [5]:

MSE (Mean Squared Error):

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y}_{tp})^2 \tag{2}$$

Represents the expected squared difference between the actual and predicted values.

RMSE (Root Mean Squared Error):

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y}_{tp})^2} \tag{3}$$

The square root of the mean squared error, representing the deviation between predictions and actual observations.

MAE (Mean Absolute Error):

$$\frac{1}{n}\sum_{t=1}^{n}|y_t - \bar{y}_{tp}| \tag{4}$$

The average absolute error between the predicted and actual values. Where y_t is the actual value and $\bar{y}_{tp}$ is the predicted value.

MSE, a commonly used regression metric, gives weight to each deviation between prediction and reality, severely penalizing larger errors, making it very sensitive to outliers. RMSE, being the square root of MSE, is interpretable in the same units as the target variable and is also sensitive to large errors, measuring the volatility of predictions. MAE, on the other hand, provides an absolute measure of fit, offering an intuitive feel of the prediction error's true magnitude.

Taken together, these metrics offer a comprehensive, nuanced, and visual evaluation of model performance. While MAE gives a direct sense of the prediction error, MSE and RMSE highlight the model's sensitivity to larger prediction errors. Used in conjunction, these metrics help us understand model performance under various scenarios. Comparative performance of the Random Forest model against other models is presented below:

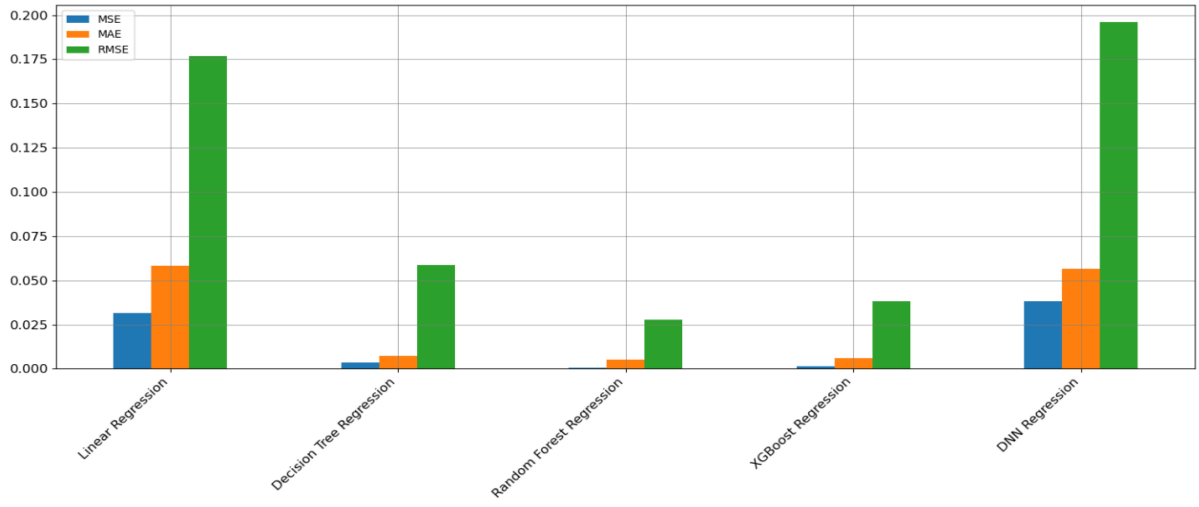


Figure 4. Model Performance Comparison.

Table 2. Model Error Data (rounded to three decimal places).

Method	MAE	MSE	RMSE
Linear Regression [6]	0.031	0.058	0.177
Decision Tree [7]	0.002	0.007	0.06
Deep Neural Networks [8]	0.038	0.054	0.196
XGBoost [9]	0.001	0.006	0.038
Random Forest [10]	0.001	0.005	0.032

Drawing from the data in Figure 4 and Table 2, and by observing the three key metrics – MSE, MAE, and RMSE – we derive the following observations and conclusions:

For MAE, the Random Forest regression recorded the lowest at 0.001089, outperforming the second-best, XGBoost regression, by approximately 26%. A low MAE score for the Random Forest indicates a high accuracy with minimal error, attributable to its ensemble method averaging out predictions across multiple trees, thus mitigating anomalies or extreme predictions. In contrast, while XGBoost is also an ensemble technique, it focuses on sequentially rectifying errors of prior trees, which may lead to slightly increased errors in certain scenarios.

In terms of MSE, the Random Forest model exhibited the best performance with 0.00532, outperforming the Decision Tree regression’s 0.006845 by about 22%. This superior performance indicates that not only does the Random Forest model have a smaller variance, but it also has a minor bias, making it adept at delivering stable predictions across varied data distributions.

For RMSE, the Random Forest model’s score of 0.032997 was again the lowest, outpacing the XGBoost’s 0.038346 by roughly 14%. This superiority suggests that the Random Forest not only offers predictions closer to the real values but also has fewer instances of significant errors.

Interestingly, Linear Regression outperformed the DNN regression across all three metrics. For instance, in terms of MSE, Linear Regression’s 0.031354 was approximately 18% lower than the DNN’s 0.038398. This could be attributed to the linear nature of the data, making linear models excel for this specific task. In contrast, the DNN, due to its depth and complexity, might overfit without sufficient data or proper tuning, resulting in greater errors on the test data.

Evidently, the Random Forest’s performance was superior amongst all models. This reaffirms its potency as an ensemble technique, excelling in numerous tasks by constructing multiple decision trees and amalgamating their predictions to reduce overfitting, enhance stability, and improve generalization.

In conclusion, the Random Forest regression showcased superior performance in this experiment, not only securing the best scores across MSE, MAE, and RMSE but also significantly outpacing other models. This demonstrates that for this specific dataset and task, the Random Forest is an exceptionally apt and potent choice. The models of Decision Tree, Random Forest, and XGBoost significantly outperformed both Linear Regression and Deep Neural Networks. The stellar performance of the Random Forest model in predicting regional risks of COVID-19 can be attributed to:

- Diversity and Robustness: Random Forest, by building multiple decision trees and averaging or voting on their results, effectively reduces biases or noise that might arise in individual models.
- Handling High Dimensionality: The post-processed dataset has 25 features. Random Forest efficiently manages high-dimensional data and inherently performs feature selection.
- Preventing Overfitting: Randomly selecting feature subsets for splits and employing bootstrapping add an extra layer of randomness, preventing overfitting and improving generalization.

4. Conclusion

The study utilized the Kaggle COVID-19 statistical dataset, which encompasses multiple features, facilitating a comprehensive analysis. Rigorous preprocessing was carried out on the dataset, ensuring the accuracy and robustness of the models. A regional infection risk prediction model based on Random Forest was adopted and juxtaposed with other machine learning models. It was revealed that the Random Forest model significantly outperformed other models in predicting the regional risks of COVID-19 infection.

While the Random Forest model demonstrated commendable performance in this study, there remains room for improvement, particularly in hyperparameter tuning and feature selection. Moreover, although the dataset is relatively extensive, it displays significant regional sampling bias. Future studies might consider utilizing a more exhaustive and representative dataset. Additionally, in real-world applications, the model could need to account for various pragmatic factors, such as policy changes, individual behaviors, and habits, all of which could considerably influence the spread of COVID-19.

References

[1] Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health.* 2021, 100178.

[2] Firza N, Monaco A. Forecasting Model Based on Lifestyle Risk and Health Factors to Predict COVID-19 Severity. *International Journal of Environmental Research and Public Health.* 2022 Oct 1;**19(19)**: 12538.

[3] Jia L, Wei Z, Zhang H, Wang J, Jia R, Zhou M, et al. An interpretable machine learning model based on a quick pre-screening system enables accurate deterioration risk prediction for COVID-19. *Scientific Reports.* 2021, **11(1)**.

[4] CoVCSD - Covid-19 Countries Statistical Dataset [Internet]. www.kaggle.com. https://www.kaggle.com/datasets/aestheteaman01/covcsd-covid19-countries-statistical-dataset

[5] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peer J Computer Science.* 2021 5;**7(5)**:e623.

[6] Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. *Google Books. John Wiley & Sons*; 2021,1-11.

[7] Song YY, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry* 2015, **27(2)**:130–5.

[8] Sze V, Chen YH, Yang TJ, Emer JS. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*. 2017 **105(12)**:2295–329.

[9] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *International Conference on Knowledge Discovery and Data Mining*. 2016, 785–94.

[10] Shafi A. Sklearn Random Forest Classifiers in Python Tutorial [Internet]. www.datacamp.com. 2023. Available from: https://www.datacamp.com/tutorial/random-forests-classifier-python