Part B –Supervised Learning

The goal of implementing the supervised learning algorithms (decision tree, gradient boosting, and random forest algorithms) was to predict the rate of change of covid cases with respect to the temperature, the date and the proportion of fully vaccinated people for a given location.

To test the performance of the algorithms, I used the test and train method, i.e., I trained the models on 90 % of the data and tested the model on the remaining 10 %. During each iteration of the training process for each model, the maximum depth of the tree is changed (10, 20, 30 or 40). This is done to detect overfitting. I also initialized the random state of the algorithms to 44 for reproducibility.

The tables below summarizes the results of the 3 algorithms.

Decision Tree Regressor:

Depth	MSE	RMSE	MAE	R2
10	1236267	1112	649	0.9699
20	829088	911	382	0.9798
30	527415	726	260	0.9872
40	464821	681	212	0.9887

Gradient Boosting Regressor:

Depth	MSE	RMSE	MAE	R2
10	267365	517	252	0.9935
20	411625	642	219	0.9900
30	571014	755	239	0.9861
40	451690	672	218	0.9890

Random Forest Regressor:

Depth	MSE	RMSE	MAE	R2
10	1243	1544703	252	0.9624
20	750	562044	219	0.9863
30	699	489146	239	0.9881
40	690	476320	218	0.9884

The table below shows the time taken to construct the models.

Algorithm	Time (in mins)
Decision Tree Regressor	30

Gradient Boosting Regressor	20
Random Forest Regressor	20

GitHub repository link – https://github.com/bsewp045/CSI4142-Phase-4

Actionable insights:

For all the 3 algorithms and for all the depths, the value of the R2 score is very high (above 0.95). This suggests that the temperature, the date, and the proportion of fully vaccinated people are good predictors for predicting the rate of change of covid cases for a given location.

The RMSE values allows us to detect some level of overfitting. For instance, for the gradient boosting regressor algorithm, the lowest RMSE was obtained when the maximum depth of the tree was set to 10. This suggests that some overfitting might be happening when the depth of the tree is increased beyond 10, which causes the complexity of the model to increase.