REGRESSION ALGORITHM

Problem Statement : Predicting insurance charges for a client from a dataset.

Dataset Information : The Dataset contains calculated insurance charges based on Age, Sex, BMI, Children, Smoker.

Pre-Processing Method : There are two columns with a string datatype which is a nominal Data, but in-order for the algorithm to read the datatype, it needs to be converted to a number type for creating a good prediction model

ALGORITHM MODELS:

1. MULTIPLE LINEAR REGRESSION

The “Simple Linear Regression” model provides a predicted r2_score value as 0.7894790349867009.

Hence this is not a good model, as this is not close to 1.

2.SIMPLE VECTOR MACHINE

KERNEL/Model	r2_Score
Linear	-0.111661287196084
Poly	-0.0642925840210553
Rbf	-0.0884273277691388
Sigmoid	-0.0899412170256757
Precomputed	**ValueError:** Precomputed matrix must be a square matrix. Input is a 936x5 matrix.

None of the r2_score values from Simple Vector Machine is closer to 1.
Hence this is not a good model.

DECISION TREE REGRESSION

CRITERION	SPLITTER	R2_SCORE
Squared_error	Best	0.706728238107016
Friedman_mse	Best	0.692330008177228
Absolute_error	Best	0.670882212019776
Poisson	Best	0.722964483205367
Squared error	Random	0.731122165680786
Friedman_mse	Random	0.749912355387741
Absolute_error	Random	0.70022203641009
Poisson	Random	0.636562860683202

None of the r2_score values from Decision Tree Regression was closer to 1. So this is not a good model.

RANDOM FOREST

n_estimator	Random state	R2_Score
50	0	0.849586047230992
80	0	0.853326157859657
100	0	0.853707449231218
200	0	0.85247678240466
500	0	0.853097852839674
50	50	0.854278839820169
80	50	0.856207643875604
100	50	0.857632867709948
200	50	0.856472132293473
500	50	0.856968223891954

n_estimator	Random State	max_features	r2_score
50	0	Sqrt	0.869919600469524
100	0	Sqrt	0.871288294739591
150	0	Sqrt	0.870481528059162
200	0	Sqrt	0.871413688772296
50	0	Log2	0.869919600469524
100	0	Log2	0.871288294739591
150	0	Log2	0.870481528059162
200	0	Log2	0.871413688772296

As we can see the r2_score value of the first four combinations of n_estimator and random state with max feature as “Sqrt” and next four combinations of n_estimator and random state with max feature as “log2” are identical.

As those two max_features doesn’t make any difference and we can either choose “sqrt” or “log2” as a parameter for criterion in Random forest model. The r2_score value closest is 0.871413688772296

CONCLUSION

Hence the closest r2_score value predicted was 0.8714136887722296, I choose the ”Random forest” model with the combination of parameters (n_estimator = 200 , random_state = 0 , max_features = sqrt / log2). Even though the predicted r2_score value is not closer to 1, but still its better than the r2_score values predicted by other models.