PRML BONUS PROJECT REPORT
FLIGHT PRICE PREDICTION

1) ABSTRACT

Flight prices are often difficult to predict and depend on a number of factors including length of journey, source, destination etc. In this project the aim is to predict the flight price depending on a number of factors as given in the dataset.

2) DATA PROCESSING AND CLEANING

Preprocessing is done. The dataset has 11 columns including the target column ‘Price’. On exploratory analysis, nan values are dealt with and several non-categorical and non-numerical data is located.

- ‘Date_of_Journey’ is converted to a proper datetime format.
- It is subsequently split to create 3 new columns: ‘Date_Month’, ‘Date_Day’, ‘Date_Year’.
- On further observation all values of ‘Date_Year’ are 2019 so it is dropped due to lack of correlation to final predicted price.
- ‘Source’, ‘Destination’, ’Route’,’Additional Info’ are label encoded.

- ‘Total_Stops’ which is given in string format is converted to int, ie from ‘no stops’ to 0 , ‘1 stop’ to 1 and so on.

- ‘Duration’ is given in an ‘x h y m’ format so it has been successfully converted into an int type column containing duration in minutes.

- ‘Arrival_Time’ and ‘Dep_Time’ are successfully converted to hour and minute format.

3) METHOD

With a fully categorical/numerical dataset, several regressors are now applied with hyperparameter tuning

1) Decision Tree Regressor

```
The best splitter is: random
The best creteria is: friedman_mse
The minimum sample split is: 5

r2 score: 0.8387710627133875
mse: 3394351.632046282
```

2) LightGBM Regressor

```
The best n_est is: 1000
The best n_leaves is: 12

r2 score: 0.9099759892304289
mse: 1895274.8372696654
```

3) XGBoost Regressor

```
The best verbosity is: 0
The best depth: 8

r2 score: 0.8906511973654991
mse: 2302119.538410828
```

4) Extra Trees Regressor

```
r2 score: 0.93916513924632
mse: 1810790.5964142554
```

5) Random Forest Regressor

```
r2 score: 0.9370125439805143
mse: 2201871.1185791176
```

4) MODEL EVALUATION

	Model	r2_score	mse
0	Decision Tree Regressor	0.853783	3.078307e+06
1	LGBM Regressor	0.909976	1.895275e+06
2	XGBoost Regressor	0.890651	2.302120e+06
3	Extra Trees Regressor	0.939165	1.810791e+06
4	Random Forest Regressor	0.937013	2.201871e+06

Thus ExtraTreesRegressor is the best model and Decision Tree Regressor is the worst.