• Abstract:-

-> Air travel has become one of the most popular means of transportation for people who have to travel from place to place within a short span of time. Clearly, this means that time is an important resource when it comes to flights and flight schedules. Various air travel terminologies like ‘Taxi in’ time, ‘Wheels off’ time, ‘Wheels in’ time, ‘Taxi out’ time, etc. denote various time schedules and durations with respect to a flight. Utilizing a dataset of these various flight-related time durations and schedules, spanning over five years, an attempt has been made to understand the correlation between various factors while result in a flight being delayed from its expected schedule. Using this historical data and various flight-related parameters/features, ‘Categorical Naïve Bayes classifier’ algorithm has been used to classify the ‘arrival delay’ (the delay which the flight faced in reaching its destination as opposed to its expected arrival time) into blocks of 45 minutes, which gives an accuracy of 90%.

• Brief methodology:-

-> In all, 13 features (out of all the features in the dataset) were chosen as being the most relevant in predicting the duration (in terms of chunks of time) of delay (or early arrival, in some cases) of a given flight. Hence, using this classifier, at the time the flight leaves from its source, its delay or early arrival at the destination can be predicted.

-> The relevance and importance of each of these features was properly weighed, and its possible contribution toward the delay of a flight was thought of.

-> 9 of these features were in textual/string format, and the remaining 4 in numerical format.

-> First, cancelled flights were filtered out.

-> To properly apply Categorical Naïve Bayes classifier (using sklearn library), all of these features had to be encoded into labels (this is known as ‘label encoding’). This was done using sklearn’s ‘Label Encoder’.

-> Then, the target feature ‘Arrival Delay Blocks’ was further discretized into 5 classes.

-> 80% of the dataset was used for training, and the remaining 20% for testing

-> Finally, these 13 label-encoded features were fit into the Categorical Naïve Bayes classifier model and it was tested, which resulted in a nice accuracy of 90%

-> Finally, evaluation metrics were applied and the results were visualized.