Project Design Phase-I
Proposed Solution Template

Date	19 September 2022
Team ID	PNT2022TMID36222
Project Name	Project - developing a flight delay prediction model using machine learning
Maximum Marks	2 Marks

Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	we focused the idea and research on LaGuardia International Airport. Compared with the data produced by all airports in USA, the data which we gathered was very limited, but it gave us a great direction on how weather plays a part in flight delays.
2.	Idea / Solution description	These delays not only cause inconveniences to the airlines but also to the passengers. The result is an increase in travel time which increases the expenses associated with food and lodging and ultimately causes stress among passengers. The repositioning, fuel consumption while trying to reduce elapse times, and many
3.	Novelty / Uniqueness	The data that I used comes from Kaggle and it consists of a multi-year dataset ranging from 2009 to 2018 separated by year, so one file per year. Each one of these files contains an average of 28 categories with a few million rows. Because of the size of each file I chose to work only with of over 7.2 million rows.
4.	Social Impact / Customer Satisfaction	At this point I had to make a pause and decide what the definition of a delayed flight would be for the project because this is what would be determining if I could drop or not any other columns and/or rows.
5.	Business Model (Revenue Model)	For the ML the workflow was pretty straight forward by starting defining the target, which was the FLIGHT_STATUS, and then dropping it alongside the DEP_DELAY (for the first set of models only) from the dataframe to define X (features). With this done, I split the data with a 25 and 75% for the test and used a typical rate of 42.
6.	Scalability of the Solution	Doing the MLP Deep Neural Network was more difficult and time consuming due to the high number of tests needed and the size of the dataset, which was reduced from +7 million to around +4 million rows by limiting the study to the top 20 destination cities.