Project Flow

TEAM ID	**PNT2022TMID50880**
PROJECT NAME	**Developing flight delay prediction model using Machine Learning and hypothesis**
TEAM MEMBER	**Muruganandhi, DIvya , Pavithra , Vishwa abirami**

Flight Data: The flight data is from the year 2007 and 2008, which is taken from Bureau of Transportation Statics. Every flight has many variables, which give detailed information about the specific flight [8].

1. Flight Number

2. Carrier {American, United}

3. Destination {Airport Code: SFO}

4. Origin {Airport Code: LAX}

5. Date {MM/DD/YY}

6. Day of Week {Mon, Tue, Wed, Thu, Fri, Sat, Sun}

7. Scheduled Departure Time {HH:MM AM/PM}

8. Actual Arrival Time {HH:MM AM/PM}

9. Actual Departure Time {HH:MM AM/PM}

10. Minutes Late {+Late/-Early}

11. Scheduled Arrival Time {HH:MM AM/PM} 7 The 2008 on-time performance data contains 7 million records. The average number of daily flights is 19,178, with 24 data elements included in the flight database.

In this section, there is an overview of the process of data mining and data modeling, from collecting the data, through the data preparation and finally the data modeling. Data cleaning and formatting can be considered as the most critical part of the whole project according to several data scientists . Figure 1 shows how the process of data mining works to extract knowledge using algorithms

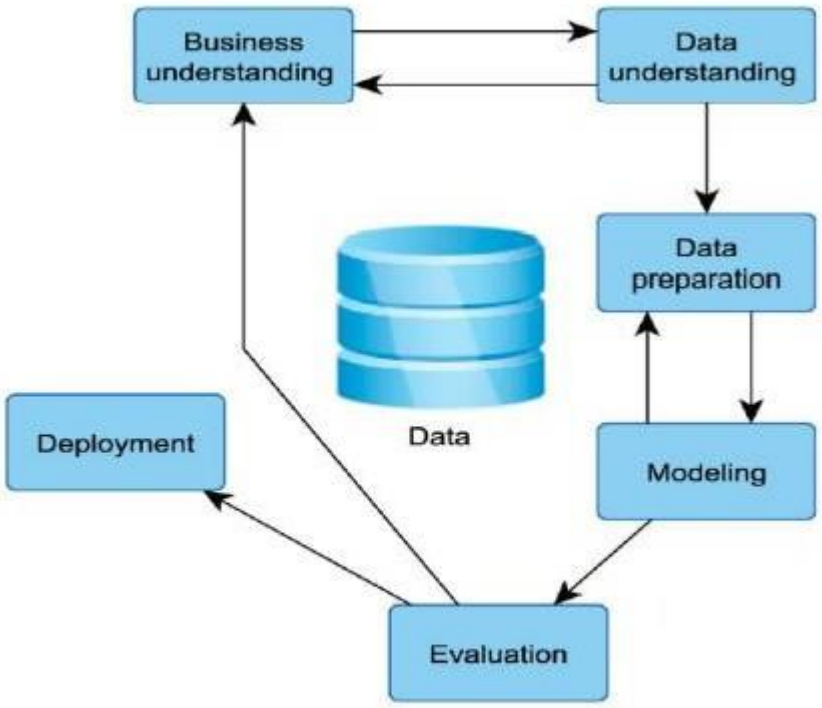

Figure 1: Project Development

DATA COLLECTION 10 Once the project undertaking is completely comprehended, our subsequent stage is to gather the information that is required for future model building. The information accumulation was an issue as data was not situated at a single source. The data was kept in unique information design. To achieve the end goal, it requires a clear understanding of the correct location of the data.

As we can see in Figure , the US Bureau of Transportation Statistics gives detailed information on every single household flight, which incorporates their booking and take off circumstances and real takeoff, origin, destination, date, and carrier. We consolidated a portion of the information properties with Local Climatological Data from National Oceanic and Atmospheric Administration (NOAA) to shape a join data set. Since the datasets for every year are very massive, we decrease our concentration to one-year, i.e., 2008, which as of now contains 1 million records for the most significant airplane terminals. In this venture, we have taken 2007 as our preparation set and 2008 as our test set. Handling speed is a noteworthy thought since the machine learning methodology that functions admirably on smaller datasets cause issues with the Jupyter Notebook establishments on our PCs.