Detailed Project Report (DPR)

Flight Price Prediction

BISWAJIT JENA

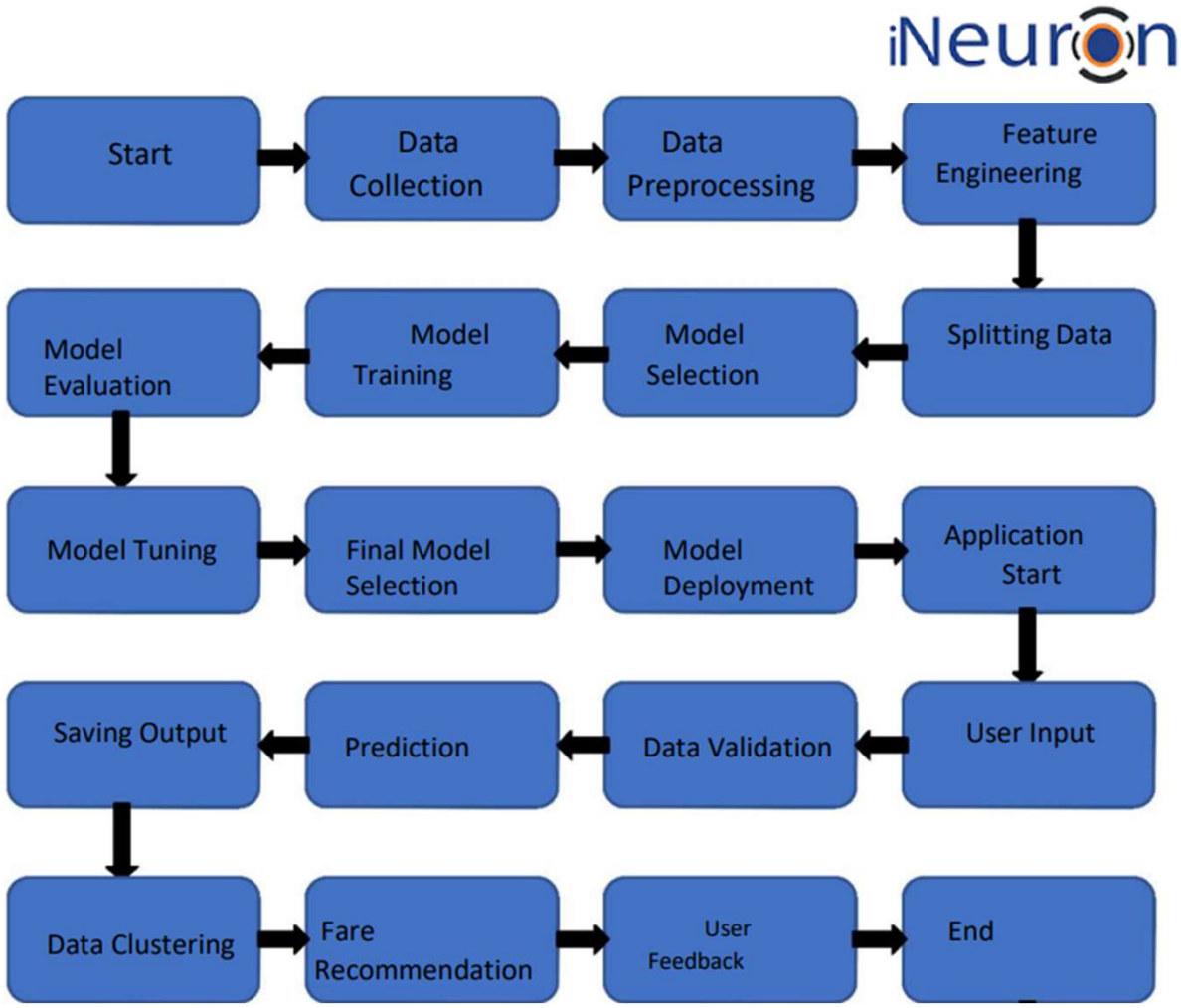
1. Introduction

This project aims to predict flight prices using machine learning algorithms. The system will process a dataset containing flight details, perform exploratory data analysis (EDA), preprocess the data, build predictive models, and deploy a web application for real-time flight price predictions. This report outlines the architecture, wireframe, and low-level design of the project.

2. Architecture Design

The architecture design document provides an in-depth overview of the system's structure and the workflow involved in the flight price prediction model. The main components are:

Architecture



Dataset

The dataset for flight price prediction includes flight details such as airline, source, destination, total stops, date and time of journey, and flight price. The dataset is preprocessed to handle null values, encode categorical variables, and generate new features such as journey day, month, and duration in minutes.

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

Data Transformation

Before sending the data into the database, data transformation is required so that data are Transform the data into a suitable format for model building. This includes filling missing values with appropriate data to ensure completeness. Categorical features are encoded into numerical values to be compatible with machine learning algorithms. Additionally, features such as the journey day and month are extracted from the journey date, and flight duration is converted into minutes for consistency. The transformation process ensures that the data is structured and clean, making it ready for the next steps in the pipeline.

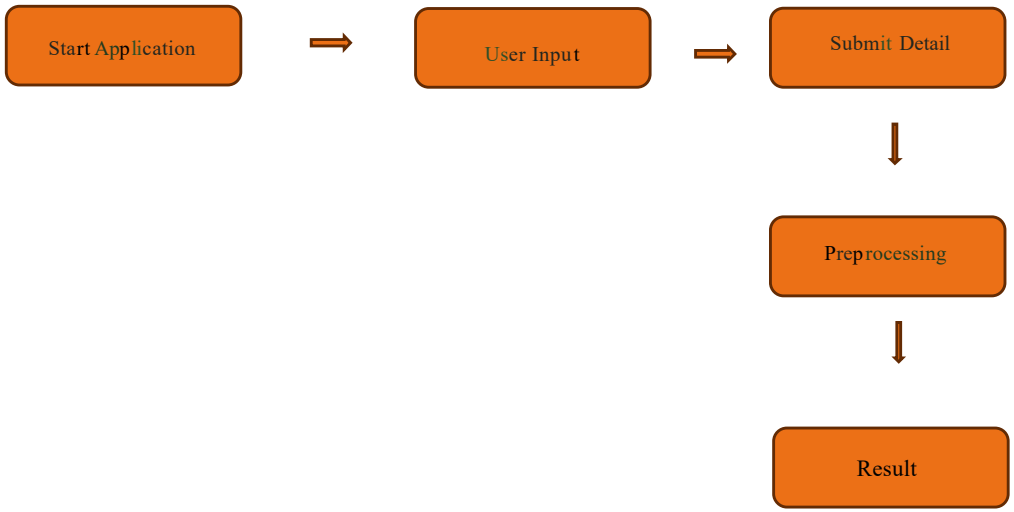
Data Preprocessing

Preprocess the data to handle inconsistencies and prepare it for model training. This involves dealing with missing values by either imputing them or removing affected rows. Categorical variables are encoded using techniques like one-hot encoding to convert them into numerical format. Numerical features are scaled to ensure uniformity across different magnitudes. This step is crucial as it ensures that the data fed into the model is clean, well-structured, and free from biases that could affect the model's performance.

Model Building

Build and evaluate multiple models to identify the best-performing algorithm. Models such as Linear Regression, Random Forest, and Gradient Boosting are trained and tested. Each model's performance is evaluated using metrics like RMSE and R^2 score. The best-performing model is selected based on its predictive accuracy and ability to generalize to new data. This step ensures that the most suitable model is chosen for deployment.

User Input / Output Workflow



. Users access a hub to input all their flight data