Project Title: **Public transportation efficiency analysis**

Problem Statement :

- Dive into the world of big data analysis with IBM Cloud Databases.
- Uncover hidden insights from vast datasets, from climate trends to social patterns.
- Consider incorporating machine learning algorithms to improve arrival
- time prediction accuracy based on historical data and traffic conditions..
- Embark on data-driven adventures, exploring the endless possibilities of big data.

Problem Definition :

- The project involves delving into big data analysis using IBM Cloud Databases.
- The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns.
- The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and preditcing the results for business intelligence.

INNOVATION

Incorporating advanced machine learning algorithms in the big data.

Introduction :

This project dives into the realm of big data analysis using IBM Cloud Databases, with a focus on extracting invaluable insights from vast datasets encompassing climate trends and social patterns. The objective is to uncover hidden connections between these domains, facilitating data-driven decision-making and fostering a deeper understanding of our changing world.

INNOVATIVE COMPONENTS

Machine Learning Algorithm:

Machine learning algorithms is essential in this project to distil meaningful insights from complex climate and social datasets. These algorithms enable pattern recognition, predictive modelling, and relationship identification, helping uncover hidden connections between climate trends and societal behaviour's. By harnessing the power of machine learning, we empower decision-makers with data-driven intelligence, facilitating proactive responses to climate challenges and societal changes.

1.Problem Definition and Understanding:

- Clearly define the problem you want to solve, including the specific insights you aim to extract from the data. Understand the business context and goals.

2.Data Collection and Integration:

- Gather relevant datasets, including climate data and social data, and integrate them into a single repository or data pipeline.

3.Data Pre-processing:

- Clean and pre-process the data. Handle missing values, outliers, and data quality issues.
- Normalize or scale features as necessary.

4.Feature Engineering:

- Create meaningful features from the data that can improve model performance.
- This might involve time-based features for climate data or sentiment scores for social data.

5.Data Splitting:

- Divide the data into training, validation, and test sets to evaluate and validate machine learning models effectively.

6.Selecting Machine Learning Algorithms:

- Choose appropriate machine learning algorithms based on the nature of the problem.
- For climate data, time-series forecasting models like ARIMA or LSTM may be suitable, while for social data, NLP models such as LSTM or Transformer-based models can be used.

7.Model Training:

- Train the selected machine learning models on the training dataset. Tune hyper parameters and monitor model performance on the validation set.

8.Model Evaluation:

- Evaluate model performance using appropriate metrics.
- For climate trends, metrics like RMSE or MAE may be relevant, while for social patterns, accuracy, F1-score, or AUC can be used.

9.Ensemble Learning (Optional):

- Consider using ensemble learning techniques, such as Random Forests or Gradient Boosting, to improve model robustness and accuracy.

10.Interpretable AI (Optional):

- If needed, use interpretable machine learning models or techniques (e.g., SHAP values) to explain how the model arrives at its predictions, especially for business stakeholders.

11.Hyper parameter Tuning:

- Optimize hyper parameters to fine-tune model performance.
- This can be done manually or using automated techniques like grid search or Bayesian optimization.

12.Model Deployment:

- Deploy the trained machine learning models in a production environment, either as batch processes or real-time services, to generate insights.

13.Continuous Monitoring and Maintenance:

- Implement monitoring to track the performance of deployed models.
- Regularly update models to adapt to changing data and maintain model accuracy.

14.Ethical Considerations and Bias Mitigation:

- Ensure that models are trained and deployed in an ethical manner.
- Detect and mitigate biases in the data and algorithms, especially for social data analysis.

15.Visualization and Reporting:

- Create visualizations and reports to present the extracted insights in an understandable and actionable format for business intelligence.

16.Feedback Loop and Stakeholder Engagement:

- Maintain a feedback loop with stakeholders to gather input on the utility of insights and any necessary adjustments to the analysis process.

17.Documentation and Knowledge Sharing:

- Document the entire machine learning pipeline, including data sources, pre-processing steps, model architectures, and deployment procedures for knowledge sharing and future reference.

CODE FOR THE GIVEN PROBLEM STATEMENT:

INPUT 1 :

```
%matplotlib inline
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import datetime
import os
from math import sqrt
import warnings

## For Multiple Output in single cell
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
warnings.filterwarnings('ignore')
```

INPUT 2 :

```
data = pd.read_csv('../input/unisys/ptsboardingsummary/20140711.CSV')
data.shape
data.head(10)
```

OUTPUT 1:

(10857234, 6)

OUTPUT 2 :

TripID	RouteID	StopID	StopName	WeekBeginning	NumberOfBoardings
0	23631	100	14156 181 Cross Rd	2013-06-30 00:00:00	1
1	23631	100	14144 177 Cross Rd	2013-06-30 00:00:00	1
2	23632	100	14132 175 Cross Rd	2013-06-30 00:00:00	1
3	23633	100	12266 Zone A Arndale Interchange	2013-06-30 00:00:00	2
4	23633	100	14147 178 Cross Rd	2013-06-30 00:00:00	1
5	23634	100	13907 9A Marion Rd	2013-06-30 00:00:00	1
6	23634	100	14132 175 Cross Rd	2013-06-30 00:00:00	1
7	23634	100	13335 9A Holbrooks Rd	2013-06-30 00:00:00	1
8	23634	100	13875 9 Marion Rd	2013-06-30 00:00:00	1
9	23634	100	13045 206 Holbrooks Rd	2013-06-30 00:00:00	1

INPUT 3 :

```
out_geo = pd.read_csv('../input/outgeo/output_geo.csv')
out_geo.shape
out_geo.head()
```

OUTPUT 3 :

(4165, 10)

Conclusion :

Transport systems play a key role in advanced societies nowadays. In this paper, we conducted a literature review study of travel and arrival time prediction models. The research scope we investigated is restricted to road networks. Our review has focused on several aspects which can influence traffic management and control. Intelligent traffic management and control could be one of the solutions to tackle the challenge of increasing travel demand, CO2 emissions, safety concerns, and wasted fuels. Travel and arrival time are very useful indicators of traffic and are vital components of intelligent transportation systems. Therefore, we notice considerable research interest in modelling and predicting these traffic indicators.