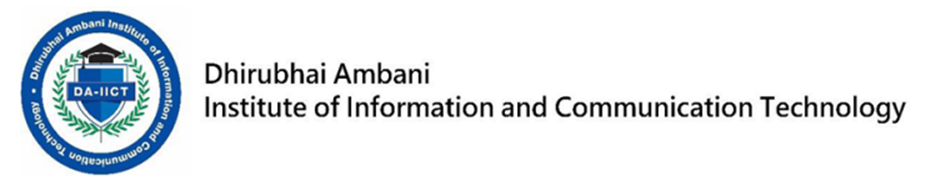
****

**IE-406 – Machine Learning**

*Project Proposal*

**“TrainTripper: ML-Enhanced Predictive Ticket Cost Estimation and Smart Train Selection”**

*Date*

**25th August 2023**

*By*

**Bhavya Rajdev (202001067)**

**Maharth Thakar (202001069)**

**Dhruv Prajapati (202001103)**

**1. Motivation**

The project is driven by the aim to simplify travel planning for users. With the increasing complexity of train travel logistics, it becomes very difficult for travelers to take into account multiple factors such as travel time, travel cost, night travels, multi-train itineraries. This project seeks to alleviate this challenge by integrating machine learning to forecast ticket prices and offer intelligent train suggestions. By doing so, users will experience a smoother journey planning process, enabling them to make well-informed decisions efficiently.

Moreover, the absence of comprehensive recommendation systems for train journeys involving multiple connections is a significant gap. Unlike flights, existing applications often lack the capability to suggest optimal multi-train itineraries. This project fills this void by not only considering direct routes but also accommodating complex travel scenarios. The focus on seamless multi-train travel aligns with the evolving landscape of transportation technology. By offering a user-centric interface and predictive insights, TrainTripper aims to revolutionize train travel planning, making it as efficient and user-friendly as other modern travel options.

**2. Literature Review**

In the context of our project, the emphasis of the literature review is directed towards evaluating existing train ticket booking platforms rather than research paper review. This approach is chosen because the project aims to address practical challenges faced by users in their travel planning endeavors. By analyzing real-world platforms like IRCTC and IXIGO, we can gain insights into the current state of train travel planning and identify opportunities for improvement that directly benefit users' experiences.

Examining existing websites such as IRCTC and IXIGO reveals a mix of advantageous features and limitations. Noteworthy advantages include the provision of user-friendly filters that allow travelers to narrow down their search by selecting specific classes (sleeper, AC, etc.) and quotas. Moreover, the ability to sort search results based on arrival and departure times, as well as distance, adds to the convenience of users. However, these platforms exhibit certain drawbacks, such as a tendency to avoid and prefer recommending night journeys and a lack of options for multi-train itineraries, which can be particularly useful for complex travel scenarios.

In this project, we intend to address the limitations identified in existing platforms. Our application will enhance the user experience by utilizing machine learning to recommend optimal train options, irrespective of travel times, and by providing multi-train travel itineraries for situations where direct routes are unavailable. These improvements aim to offer users a more comprehensive and user-centric approach to train travel planning, thereby bridging the gaps identified in the current offerings and elevating the overall travel planning experience.

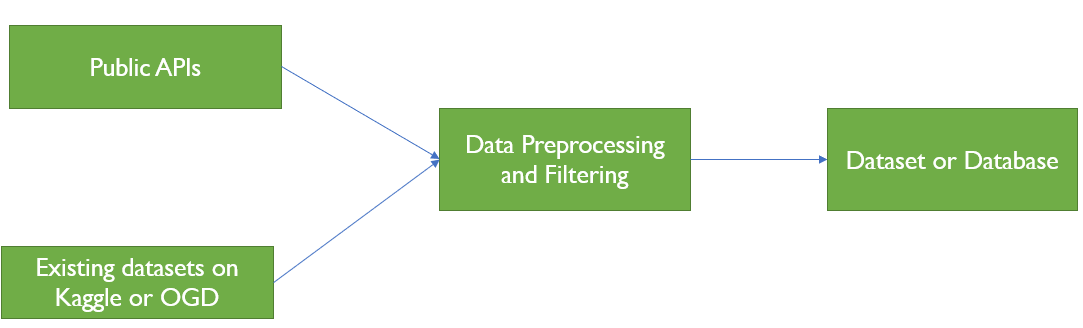
**3. Proposed Methodology**

Our goal is to develop a software application that recommends train options to users based on their source and destination stations, as well as departure or arrival dates. The machine learning component of our project is focused on predicting ticket prices for the recommended trains.

**3.1 Data Collection and filtering**

We will explore online government datasets and public APIs provided by clients of IRCTC (Indian Railway) to gather relevant data. This data will include information about train stations, cities, states, trains, their schedules, and ticket prices.There will be two phases, data collection and preprocessing. In the data collection phase, we plan to investigate various sources like **Open Government Data** (OGD) and datasets available on platforms like **Kaggle**. We will also explore **public APIs** like IndianRailways API to extract relevant train-related information. Our goal is to gather comprehensive data sets that align with the project's requirements, ensuring a robust foundation for accurate recommendations.

Also we would require preprocessing of the data, since API data will always be in **JSON format**, we will have to clean and combine the relevant data for further price prediction model and train selection algorithm. We will be using normal **Python** Libraries like **requests, pandas, matplotlib.**

****

**3.2 Price Predictor Model:**

We have planned to create a machine learning model that predicts ticket prices. This model will take inputs such as the train name, source and destination stations, and travel date. It will output the predicted ticket price for the given parameters.

* **Model selection**: The process of model selection will be done, considering the **linear nature of the data matrix**—where examples of potential models such as **linear regression,** **neural networks**, support vector regression (**SVR**), decision tree regression (**DTR**), and random forest regression (**RFR**).
* **Validation**: We intend to validate the output through various ways. Firstly, **human validation** will be conducted, involving the **cross-checking** of predicted prices for nearby stations such as Ahmedabad (Ahm) and Surat (Sur) against ground-truth values. Additionally, we will adopt the **train-test split** methodology, effectively **segregating the dataset** for comprehensive testing. This robust validation strategy ensures the reliability and accuracy of our model's predictions.Furthermore, we will introduce **k-fold cross-validation**, which involves partitioning the dataset into 'k' subsets and iteratively using 'k-1' subsets for training and the remaining subset for validation
* **Deploying** : In the deployment phase, we will deploy the model as a **'.h' file** locally, serving as a tool for making predictions. This local deployment will enable the model to generate predictions based on various inputs.

**Tools and Technologies Used**: In our project, we will be using **Python** and its libraries **pandas, matplotlib, sklearn,tensorflow etc.**  We will also be using **PyCharm** and **Jupyter** Notebook for our project

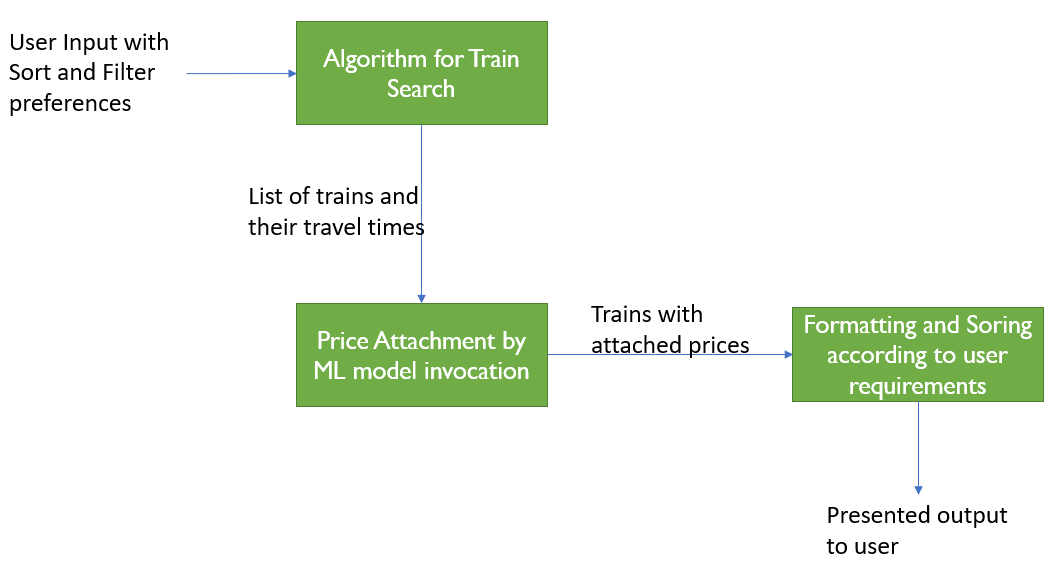
**3.3 Algorithm for Train Recommendation**

We'll design an algorithm that takes user inputs including arrival and departure stations, as well as the travel date. This algorithm will then generate a list of train names that are suitable for the user's journey.

* We will implement single train or double train journey.Also provide the nearest train station travel option in the output.
* **Direct Train Connection Search**: In this feature, the algorithm will examine trains departing from a given station at the user's specified time and evaluate whether any of these departing trains have a **direct route to the intended destination station.**
* **Double train**: In this feature, the algorithm focuses on optimizing train changes during the journey by specifically considering major junctions. When recommending train options involving transfers, the algorithm will prioritize connections that **require changing trains** at significant railway junctions. We will be selecting major junctions which will be easier for computer and other optimizations will be caught while performing the experiments.
* By incorporating Django's capabilities, we'll **construct APIs**, setting up essential endpoints which will help further in the project like the ML model and also harness **database creation** to enhance programming efficiency, ensuring seamless data management.

**3.4 Integration of ML Model and Algorithm**

The train names recommended by our algorithm will be **integrated into our machine learning model**. This integration enables the model to enrich each recommended train option with anticipated ticket prices. Subsequently, the final data will undergo **formatting** to yield a comprehensive output. The final output will encompass a catalog of suggested train choices, accompanied by their **forecasted ticket prices**.



**4. Computational Requirements**

For the entire duration of the project, we will be leveraging our personal computing resources. These resources prove adequate for effectively collecting, storing, and managing the necessary dataset, as well as constructing, deploying and invoking machine learning models. By employing our local machines, our goal is to ensure the streamlined handling of data throughout all phases of the project.

The computational demands of our project are well-contained within our current capabilities, negating the need for additional computational resources from the college. However, it's worth noting that as part of the project's future scope, the transition to a cloud environment is plausible. Such a shift could enhance the software application's accessibility, enabling users to interact with it more readily. As of now, our project's computational needs are well-matched with our personal resources, paving the way for seamless execution.

**5. Tentative Timeline**

The project encompasses several key sections that collectively contribute to its successful realization. These sections include "Data Collection and Filtering," "Price Predictor Model," "Algorithm for Train Recommendation," and "Integration of ML Model and Algorithm."

The development of the "Price Predictor Model" constitutes a crucial element of the project. This stage involves the design, training, and fine-tuning of a machine learning model capable of predicting ticket prices based on multiple factors. The accuracy and reliability of this model will significantly impact the accuracy of the cost predictions provided to users. In the "Data Collection and Filtering" phase, extensive efforts will be dedicated to sourcing and compiling relevant data, forming the foundation for subsequent stages of the project. This phase is also expected to take more time as the data required is not exactly available in the format that the model requires.

The other sections will also affect the timeline of the project but it is expected to affect not as much as the model creation section. It is anticipated that the project will be completed before the presentations. We will ensure that the project remains on track for completion before the final presentation.

**6. Results Expected**

The primary objective of this project is to develop a machine learning model capable of predicting future prices of train tickets based on various factors such as train name, source and destination stations, and travel date. The developed price prediction model is expected to give promising results during evaluation as the dependent factors of the train ticket prices such as distance, coach class etc. seem very clear and many many models are available in the market for prediction of this kind of standard dataset.

The secondary goal of the project is to create a software application that assists users in selecting the most suitable train options based on their preferences, such as source and destination stations, as well as departure or arrival dates. The algorithm developed for train recommendation will sort and filter available train options to provide tailored recommendations. We are selecting the Command Line creation route for using the algorithm of the project as this is easier compared to interface creation and it is not part of the primary goal. The primary goal is data handling and modeling which would not be compromised.

While the project has achieved significant milestones, there is one aspect that we will not be addressing in this phase due to its complexity. The interface development, which requires frontend skills, is beyond the scope of this course. As a result, the project will solely operate using command-line inputs provided by users. However, it's important to note that this limitation is not a permanent constraint and offers opportunities for improvement in the future.

An inherent limitation of the project lies in the fact that the prices provided to users are not direct reflections of the actual train ticket prices. Instead, they are predictions generated by our model. Addressing this limitation is currently unfeasible due to potential constraints in coupling our system with the API providers of Indian Railways. Incorporating real-time price retrieval would require a level of interaction that may exceed the bounds of a free account, impacting the scalability and sustainability of the application. Nonetheless, this project provides a foundation that can be expanded upon and refined in subsequent iterations.