

## **Predicting Tourist Arrivals using Machine Learning and Internet Search Index**

### **Problem statement**

Design and implement a machine learning model that predicts tourist arrivals to a specific destination using Internet search index data as a key input. The goal of this competition is to develop an accurate and robust predictive model that can assist tourism authorities and businesses in forecasting tourist arrivals, thereby enabling better resource allocation, marketing strategies, and overall destination management.

### **Dataset**

The team will extract data on tourist arrivals, internet search indexes, and other relevant variables such as hotel booking and flight search. The data will cover the weekly intervals between 2010 and 2022. The sample will include popular tourist destinations in India, such as Shimla, Jammu and Kashmir, and Goa.

The models will use the collected data to predict future tourist arrivals. The internet search indexes will be collected from Google Trends and will be included in the models as a predictor variable.

### **Task**

The task is to build a predictive model that can forecast the number of tourist arrivals for future time periods based on the provided historical tourist arrival data and Internet search index data.

### **Evaluation Criteria:**

1. **Prediction Accuracy:** Participants' models will be evaluated primarily based on their ability to accurately predict tourist arrivals for future time periods. The evaluation metric for prediction accuracy could be Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or any other appropriate metric.
2. **Model Robustness:** Models should be evaluated for their robustness and generalizability. They should perform well on unseen data and avoid overfitting.
3. **Feature Engineering:** Participants are encouraged to perform creative feature engineering to extract meaningful insights from the provided data. The quality and relevance of features used in the model will be considered.

4. **Data Preprocessing:** Data preprocessing steps, such as handling missing values, outlier detection, and data scaling, will be evaluated. Well-documented and thoughtful preprocessing methods will be rewarded.
5. **Interpretability:** While predictive performance is crucial, models that provide interpretable insights into the relationship between tourist arrivals and Internet search index data will be favoured.
6. **Model Documentation:** Participants should provide clear documentation of their model, including explanations of algorithm choices, feature selection, and any relevant assumptions.
7. **Code Quality:** The quality of the code, including readability, structure, and comments, will be taken into account during evaluation.
8. **Presentation:** Participants may be required to present their findings and models to a panel of judges, so clear and effective communication of results will be important.
9. **Innovation:** Bonus points may be awarded for innovative approaches, techniques, or insights that go beyond traditional machine learning methods.
10. **Scalability:** While the primary focus is on prediction accuracy, participants are encouraged to consider the scalability and efficiency of their models for potential real-world deployment.

The winning solution will be the one that demonstrates the highest prediction accuracy and overall model quality based on the evaluation criteria outlined above.