

**Post Graduate Diploma in Big Data Analytics (PG-DBDA)**

**March 2023 Batch (New Delhi)**

|  |  |
| --- | --- |
| **Group Number** | Group – 1 |
| **Guide** | Dr. Priyanka Jain |
| **Group Members** | 1.Monica Jha  2.Lokesh Sali  3.Shivani Phuke |
| **Title** | **Taxi Hotspot Clustering and proximity recommendation system** |
| **Objective** | The objective of the project is to solve the business problem of identifying areas where they can build up the waiting spots for its customers, so that there is a surety of getting a cab.  And allowing their customers to know what could be the possible nearest waiting points for them. |
| **Abstract** | Through this project, we try to predict the best possible way to find the number of hotspots feasible for taxis in a city.  As input, we take the pickup point’s latitude and longitude as coordinates. Using these coordinates, we try to predict the regions, where there could be highest possibility of getting a taxi.  Using this, the business can report the waiting spots/hot spots to be made. So that the customers can go to these hotspots to get a taxi booked.  Along with this, the project can be further extended as an application, where based on customers current latitude and longitude, the model can recommend him/her the nearest hotspot. |
| **Project Workflow/**  **Architecture** |  |
| **Scope Of Work** | The project is divided into following phases:   1. Collecting pickup and drop off points data for a city of a year. 2. Based on this collected data, preprocessing it to mould it as per the need. 3. Once the data is ready, we will be using the concept of K means clustering to train the model so that, it can first map out the Regions where the waiting points can be created. 4. Once the regions are mapped out, we can plot the Hot spots/waiting spots using centroids. 5. Lastly, using this trained model, we can test out the prediction for the customer to let him/her know the nearest hot spot |
| **Technologies Used** | 1. Machine Learning: Google Collab for generating model 2. Programming language: Python with libraries: NumPy, Pandas, matplotlib and seaborn for visualization. 3. Algorithm used: K means algorithm 4. Dataset source: Kaggle |
| **Application** | The model can be used for:   1. Predicting Number of hotspots feasible in a city. 2. Clustering the regions basis of frequent pickup coordinates 3. Recommend the nearest hotspot to the customer based on his current location |
| **Project Timelines**  **(Total:120 hours)** | Group formation (Guide Allocation and Team Selection): 14 August 2023.  Project Completion date: 4th September 2023 |