***Regression Assignments***

1. **Car price prediction:**

* Kaggle problem statement **link**: <https://www.kaggle.com/tunguz/used-car-auction-prices/code>
* We need to predict car prices based on various features provided.

1. **Diamond price prediction:**

* Kaggle problem statement **link**: https://www.kaggle.com/datasets/shivam2503/diamonds

The problem is little bit straight forward regarding regression.

We have to predict the price of diamond on basis of different features.

1. **Insurance price prediction:**

* Kaggle problem statement **link**: https://www.kaggle.com/code/hely333/eda-regression/data

The problem is highly unbalanced.

In this dataset smokers having high insurance charges but their counts are less and non smokers have low insurance charges increases with age but their counts are much greater than non smokers.

We have to predict insurance charges on basis of different features.

1. **Real Estate price prediction:**

* UCI problem statement **link**: <https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>
* Collab Link:

<https://colab.research.google.com/drive/14g2rkvl5fzlYAtjhztUHsGSejNiloPs6>

https://colab.research.google.com/drive/1KOjLEhjS\_PQeQyPGtClcKGTtMJARsrQB

1. **Spotify Dataset:**

* This is the dataset of Spotify songs picked from the Kaggle.
* **Link**: <https://www.kaggle.com/pepepython/spotify-huge-database-daily-charts-over-3-years>
* This contains 2 csv files.
* "Database to Calculate Popularity" includes all the daily entries (8mln+) for the songs which made it to the top 200. Among these data, quite intuitively, you will find the same song being in the charts for more than one day. We then created a popularity score, unique for a given song in a given country, which took into account the position in the charts and the days it stayed there
* "Final Database" includes many data for each song. It aggregates the popularity for songs into a single score for each. For each song several variables were retrieved by using Spotify's API (such as artist, country, genre, …)
* Working on the pre-processing, data cleaning, reduction. Transformation : https://colab.research.google.com/drive/17RUWyF-b\_RkRU-IDwEHOVFqOZ1QtK0YK#scrollTo=Z-VWLMTHL0wm

1. **US car price prediction:**

* **Description**: This machine learning problem is for developing a model using different regression algorithms for predicting the final price of the car depending on various factors like brand, colour, model etc.
* **Dataset references**:
* <https://medium.com/@chenlemuge/a-houswifes-journey-toward-data-scientist-d1354288e9ac>
* <https://github.com/Tselmeg-C/Udacity-Data_Scientist_Nanodegree_Project_One-Car_Auction_Dataset_Analysis>
* <https://www.kaggle.com/doaaalsenani/usa-cers-dataset>
* **Collab link:** <https://colab.research.google.com/drive/138aohUaWZD0kio1ztRq6FgdTRCpKSwrD>

1. **NYC Sales prediction:**

* Kaggle problem statement **link** : https://www.kaggle.com/datasets/new-york-city/nyc-property-sales

The problem is having large number of null values .

In this dataset we have many columns of features so, we have to select best features on which our result is dependent

I have applied Ensemble model random forest regressor without applying encoder to labels.

We have to predict NYC sales on basis of different features.