REPORT MINI PROJECT

**HOUSING PRICE**

**PREDICTION ML**

**PROJECT**

**SEM 3**

**Batch- 2019-2023**

**Dhruv Punetha**

**Univ R No- 2014639**

**Section- ML/AI**

**PROBLEM STATEMENT**

Developing a machine learning model that can predict housing prices in a certain region (California in our case) based on various factors such as age of house, number of beds, number of rooms in the house, population in that region ,households in surroundings, median house price in that area as well as median income on the input provided by the user. Some other values are provided such as longitude as well as latitude and the ocean proximity of the property.

**MOTIVATION FOR THIS TOPIC**

The motivation behind picking this topic was the fact that this is a good project for students who are new in machine learning. It lets them take a grasp machine learning concepts as this lets them hands-on practice on regression concepts, various python libraries as well as other machine learning concepts. As for its real life implementation. A model like this would be very valuable for a real estate agent who require information like this on a daily basis. Predicting housing prices also help people who plan to buy a house so they can know the prices in the future, then they can plan their finance well as well as let them find a perfect house that fulfills all their requirements in the best price possible. Not just buyers the knowledge about a house’s value can help someone planning on selling it to get a fair price and not be cheated.

**METHODOLOGY**

SOFTWARE USED:

* python 3.8
* windows 10 operating system
* jupyter notebook (From Anaconda navigator GUI)

DATASET USED:

For this project California Housing dataset was used from kaggle (<https://www.kaggle.com/camnugent/california-housing-prices>) to train our model to predict prices for input given by the user.

LIBRARIES USED:

To work on this project we need various libraries such as pandas, matplotlib, numpy, seaborn and sklearn from python.

These libraries help us modify as well as visualize our dataset as per our requirements so that we can have a better grasp over the data and can modify it as per our requirements.

MODELS USED:

While working on machine learning projects we never know for certain which model will perform the best for this reason we have to use various models and based on their performance decide the model best working on our dataset. For this project Linear Regression, Decision Tree Regressor and Random Forest Regressor were used. (Based on their results we find Random Forest to work the best hence we use it in the end).

STEPS USED:

* Get the dataset and download required libraries then import them in your project.
* Visualize your dataset to get a better grasp of it. After that clean the missing data by replacing it with some mean or removing it.
* Find out correlation of different features to see how much they impact the outcome. This helps in selecting important features that have major impact on the answer.
* Split your data into training and testing part by stratified shuffling train part is the part which models use to train while test data being the part which is used to see the performance of the model in the end. Test part is kept hidden from the model till the end.

* Create pipelines to do transformation on the data in a particular sequence
* Try out various models :Linear Regression, Decision Tree Regressor and Random Forest Regressor on train data and see which performs best.
* Select the best model that gives satisfactory output to be used to train our model and give it to the test data. See how it performs on the test data . If it works fine then save the model for further use. Now the model is ready
* In the end we use the model to predict housing prices for custom input.