\_\_\_

**FINAL PROJECT PROPOSAL \_\_\_\_**

|  |  |
| --- | --- |
| **Name** | **NUID** |
| **Sai Sindhu Thondapu** | **002785016** |
| **Vijaya Bharti** | **002754378** |

#### Housing Prices Prediction Project

* **DATASET:** <https://www.kaggle.com/datasets/camnugent/california-housing-prices?resource=download> ****

The dataset we will use is the "California Housing Prices" dataset which is based on data from the 1990 census. This dataset offers great opportunities for learning. The prediction task for this dataset will be to predict housing prices based on several features.

* **PROJECT IDEA:**

In this project, we will build a machine learning model that predicts the median house value for a given "district". This data has metrics like population, median income, and median housing price for each block group in California. Block groups are smallest geographical unit for which the US Census Bureau publishes data. We will here call them "districts" for the scope of this project.

Our problem is clearly a supervised learning task, because you have labelled input data. It is also clearly a regression task since we have to predict a numeric value..

* **The step-by-step process of this project includes**:
* We’ll describe and check the shape of the dataset to see the main statistical points in the dataset.
* We’ll perform extensive data exploration (by visualizing data using Scatterplot, Seaborn and Matplotlib) and then processing the data by using Custom transfer, Feature Scaling and using Pipelines for transformation.
* We will create a model to predict the house prices, for which we will train the machine using the train data by using multiple Machine Algorithms models such as Linear Regression with K-Fold Cross-Validation, Random Forest Regressor and Decision Tree Regressor.
* We will fine tune the data using our hyper parameters with grid search.
* Finally we can evaluate the final model on the test set.
* **Why we chose the same project?**

We chose the same project because in the mid-term project proposal we got the feedback to use at least 10k rows dataset, hence in this project we will be keeping the suggestion into consideration. We also feel that as it is a classification problem, last time we only used linear regression model to train and test, whereas this time we are planning to explore other Machine Learning algorithms as well as such Random Forest Regressor and Decision Tree Regressor etc.

* **REQUIRED TOOLS & LIBRARIES:**
  + Jupyter Notebook
  + Python
  + Numpy
  + Pandas
  + Scipy
  + Matplotlib
  + Seaborn
  + Sklean
* **REFERENCES:** 
  + <https://www.kaggle.com/datasets/camnugent/california-housing-prices/discussion/250628>
  + <https://kathavachhani.medium.com/data-preprocessing-using-scikit-learn-california-housing-prices-dataset-f09187c073f6>