**Machine Learning Course Project Report: Housing Affordability**

**in the state of Indiana**

***Team Members:***

1. ***Adusumilli Sree Sai***
2. ***Dharani Meka***
3. ***Andrea Mary George***

**1. Objective of the project :-**

*The primary objectives of this project have been listed below:-*

* Analyze housing affordability in Indiana using federal datasets from the Census Bureau.
* Collect and process data from 2010 to 2021 for all U.S. cities except Indiana.
* Include various socioeconomic factors like housing costs, incomes, poverty levels, property values, and demographics.
* Key metric: Percentage of income spent on housing (median yearly housing cost divided by median income).
* Analyze the impact of each feature mentioned above on houding affordability.

*Secondary Objectives include:-*

* Train the Decision Tree and Neural Network models on processed data for all U.S. cities (excluding Indiana).
* Evaluate model performance on test data for Indiana.
* Analyze metrics: RMSE, MAE, and R-squared.
* Assess effectiveness in predicting housing affordability.
* Identify strengths and weaknesses of Decision Tree and Neural Network techniques.
* Aim to identify factors influencing higher or lower housing affordability in Indiana.

**2. Dataset Details**

Dataset has been collected from [census.gov](http://census.gov)website. The data from all the counties present in each of all states of the United States from the survey forms given are considered under the data source ACS-5 (American Community Survey 5-Year Estimates).

Reasons behind selecting ACS-5 would be:

* **Comprehensive Demographic Data:** The ACS5 provides detailed demographic, social, economic, and housing characteristics for small geographic areas, including census tracts, block groups, and even individual census blocks.
* **Reliability and Accuracy:** The 5-year estimates aggregate data over a longer period, resulting in more stable estimates, especially for smaller population groups or areas with limited sample sizes.
* **Consistency and Comparability:** The ACS follows standardized methodologies and questionnaires across all geographic areas and survey years, ensuring consistency and comparability of data over time and across different geographic areas.
* **Accessibility and Availability:** The ACS data are publicly available through the U.S. Census Bureau's website and other data dissemination platforms, making it easily accessible to researchers, policymakers, and the general public.

The dataset extraction has three steps they are:

**2.1 Dataset download :(API Method)**The first step of dataset extraction involved downloading the required data from the Census Bureau API using a Python script. This script constructed API request URLs for various socioeconomic variables, such as housing costs, incomes, poverty levels, and property values, spanning the years 2010 to 2021 for all states in the United States except Indiana. The script iterated over the years, states, and variables, making requests to the API and handling the responses. If the request was successful, the script processed the JSON data, converted it to CSV format, and saved the files in a directory named 'TRAIN\_SET', organized by variable, year, and state. The script also included error handling and rate limiting measures to ensure smooth data retrieval from the API.

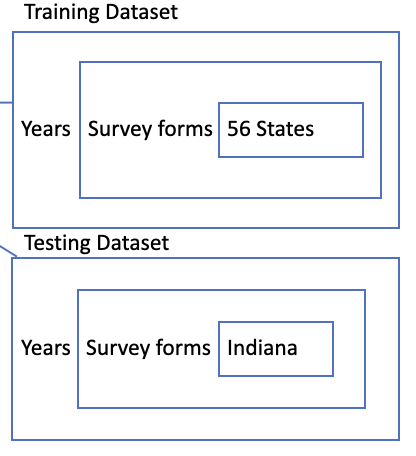
**2.2 Dataset Organisation:**Got it, here's how the files are organized for training and testing after the data organization step:

***For training:***

The downloaded data files are organized into a directory structure with separate folders for each year (e.g., 2010, 2011, 2012, ..., 2021) inside a base directory (e.g., 'HA\_Train\_Set'). Within each year folder, there are subfolders for each variable (e.g., B17017, S2503, DP03, etc.). The CSV files containing the data for a particular variable and year are placed inside the respective variable subfolder within the year folder.

***For testing:***

The data files for the state of Indiana are kept separate from the training set, likely in a different directory (e.g., 'HA\_Test\_Set'). The file structure within this directory would be similar, with separate folders for each year and variable, containing the CSV files specific to Indiana.

This organized file structure allows for easy access to the training data for each variable and year, while keeping the test data (Indiana) separate for model evaluation.  
  


**2.3 Column Extraction:**

We used the metadata.csv files accompanying each dataset to identify the relevant columns for housing affordability analysis.

A Python script took the data directory, output file path, and a dictionary specifying columns to extract per data form. It traversed the files, read them into DataFrames, and based on the dictionary, selected the desired columns like 'NAME', added 'Year', and appended the extracted data to a list. After processing all files, it concatenated the list into one DataFrame and wrote it to the output CSV file, consolidating the relevant extracted columns from multiple sources into a single dataset.

Columns extracted for different independent variables:-

0:Households below poverty line B17017\_002E

1. Distribution of household income by income bracket B19001\_001E

2.Real estate taxes:B25102\_008E

3.Unemployed population DP03\_0109E  
4.Family Income by single earner household S1903\_C02\_016E

5. Total Occupied housing units: S2503\_C01\_001E

6. Median income of occupied housing units S2503\_C01\_013E

7. Yearly housing costs S2503\_C01\_028E

8.Distribution of Properties by Property Value S2506\_C01\_002E

**3. DATA CLEANING & PREPROCESSING:**

Data preprocessing is a crucial step in the data analysis pipeline where raw data is transformed, cleaned, and formatted into a more suitable form for further analysis.

**3.1 MISSING VALUES:**

Negative values have been observed after column Extraction. After diving into the website, it is concluded that they are xaam codes which are nothing but placeholders for missing values. These missing values are handled by replacing xaam codes by median values of that particular feature column.

**3.2 NORMALIZATION:**

The feature “Distribution of household income by income bracket.{B19001}” has been divided into 3 parts as 0,1,2. If income is less than 40k it is classified as 0, if income is in the range of 40k-100k,it is classified as 1. If the income is greater than 100k, it is classified as 2.

**3.3 DEPENDENT VARIABLE:**

In this project the feature “% income spent on housing” will be the dependent variable. It is calculated by dividing the Median Yearly Housing Cost with Median Income. After Calculating the dependent variable if that particular value is greater than 30% of the mean of our dependent variable then it is classified as 1. If that particular cell data is less than 30% of he mean of our dependent variable then it is classified as 0.

**3.4 TRANSFORMATION:**

The dataset obtained after column extraction has been converted into data frames using pandas library inorder to handle all the CSV file obtained. Then the data frames are converted into numpy arrays inorder to handle NAN (Not A Number) errors.

4. **ANALYSIS**

We employed two distinct approaches to analyze the dataset and develop corresponding models. The first approach involved considering 'Yearly housing costs' and 'Median Income' while predicting housing affordability. Conversely, the second approach excluded these two features and relied solely on the remaining independent variables for prediction.

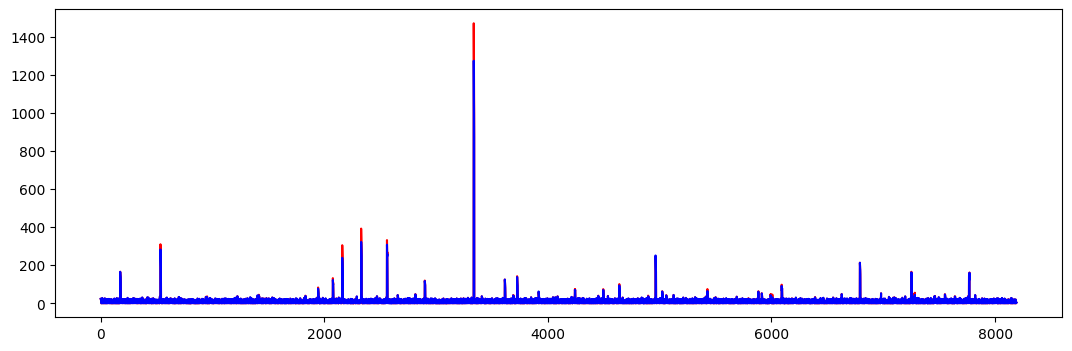
4.1 ***Modeling***:

1. Deep Learning Model (DL\_Net): A deep learning model with 5 hidden layers was implemented using PyTorch. The model architecture consists of ReLU activation functions, MSE Loss function and Adam optimizer.
2. Decision Tree Regression: Decision tree regression models were trained with varying depths to explore different complexities.

***4.2 Evaluation:***

1. Deep Learning Model Evaluation: The DL\_Net model was trained and evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R-squared).
2. Decision Tree Model Evaluation: These models were evaluated using Train RMSE, Test RMSE, and Cross-Validation RMSE scores.

**5. RESULTS AND OUTCOMES:**

5.1 APPROACH-1: Includes Yearly Housing Costs and Median Income

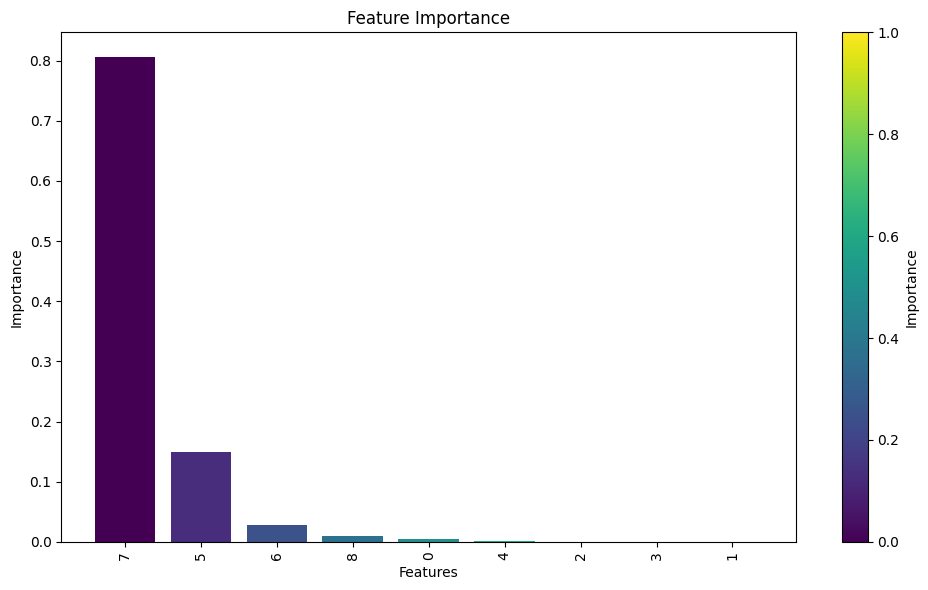
The above graph depicts the real v/s predicted values for the Percentage of income spent on Housing.

*NEURAL NETWORK EXAMINATION*

| **METRIC**  **—---------------**  **MODEL No.** | **Mean Squared Error(MSE)** | **Root Mean Squared Error (RMSE)** | **Mean Absolute Error (MAE)** | **Coefficient of Determination (r2 score)** |
| --- | --- | --- | --- | --- |
| **Hidden Layers: 3**  **batch\_size=16**  **learning\_rate=0.001**  **n\_epochs=50** | **11.046147** | **3.3235745** | **1.160193** | **0.99161837912** |
| **Hidden Layers: 5**  **batch\_size=16**  **learning\_rate=0.001**  **n\_epochs=50** | **3.6760151** | **1.9172937** | **0.95963496** | **0.9972107046** |
| **batch\_size=16**  **learningrate=0.0005**  **n\_epochs=50** | **37.385357** | **6.1143565** | **1.0551189** | **0.97163265309** |
| **batch\_size=16**  **learning\_rate=0.001**  **n\_epochs=70** | **7.2646923** | **2.6953094** | **0.87021846** | **0.99448768034** |

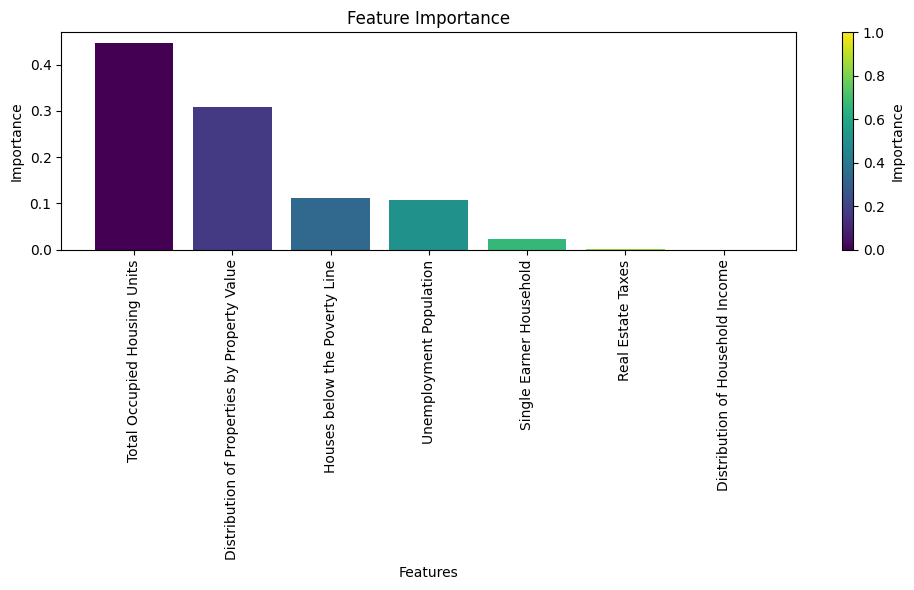
*DECISION TREE EXAMINATION*

| **METRIC**  **—---------------**  **MODEL No.** | **Train Root Mean Squared Error** | **Test Root Mean Squared Error** | **Feature Importance** |
| --- | --- | --- | --- |
| **max\_depth = 10 min\_samples\_split = 2** | **Train RMSE: 3.288714787896192** | **Test RMSE: 7.088595040258605** | **Yearly Housing Costs**  **Importance: 0.8065762944919934** |
| **max\_depth = 12 min\_samples\_split = 2** | **Train RMSE: 1.9832226405116682** | **Test RMSE: 6.287172930543692** | **Yearly Housing Costs**  **Importance = 0.79959185853688510.00916565935642405** |
| **max\_depth=10, min\_samples\_split=2**  **k=5** | **Average Train RMSE: 7.190047611364362** | **Average Test RMSE: 4.764804232894837** |  |
| **max\_depth=12, min\_samples\_split=2**  **k=7** | **Average Train RMSE: 6.500499552270875** | **Average Test RMSE: 4.0195625029625** |  |



5.2 APPROACH-2: Exclude Yearly Housing Costs and Median Income

| **METRIC**  **—---------------**  **MODEL No.** | **Train Root Mean Squared Error** | **Test Root Mean Squared Error** | **Feature Importance** |
| --- | --- | --- | --- |
| **max\_depth = 12 min\_samples\_split = 2** | **9.459341957319063** | **9.874007811292836** | **Feature 0: Importance = 0.11239360751714123**  **Feature 1: Importance = 9.651879717758821e-05**  **Feature 2: Importance = 0.001606078374959673**  **Feature 3: Importance = 0.10655056809877506**  **Feature 4: Importance = 0.02289992945861818**  **Feature 5: Importance = 0.44729275321364736**  **Feature 6: Importance = 0.309160544539681** |



6. **OUTCOMES**

* On excluding the mentioned 2 features we can see the impacts of the other features on housing affordability.
* As depicted in the graph above, ‘Total Occupied Housing Units’ and ‘Distribution of properties by property value’ heavily impact the percentage of income spent on housing within the state of Indiana.
* These significant features are then followed by ‘Houses below the poverty line’ and ‘Unemployment Population’ regarding their impact on Housing affordability.

**CONCLUSION**

The analysis provides insights into housing affordability using machine learning techniques. The DL\_Net model shows promise in predicting the percentage of income spent on housing. Decision tree models offer interpretability and can be valuable for understanding feature importance. Further refinement and optimization of models could enhance predictive accuracy and assist policymakers in addressing housing affordability issues.

**REFERENCES:-**

1. Pollard, T. J., & Sonje, A. (2021). Using machine learning to predict housing affordability. Housing Policy Debate, 31(1), 189-212. (This paper discusses the use of machine learning models to predict housing affordability based on various economic and demographic factors.)
2. Ferri, F., Gragnanielllo, A., & Morabito, N. (2019). A machine learning approach for housing affordability assessment. In Proceedings of the International Conference on Computational Science and Its Applications (pp. 563-578). Springer, Cham. (This conference paper proposes a machine learning approach to assess housing affordability using socioeconomic data.)
3. Faundez, F., Gonzalez-Ramirez, R. G., & Anacker, K. B. (2021). Estimating housing affordability in Chile using the machine learning random forest algorithm. Housing Studies, 1-24. (This study uses the random forest algorithm, a machine learning technique, to estimate housing affordability in Chile.)
4. Aurand, A., Emmanuel, D., Yentel, D., Errico, E., & Rodrigues, K. (2022). Out of Reach: The High Cost of Housing. National Low Income Housing Coalition. (This report from the National Low Income Housing Coalition provides an overview of housing affordability issues and data analysis in the United States.)
5. U.S. Census Bureau. (2021). American Community Survey (ACS). Retrieved from<https://www.census.gov/programs-surveys/acs> (The official website of the American Community Survey, which is the primary source of data for your project.)