**User Guide:**

**a. Hardware and Software Requirements**

**Hardware Requirements:**

Minimum: 8GB RAM (16GB recommended for large datasets)

Processor: Intel i5/i7 or AMD equivalent

Storage: At least 20GB free disk space

**Software Requirements:**

Operating System: Windows 10/11, macOS, or Linux

Python Version: 3.8 or later

Development Environments:

Anaconda (Recommended for package and environment management)

Jupyter Notebook (For interactive coding and visualization)

Google Colab (Cloud-based execution, optional)

**b. List of Python Libraries Used**

This project utilizes the following Python libraries for data analysis and machine learning:

|  |  |
| --- | --- |
| Library | Purpose |
| Pandas | Data manipulation and analysis |
| NumPy | Numerical computations |
| Matplotlib | Basic data visualization |
| Seaborn | Advanced data visualization |
| Scikit-learn | Machine learning models |
| Statsmodels | Statistical analysis |
| TensorFlow | Deep learning and ANN |
| Keras | High-level API for deep learning |
| PyTorch | Deep learning and ANN |
| XGBoost | Gradient boosting for ML |
| LightGBM | Efficient gradient boosting |
| CatBoost | Categorical boosting for ML |

**c. Environment Details**

Python Version: 3.8 or later

Execution Platforms:

Anaconda Environment: Recommended for managing dependencies

Jupyter Notebook: Provides an interactive environment for coding and visualization

Google Colab: Can be used for cloud-based execution, eliminating the need for local set

**Flowchart of the Research Methodology**

**Data**

Person Level Data, Household Level Data  
  
↓  
  
**Data Preprocessing**

Dataset Merging, Missing Value, Encoding, Standardization   
  
↓  
  
**Feature Engineering**

Lasso Regression for Feature Selection

Multiple Correspondence Analysis (MCA)  
  
↓  
  
**Modelling Phase**

↓

Linear Regression

Decision Tree Regression

Random Forest Regression

Extreme Gradient Boosting – XGBoost

Light Gradient Boosting Machine – LightGBM

Stacking Ensemble

ANN – Artificial Neural Network

↓

**Interpretation**

MAPE and R^2

**Machine Learning & Deep Learning Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R^2 Score | | MAPE Score | |
|  | Train | Test | Train | Test |
| Linear Regression | 0.654 | 0.651 | 0.0074 | 0.076 |
| Ridge Regression | 0.654 | 0.651 | 0.02 | 0.0276 |
| XGBoost | 0.698 | 0.68 | 0.025 | 0.026 |
| Gradient Boosting | 0.71 | 0.68 | 0.025 | 0.026 |
| ANN Regression | 0.7093 | 0.6893 | 0.025184 | 0.026011 |
| LightGBM | 0.6953 | 0.6890 | 2.57 | 2.60 |

The best model accuracy was obtained by Gradient Boosting.

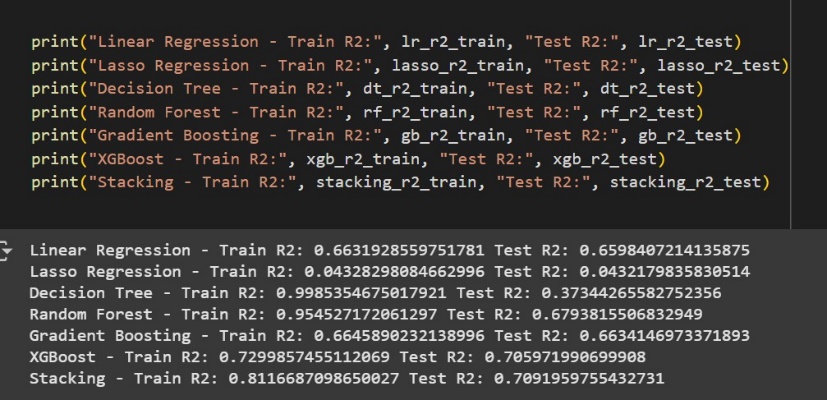
**Files Required**

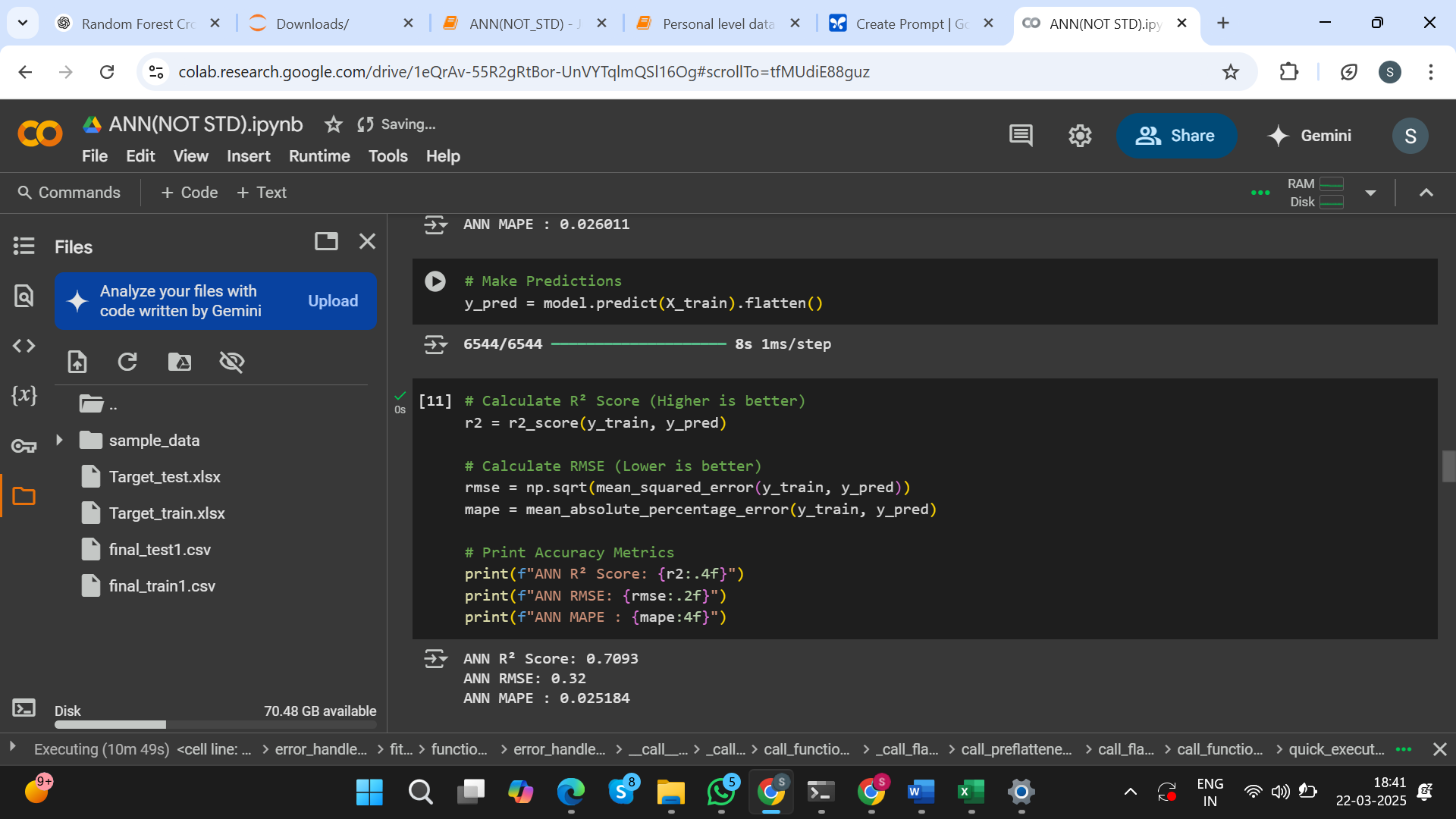
[**link to data**](https://drive.google.com/drive/u/0/folders/1qlEVna0zi-CEdc5bYsafJS9nYZdDzeXK/view)

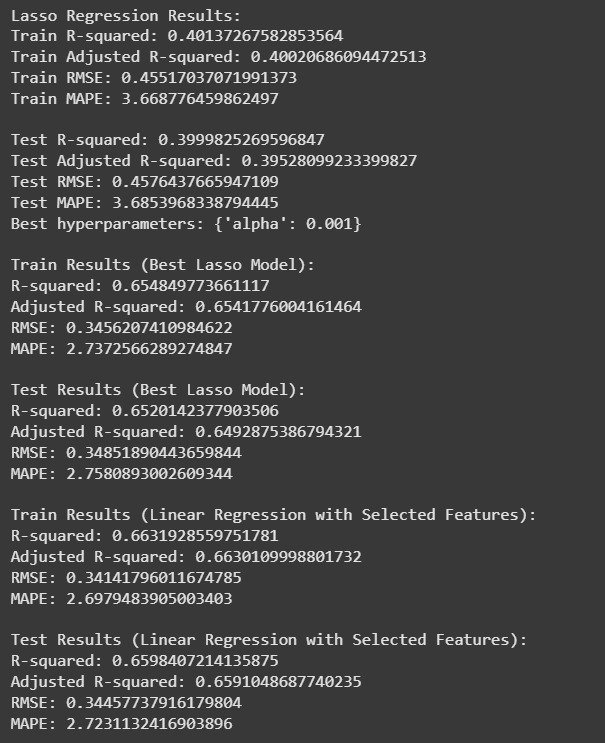
**Steps**

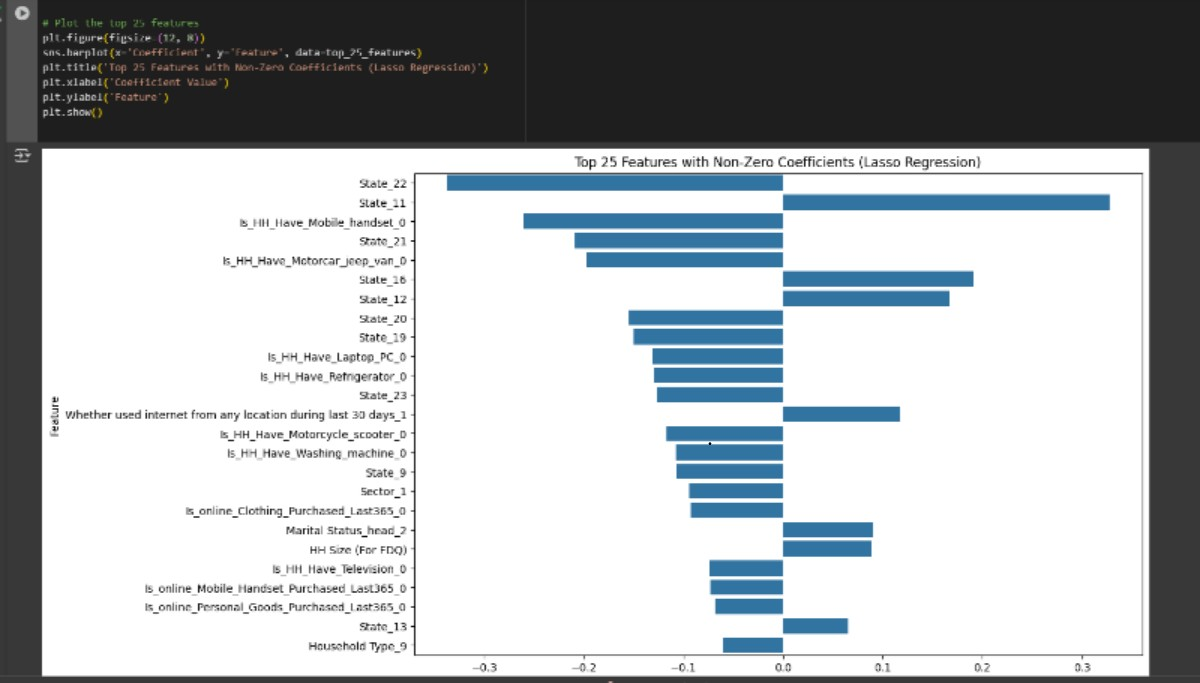
1. Convert the person level dataset to household level. We considered Marital status, Highest educational level attained (code), Total year of education completed for the head of the family. To reduce the bias, we also calculated a column of median Highest educational level attained (code) and median of Total year of education completed for the member of the household of the family. We then took Whether used internet from any location during last 30 days in a household & took the average of the other columns.
2. We then merged the Household level data with the Person level data with Household ID.
3. InHousehold data it had many categorical columns we used OneHotEncoder to convert it into numeric columns. For NIC and NCO, we use broad scale classification.
4. We then tried to reduce the data by dimension reduction technique MCA (Multiple Correspondence analysis) and tried to fit ML models but the value of R2 and RMSE were not very great.
5. We are considering our data on the basis of states and sectors.
6. Then the next step was to use Lasso regression for the feature selection and find variables impacting the Total Expenditure in positive and negative way.
7. Using the variables obtained we tried to fit ML models such as Multiple Linear Regression, Decision Tree, Random Forest, XGBoost, GradientBoosting Regressor, Artificial Neural Network Regression (ANN). (We also checked the assumptions of regression).

**Screenshot of execution**









**Team Menbers**

1. Shruti Khapekar –

Email- [khapekarshruti09@gmail.com](mailto:khapekarshruti09@gmail.com)

Contact no. – 7666800758

Institute – NMIMS, Mumbai

Year of passing – 2026

Contribution – Model building

2. Sandesh Tayde

Email- taydesandesh17@gmail.com

Contact no. – 8830981122

Institute – NMIMS, Mumbai

Year of passing – 2026

Contribution – Model building & feature scaling

3. Niraj Patil

Email- nirajpatil9892@gmail.com

Contact no. – 9136196225

Institute – NMIMS, Mumbai

Year of passing – 2026

Contribution – Feature Selection

4. Aditi Pawar

Email- aditipawar882003@gmail.com

Contact no. – 9969179289

Institute – NMIMS, Mumbai

Year of passing – 2026

Contribution – Data Preprocessing

5. Ronald Chettiar

Email- ronald.chettiar07@gmail.com

Contact no. – 9136123870

Institute – NMIMS, Mumbai

Year of passing – 2026

Contribution – Data Preprocessing & Documentation