Dhirubhai Ambani University
Technology

Formerly DA-IICT

# DS605 - Fundamentals of Machine Learning

## Team Lambda
## (Autumn 2025)



| Name | Student ID |
|---|---|
| Het Katrodiya(Leader) | 202518005 |
| Gaurang Jadav | 202518012 |
| Aksh Patel | 202518046 |
| Vrushil Panwala | 202518010 |

## Introduction

The project contains different places of residence, including income levels, deductions, taxable income, and final tax liability. It also includes information on the number of tax returns filed and migration-related attributes. This dataset provides a comprehensive view of how financial indicators vary across regions and how they influence overall tax liability. Using this data, we aim to analyze tax patterns, identify meaningful relationships, and build machine learning models to predict tax liability and classify tax changes due to migration.

## Problem Statement

It solves two statements :
1. **Regression** → Predict total tax liability using financial and demographic indicators.
2. **Clustering + Classification** → Identify migration patterns and classify whether tax liability increases or decreases after migration.

## Objective

→ Build a regression model to forecast total tax liability.
→ Use clustering to group migration behaviors.
→ Develop a classifier that predicts tax increase/decrease after migration.

## Dataset Description

The dataset contains tax-related information for different places of residence, including indicators such as Adjusted Gross Income (AGI), deductions, dependent exemptions, taxable income, and total tax liability. It also includes migration-related attributes and income class distributions, enabling analysis of how financial and demographic factors influence tax outcomes. The dataset provides a structured foundation for performing regression, clustering, and classification tasks in the project.

## Data Preprocessing

1. **Data Cleaning :-**
   - Dataset features have less than 300 data missing . so , we dropped it .
   - Features have many random values . we also remove those data .
   - We have “Disclosure” feature having more than 75% missing . We drop that feature .because it will not make any effect on prediction .
   - We have many features with some confusing names . we rename those columns .

2 . **Data Transformation :-**
   - We are having finance data . Instead directly removing outlier we transform data using log-lose then we remove the outliers .

3. **Target Variable Construction :-**
   - Tax_Change = tax liability of returns - prev tax liability
   - Based on tax_Change we created a tax_increase_flag .
   - It will use for the classification problem .
   - Tax_Increase_Flag → Tax Increase → 1
     Tax Decrease or same as prev → 0
4. **Exploratory Data Analysis :-**
   - We draw the correlation matrix,histogram,boxplot and elbow method to know the optimal cluster number .

5. **Train - test split :-**
   - We split the data into 80-20 . 80% data for training the data and 20% for the testing the data. We remove the target variables and highly correlated features from the training data to avoid the leakage data .
6. **Final Dataset for Modeling :-**
   - It contains the numerical features and categorical features for prediction .

## Model Training

- **Regression Problem :-**

- Ridge
- Lasso
- Random Forest
- Gradient Boost
- SVR

- **Classification problem :-**
  - Logistic
  - Decision Tree
  - Random Forest Classifier
  - XG Boost Classifier

# Model Evaluation

### Regression Problem :

Ridge

| | Train Performance | Test Perfomance |
|---|---|---|
| RMSE | 18929.965750626263 | 21825.51355887156 |
| R2 Score | 0.9254036022335829 | 0.9152283816540651 |

Lasso

| | Train Performance | Test Perfomance |
|---|---|---|
| RMSE | 18924.367621048706 | 21772.70215531246 |
| R2 Score | 0.925447716265068 | 0.9156381307054282 |

Random Forest

| | Train Performance | Test Perfomance |
|---|---|---|
| RMSE | 378.1550039041244 | 5685.493855364336 |
| R2 Score | 0.998822669174423 | 0.9942474828761344 |

Gradient Boosting

| | Train Performance | Test Perfomance |
|---|---|---|
| RMSE | 2520.672270957436 | 5926.115202650117 |
| R2 Score | 0.9986773316319396 | 0.9937502634467963 |

SVR

| | Train Performance | Test Perfomance |
|---|---|---|
| RMSE | 70146.06918894457 | 75986.24624826461 |
| R2 Score | -0.024294588221893676 | -0.027522906982932138 |

### Classification Problem :

| Model | Accuracy |
|---|---|
| Logistic Regression | 0.5678206583427923 |
| Decision Tree | 0.6912599318955732 |
| Random Forest Classifier | 0.7392167990919409 |
| XG Boost Classifier | 0.7812145289443814 |

## Application and Impact

This project helps in effective tax planning by predicting tax liability based on key financial indicators. It also provides insights into how migration influences changes in tax contributions across regions. These findings support better decision-making.

## Conclusion

The project successfully predicts tax liability using key financial indicators and accurately classifies whether tax liability increases or decreases after migration. The analysis reveals strong relationships between AGI, taxable income, and tax outcomes, while clustering provides meaningful insights into migration patterns. Overall, the models help improve understanding of tax behavior and support data-driven tax planning.