# Predict Poverty using machine learning

**Your Name here**

**Your Student ID**

**Semester**

Contents

[Predict Poverty using machine learning 1](#_Toc102088333)

[Introduction: 3](#_Toc102088334)

[Significance: 4](#_Toc102088335)

[Methodology: 4](#_Toc102088336)

[Analysis/Findings: 4](#_Toc102088337)

[Missing values: 4](#_Toc102088338)

[Outliers: 5](#_Toc102088339)

[Data visualization: 6](#_Toc102088340)

[Feature Selection: 7](#_Toc102088341)

[Data Splitting: 7](#_Toc102088342)

[Model Building: 8](#_Toc102088343)

[Decision Tree: 8](#_Toc102088344)

[Bayesian: 8](#_Toc102088345)

[Random Forest: 9](#_Toc102088346)

[XGBoost: 9](#_Toc102088347)

[Gradient Boosting: 10](#_Toc102088348)

[Comparison: 10](#_Toc102088349)

# Introduction:

The reason for this task is to introduce the usage of machine learning for the examination of enumeration information by proposing a component designing cycle to make family attributes related to the Synthetic Minority Over-Inspecting Technique (SMOTE) for foreseeing populace neediness. The machine learning models used in the poverty prediction from the census data, include Decision Tree, Bayesian, XGBoost, Random Forest, and Gradient Boosting. After training the above models, we got the following accuracies.

Accuracies of the used models

We can see that the XGBoost and Random Forest performed very well by giving us the highest accuracy.

# Significance:

The significance of this project is that it’ll be helpful for the government authorities to see the Poverty based on different factors and take immediate actions to help the population.

# Methodology:

The overall methodology of this project is given below:

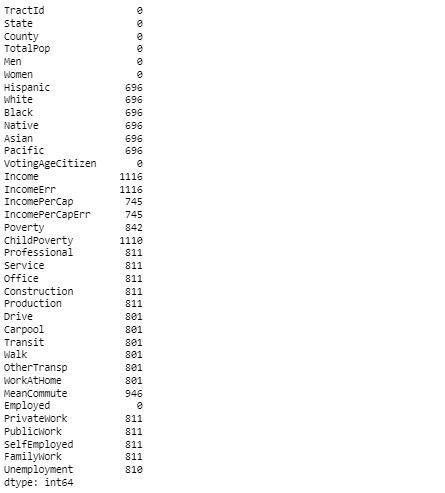
# Analysis/Findings:

We have the data of census of the whole USA having states and regions, the dataset has features like races, gender, income, poverty, etc. The goal is to predict poverty.

The first step was to perform the advanced exploratory data analysis on the data. In the basic EDA, we found out that we have a total of 37 features and 74001 rows.

## Missing values:

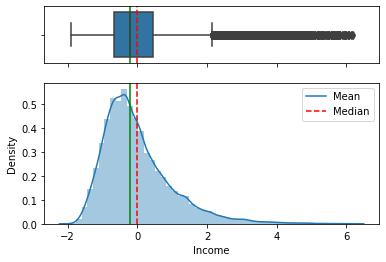
The next step in the analysis was to check if we have missing values or not? This step is taken into consideration in every real-life project. Raw data have many missing values and outliers, these values need to be treated very carefully because they’ll impact the machine learning side and analysis. Initially, we checked the missing values and we found out that in the dataset, we have a huge number of missing values.



We can see the missing values in every feature in the data in the above image. There are many ways to handle missing data, we sometimes perform statistical methods to fill all the NaN values, and sometimes we simply remove all the missing values. In our case, we removed all the missing values from all the features.

## Outliers:

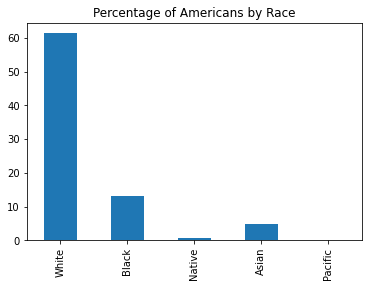
Outliers are the values that are out of the boundary from the rest of the values, either they are very big values or very small. Outliers are present in every data we deal with the outliers in many ways like we use the IQR (Inter Quartile Range) method and some other methods also to remove the outliers. We checked the outliers in our data and found out that we have a huge number of outliers also.



Just like in the above graph, we can see that we have outliers in the Income column. We are using the IQR method to detect the outliers and will remove them using the same method. The IQR method divide the data into four sections (25%, 50%, 75%, 100%). The values above 75% and below 25% are taken as outliers in this method. The IQR method will take all those values which are between 25% and 75%. This way we removed more than 5000 outliers.

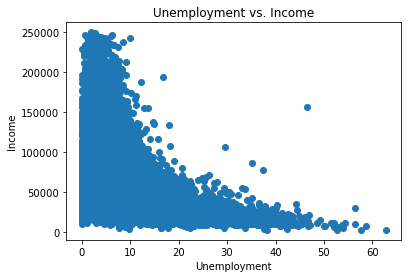
## Data visualization:

Data visualization is the pictorial way to see the data, this way we can do some deep-down analysis and get some stunning insights from the data. We tried to plot the Races of people present in our data by using the bar graph.

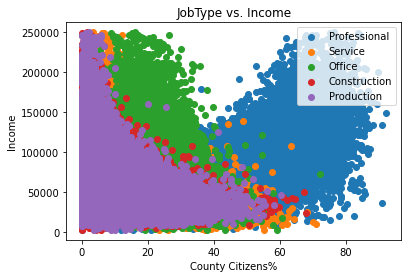


In the above graph, we can see that the number of white people is higher than the other.

Similarly, we plotted the unemployment vs Income scatter plot.



We can see the relationship between unemployment vs Income in the above graph.



In the above graph, we can easily see the type of job people have vs the Income they have.

## Feature Selection:

Feature selection is the process to select the best features for further machine learning model training. We are predicting the Poverty based on different features, so the Poverty feature is dependent and all the other features are independent.



We are storing all the independent features in the X variable and the dependent feature which is Poverty in the Y variable.

## Data Splitting:

The most important step in any machine learning project is data splitting. We split the data into test and train sets, these data sets will be used to train and test our models. We are splitting the data into training and testing sets.

* The training set will be used to train the models
* The testing set will be used to test the models



We are splitting the data in the ratio of 80:20, which means that 80% data will be used to train the models and the rest 20% of the data will be used to test the models and perform the predictions on the testing data.

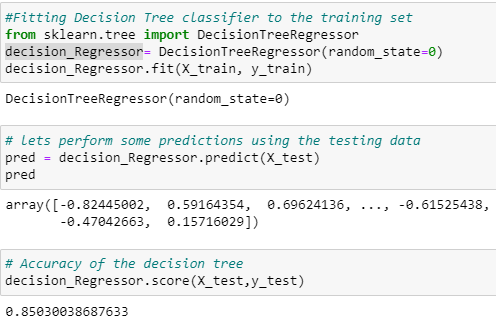
## Model Building:

This is the final step of our project, we are finally building our machine learning models, we will train them and test them. The goal is to train the following ML models.

* Decision Tree
* Bayesian (Bayesian Ridge)
* Random Forest
* XGBoost
* Gradient Boost

### Decision Tree:

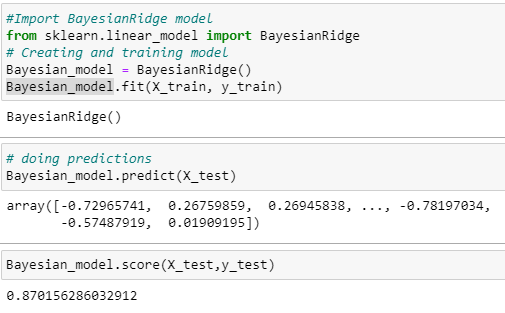
The first model that we trained is the Decision Tree regressor.



We are importing our model from sklearn library, after importing the model, the next step is to train it bypassing the training data into the fit() method, this method will feed the data into the model. After the model is trained, we performed predictions by feeding the testing data into the model by using the predict() method. After performing the predictions, we checked the accuracy of the model by using the score() method. The model gave us an accuracy of 85%.

### Bayesian:

The next model is the Bayesian Ridge regressor.



We are importing the model in the same way as we imported the decision tree model, we are feeding the data into the model and performing the predictions. Bayesian Ridge model gave us the accuracy of 87% which is higher than the Decision Tree.

### Random Forest:



We trained the Random Forest model by passing some of the parameters like n\_estimators = 100 and random\_state = 0, the Random Forest gave us the highest accuracy which is 92%.

### XGBoost:

The next model is the XGBoost model, we are training it the same way as we have trained the above models. This model also gave the same accuracy as Random Forest gave which is also 92%.



### Gradient Boosting:

The last model that we have trained in is the Gradient Boosting model, this model gave us an accuracy of 88%.



### Comparison:

We have trained 5 machine learning models and their accuracy comparison is given below.