# CAPSTONE PROJECT

## US CENSUS DATASET-2017

**Data Analytics for Business**

**Submitted by:**

**Gagandeep Singh 763543**

**Submitted to:**

**Mrs. Savita Sheravat**

**St. Clair college Mississauga**

## Contents

## Abstract

The dataset acs2017_census_tract_data is taken from 2017 American Community Survey 5-year estimates, illustrates the survey of the population counted in United states. The census data can be used to predict things such as state or income, clusters, or other datasets, and can be combined with other methods of analysing the data.

## Purpose

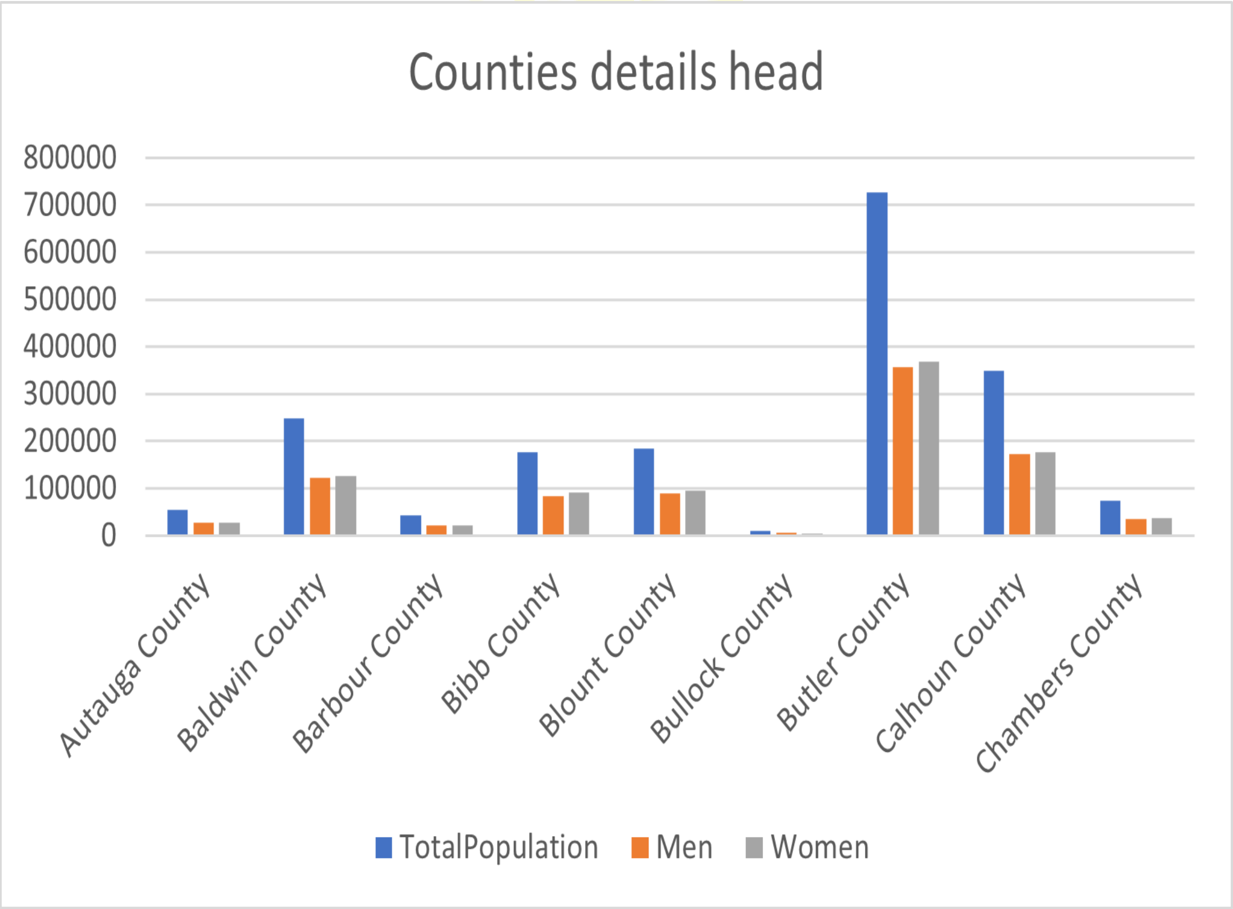
The government census determines the scope and nature of the national and local government areas; provides authoritative standards in public finance and public employment. The main motive of this dataset is representing overall census of the US including people's education, employment, poverty, income, construction and drive for transportation and other developments and may help local authorities to take economic decisions. The total number of observations in this dataset are 72743 and 32 variables.

## Table of attributes

| Attributes | Data Type | Description | Data values |
|---|---|---|---|
| **Trackid** | Numeric | Tracking id of each citizen's survey | 1001020100 to 72153750602 |
| **State** | character | State name in US | 52 states in the US |
| **County** | character | County name | Up to 2000 |
| **TotalPop** | Integer | Total population in county | 0 to 65528 |
| **Men** | Integer | Number of men | 0 to 32266 |
| **Women** | Integer | Number of women | 0 to 33262 |
| **VotingAgeCitizen** | Integer | Ratio of citizens able of voting | 0 to 39389 |
| **Income** | numeric | Total income | 2692 to 249750, NA's 1116 |
| **IncomePerCap** | Integer | Income per capita | 32 to 220253, NA's 745 |

| Attributes | Data Type | Description | Data values |
|---|---|---|---|
| **Poverty** | Numeric | Ratio of poverty in county | 0 to 100.00, NA’s 842 |
| **Childpoverty** | Numeric | Ratio of poverty in children | 0 to 100.00, NA’s 1110 |
| **Professional** | Numeric | Number of professionals | 0 to 100.00, NA’s 811 |
| **Employed** | Integer | Number of people employed | 0 to 28945 |
| **Unemployed** | Numeric | Number of citizens unemployed | 0 to100.000, NA’s  810 |

## Sample chart



## Introduction

The United States Census Bureau (USCB), officially the Census Bureau, is a major agency of the United States Federal Statistics System responsible for the production of data about the United States people and economy.

The Government Census is a three-phase program that collects state and local government data every five years as part of the government census for years ending in "2 " and "7". Between censuses, comparable data on employment and financial performance are generated from quarterly and annual internship sample surveys.

Between 2013 and 2017, the total population of the United States was 321.0 million, with 163.0 million (50.8 percent) females and 158.0 million (49.2 percent) males. For those reporting only one race, 73.0% were white; 12.7 percent are black or African-American; 0.8 percent are Native Americans and Alaska Natives; 5.4 percent are Asian; 0.2% were Native Hawaiians and other Pacific Islanders, and 4.8% were of another race. The 3.1 percent report having two or more races. An estimated 17.6 percent of people in the United States are Hispanic. And, almost 61.5% of people in the United States are non-Hispanic white. Hispanics can be of any race.

The population statistics are derived from decennial censuses, which count the entire population of the United States every ten years, as well as several other surveys. The median household income in the United States was $57,652. An estimated 6.7 percent of households earned less than $10,000 per year, while 6.3 percent earned more than $200,000 per year. From 2013 to 2017, 14.6 percent of the population lived in poverty. An estimated 20.3 percent of children under the age of 18 were poor, compared to 9.3 percent of people 65 and older. An estimated 13.7 percent of people aged 18 to 64 lived in poverty.

## Business questions

- Which county of state counted for higher population?

  This question helps us to find the most crowed areas which can be targeted by companies to sell the products. Government can provide more facilities like schools, hospitals and public transport on the basis of population.

- Compare the income of men and women by county?

  This question will help to determine the income of individuals that outcome will give an idea about the taxes and growth of individuals. We will get the state wise details about the economy which will contribute to the growth of county.

- In which county most people eligible for voting.

  This will help political parties to know the voting criteria and calculate the estimation of votes. Parties can provide services and agenda to the residents for better living style.

- Determine the highest ratio of poverty in the US?

  This data can determine the population that are under poverty line. This will help the government officials to provide necessary facilities to the poor generation such as health care, education facilities, food etc because they are future of the county.

## Theme

- Data mining and Knowledge Discovery
- Predictive Analytics

  The dataset we choose that is intensive so we will use two themes for analysing the data. First one is data mining and knowledge discovery that will help to handle this dataset and provide some patterns, and the second one will use for predictions.

## Technique and tools

We will use different techniques and tools in python to solve our problems like libraries (SK learn, pandas, NumPy). Along with it, r programming is being used for analysing the linear regression model, decision tree, correlation between the attributes, and random forest.

## Reading file and data preparation in Python

Use the built-in open() function to open the file. The open() function returns a file object with a read() method for reading the file's content.

## Data frame and looking for data type

In python, you can use type() to determine the data type of each DataFrame column. To create a DataFrame; and In the DataFrame, check the data type of each column.

## Removing and changing attribute name

There are a few cases where you want to dispose of unusual attribute names, for example, "IncomeErr", and “IncomePercapErr”. On the off chance that we experience information, like this, tidying up the names of the factors in our dataframes might be required and will definietly make work more meaningful. We have changed the “totalPop” attribute name as TotalPopulation.

### Checking missing values

In the dataset, there may be missing values, means the incomplete data attributes or the records. Therefore, we can not assume the correct modeling of dataset. So, we need to check the NA,s and handle them by replacing with a specific value.

### Recode missing values with mean

Missing values in the dataset emerge when a piece of record is absent in a segment of an information outline or contains a person esteem rather than numeric worth. Missing values

should be dropped or maintained by replacing it with a specific value. In this dataset, we have replaced the NA's with the mean of that particular attribute.

### Split data

if you split the dataset into a training set and a test set, the training dataset does not have enough data for the model to learn an effective mapping from input to output. Also, there is not enough data in the test set to effectively evaluate the performance of the model.

### Train set and test set

Train/Test is a technique for determining the accuracy of your model. It is called Train/Test because the data set is divided into two parts: a training set and a testing set. The important task in machine learning algorithms is splitting the dataset into training set and testing set. Here, most of the known data is separated into train set and rest of the test set is compared with the train set to check its similarity. As a result, one can avoid the incompatibility of the dataset and can better understand the features of the data.

### Prediction and Test set

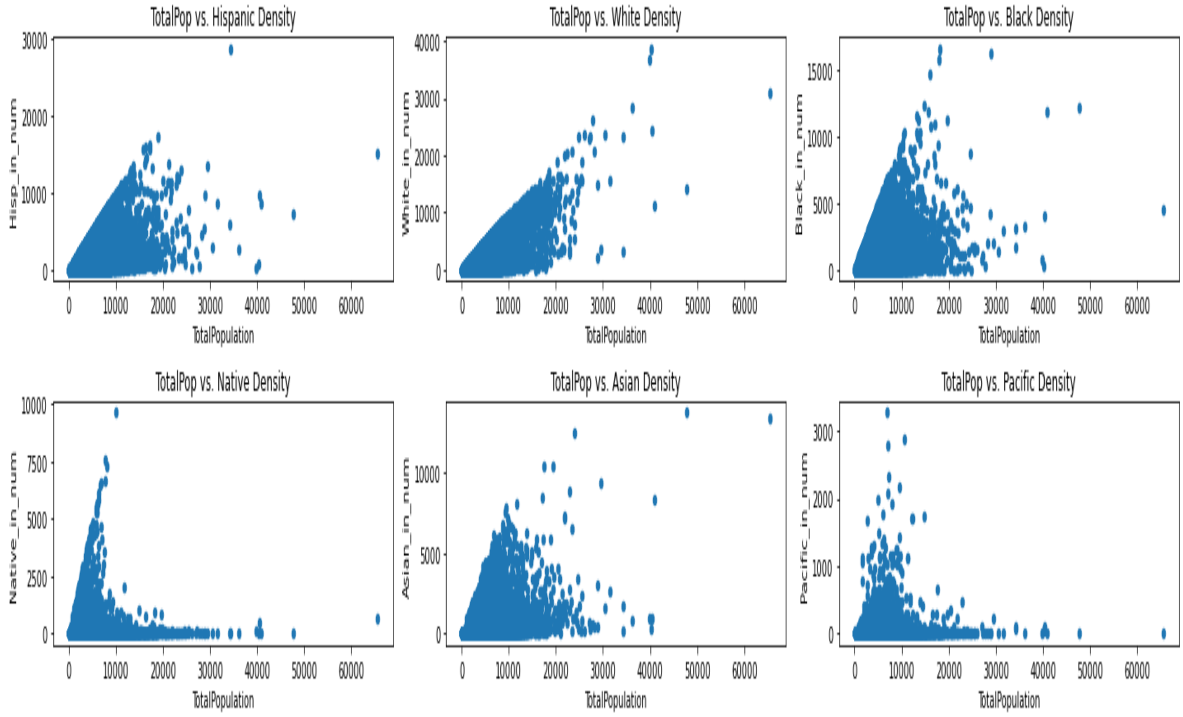
The training set should be a random sample of 80% of the original data. The remaining 20% of the testing set should be used.

## Accuracy

After training and testing the model of the dataset, it is very important to check the accuracy of the models. In this training and testing models, the accuracy to be calculated is 87.9.

## plots about Race

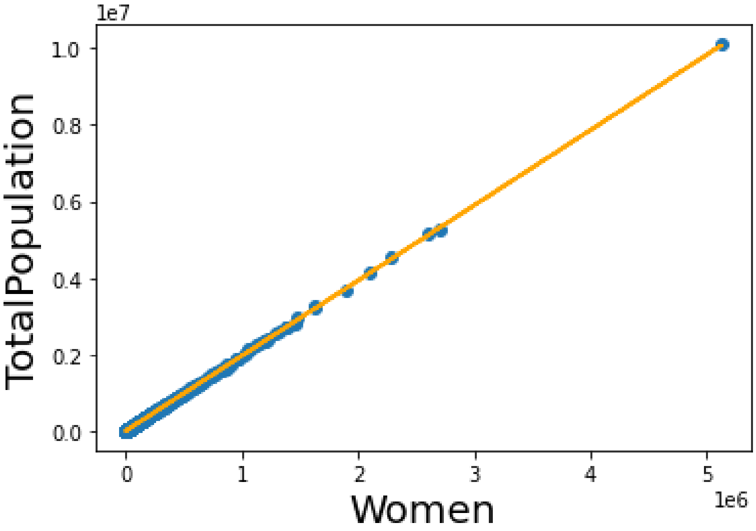
Approximately 40% of Americans perceive as racial or ethnic minorities. Nationally, the most important racial demographic corporations as of 2017 were White Non-Hispanic: 60.6%, Black Non-Hispanic: 12.3%, Asian Non-Hispanic: 5.5%, American Indian/Alaska Native Non-Hispanic: 0.7%, and, Hispanics of any race accounted for 18.1% of the U.S. population.



## Plotting regression line

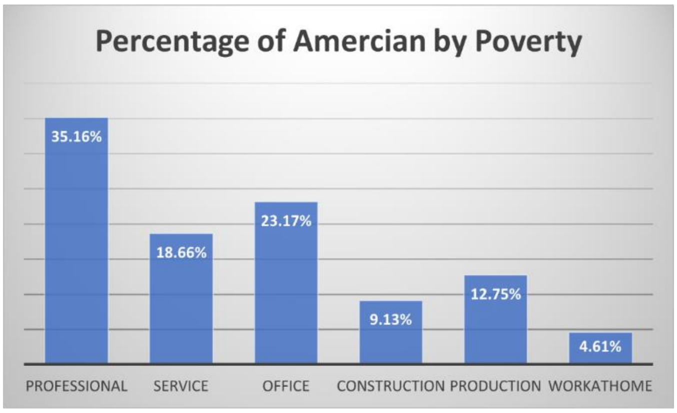
One of the best techniques to find out the linearity between the variables, is linear regression model. There are two types of variables- independent and dependent variables. In the predictive analysis, we use linear regression model, where the predicted variable helps to find out the possible outcome. For example, in our regression model, total women are related to the dependent variable (total population). According to the model, if the number of total men and women are almost equivalent ,while increased, total population will definitely increase.

The linear equation is: Y=a+bX, where, X is independent variable, Y is dependent variable and b is the slope of the line and the intercept is a.



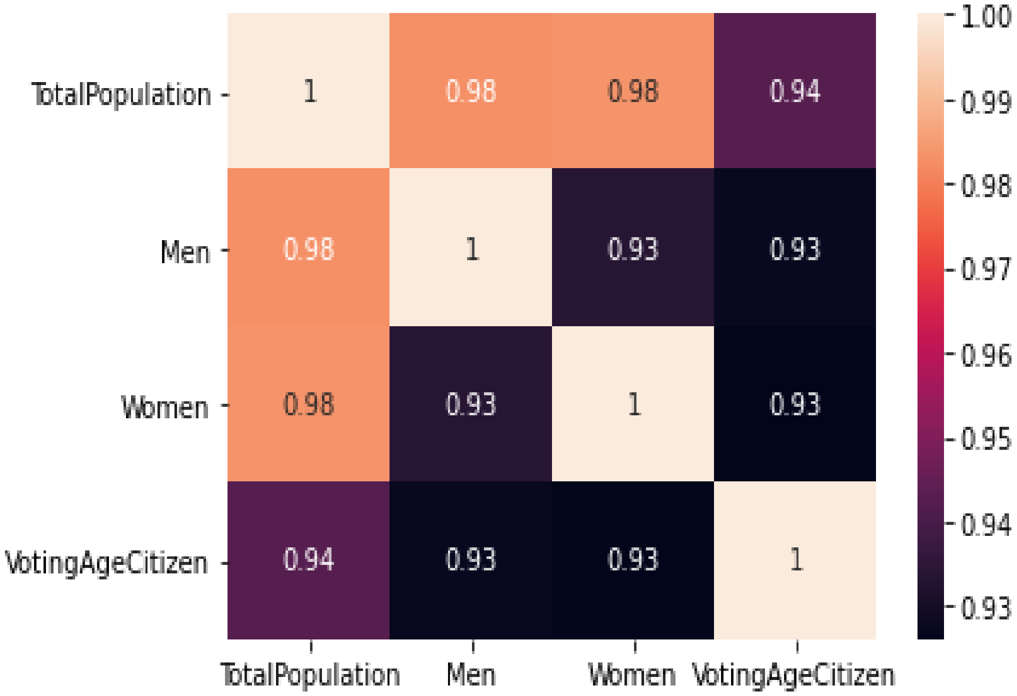
## Plot about poverty in US as percentage of total

From the bar chart of comparison between percentage of Americans’ poverty, maximum percentage of the poverty is under professional’s category  as 35%, followed by the office employees calculated for 23.17%.



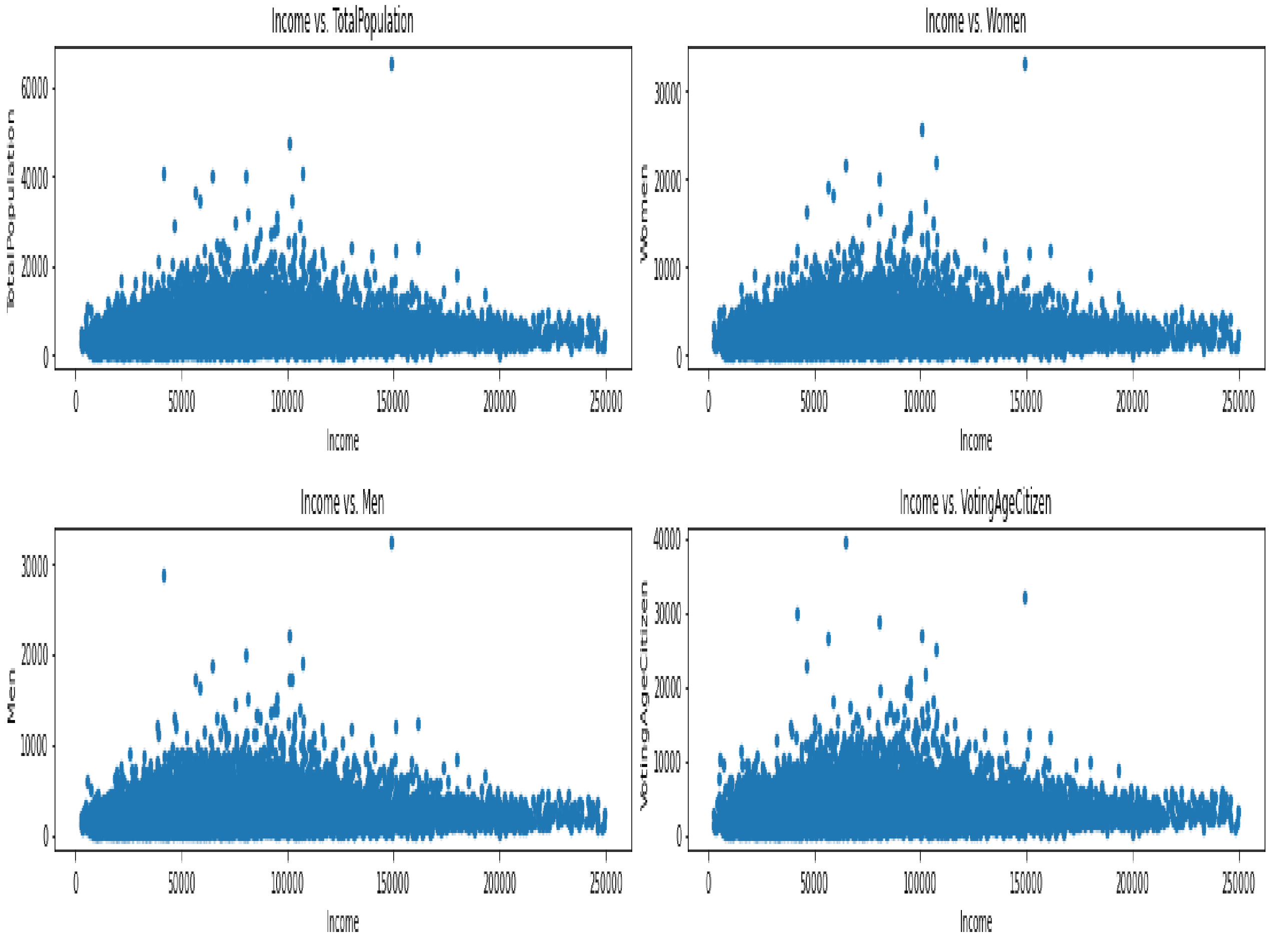
### Heatmap with correlation

Corelation shows the relationship between two or more variables. If two variables are highly corelated, we can make future prediction about one variable from the second one. It reveals how the change in one variable reflects in other one.



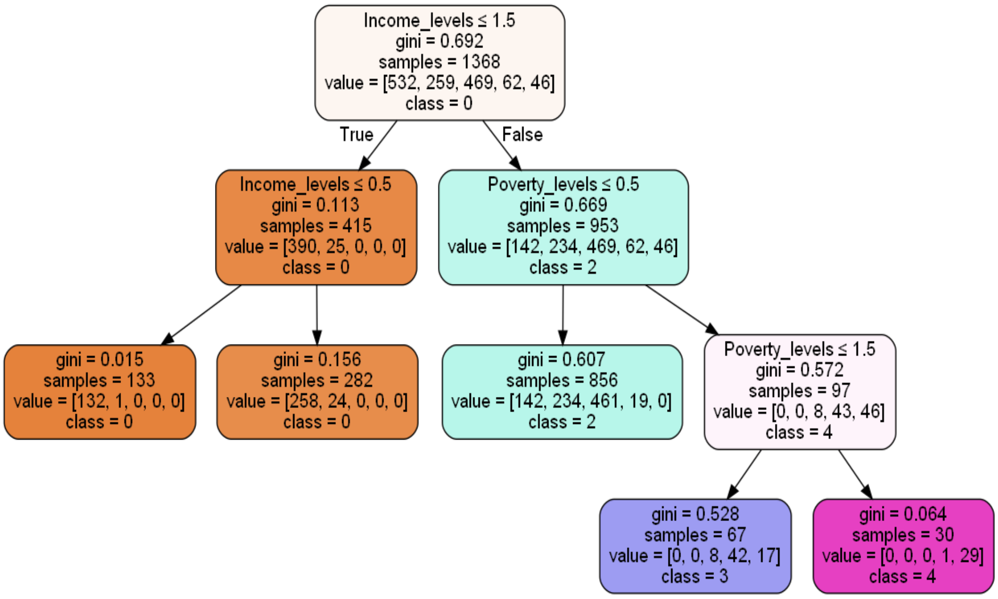
### Comparison between attributes

According to the plot evaluating overall income with overall population, women and men, the most earnings, overall populace, total women and men lies among 10,000 and 15,000. While comparing income with the total number of VotingAgeCitizen, the maximum count for income is considered 5,000 to 15,000.



## Decision tree

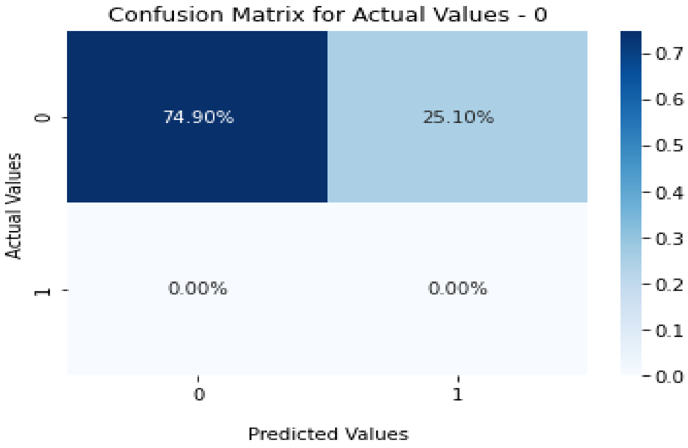
Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems. While designing decision tree, training model predict the class or value of the target variable by learning simple decision before data training.

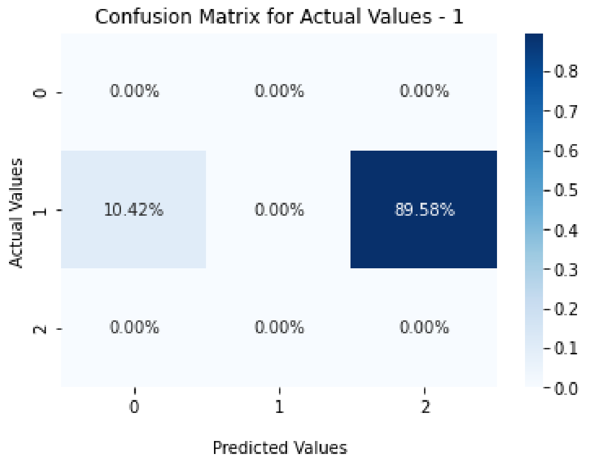


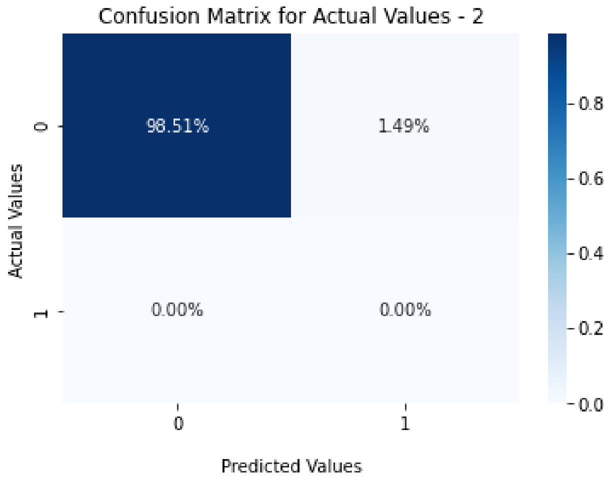
## Confusion matrix

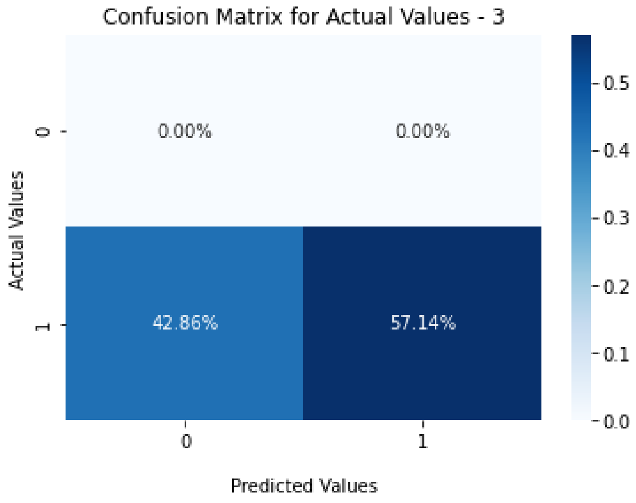
Below are the graphs for confusion matrix, there are five levels of total population that are

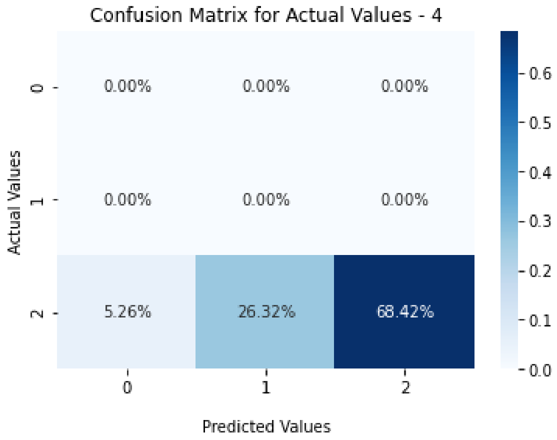
- 0-5000 is 0,
- 5000-25000 is 1
- 25000-50000 is 2
- 50000-500000 is 3
- 500000-1000000 is 4




Confusion Matrix for Actual Values - 1
Actual Values
Predicted Values
0
1
2
0.00%
0.00%
0.00%
10.42%
0.00%
89.58%
0.00%
0.00%
0.00%
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Confusion Matrix for Actual Values - 2
Actual Values
Predicted Values
0
1
98.51%
1.49%
0.00%
0.00%
0.8
0.6
0.4
0.2
0.0


Confusion Matrix for Actual Values - 3
Actual Values
Predicted Values
0
1
0.00%
0.00%
42.86%
57.14%
0.5
0.4
0.3
0.2
0.1
0.0

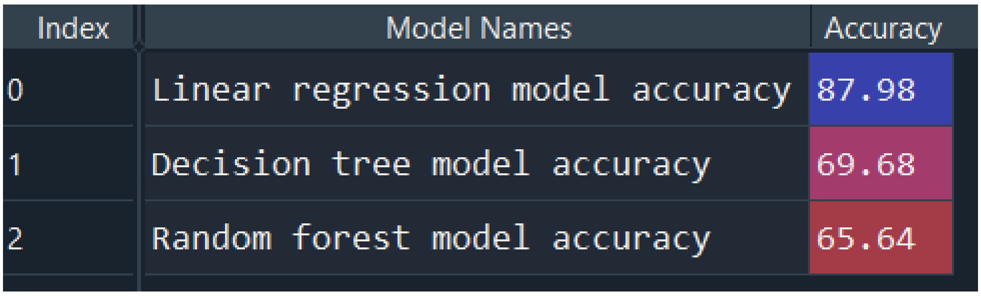


## Random Forest

In supervised machine learning, random forest is being used for solving classification and regression problems. It is beneficial for solving complicated problems by combining several classifiers. A random forest method is designed where, the decision is made depending on the high number of votes for classification and in case of regression, it consider the average number of votes.

## Conclusion

The census bureau of America helps native officials, community leaders, and businesses perceive the changes happening in their communities. it's the premier supply for elaborated population and housing info regarding our nation. To forecast and analyse the data from the 2017 US census dataset, we employed linear regression, k-nearest neighbour, decision tree, and random forest. We separated the data into a training and testing set and used these models to predict future values as precisely as possible.



| Index | Model Names | Accuracy |
|---|---|---|
| 0 | Linear regression model accuracy | 87.98 |
| 1 | Decision tree model accuracy | 69.68 |
| 2 | Random forest model accuracy | 65.64 |

### Codes are in GitHub

https://github.com/gagansingh033/Census_data_project

### Source of Reference

https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2015_census_tract_data.csv

https://github.com/gagansingh033/Census_data_project

https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

https://www.vedantu.com/maths/linear-regression

https://www.census.gov/acs/www/data/data-tables-and-tools/narrative-profiles/2017/

https://www.census.gov/newsroom/press-releases/2019/governments.html

U.S. Census Bureau,

2016 American Community Survey (ACS) 5-year estimates https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm

## Research paper reference

Income and Poverty in the United States: 2017

https://www.census.gov/library/publications/2018/demo/p60-263.html

The Supplemental Poverty Measure: 2017

https://www.census.gov/library/publications/2018/demo/p60-265.html

CES and Research Data Centers Research Report: 2017

https://www.census.gov/library/publications/2018/adrm/2017-ces-research-report.html

Market Absorption of Apartments Annual: 2017 Absorption

https://www.census.gov/library/publications/2018/demo/h130_17-a.html