**Data Analytics for Business**



**Capstone Project Final Report**

**Analysing and predicting the income and poverty level of states in US**

Jashanpreet Singh (0776982)

*Supervised by*

Mrs. Savita Seharawat

St. Clair College Ace Acumen , Mississauga

Winter 2022

**ACKNOWLEGMENT**

The completion of this project could not have been possible without the participation and assistance of a lot of individuals contributing to this project. However, I would like to express our deep appreciation and indebtedness to my supervisor and my classmates or their endless support, kindness, and understanding during the project duration.

First and foremost, I want to thank my supervisor, **Mrs. Savita Seharawat**, for her thoughtful guidance and unwavering support in accomplishment of my project on **Analysing and predicting the income and poverty level of states in US.** Her vast knowledge, extensive experience, and professional competence in Data Analytics have enabled me to effectively complete my project. This endeavour would not have been possible without her help and supervision. In my studies, I couldn't have asked for a better supervisor.

I also wish to express my sincere thanks to the St. Clair Ace Acumen  for accepting me into the Post-graduate program.

I hope I will achieve more in our future endeavours.

Thanks for all your encouragement!

Table of Contents

[Acknowlegment………………………………………………………………………………3](#_Toc452457337)

Abstract……….………………………………………………………………………………6

Introduction…..………………………………………………………………………………7

Review of the Literature…..…………………………………………………………………8

[**Theme……………………………………………………………………………………………10**](#_Toc452457339)

[**Keywords………………………………………………………………………………………..10**](#_Toc452457339)

**Tools**[**……………………………………………………………………………………………..10**](#_Toc452457339)

Detail of the US Census Demographic Data 2015…..……...…………………..…………11

[**Detailed Data Dictionary…………………………………………………………………....….11**](#_Toc452457339)

**Table 1:**[**Summary of Attributes………………………………………………………………,.11**](#_Toc452457339)

**Research Questions**[**……………………………………………………………………………..14**](#_Toc452457339)

[**Methodology………………………………………………………………………………….…15**](#_Toc452457339)

Data Preprocessing……………………………….…..…………………………..………....16

[**Import the libraries…………………………………………………………………………….16**](#_Toc452457339)

[**Load the file and data preparation……………………………………………………………16**](#_Toc452457339)

[**Dataset Shape …………………………………………………………………………………..16**](#_Toc452457339)

[**Descriptive Satistics…………………………………………………………………………….16**](#_Toc452457339)

[**Checking the missing values…………………………………………………………………...16**](#_Toc452457339)

[**Table 2:Missing Values ………………………………………………………………………..16**](#_Toc452457339)

[**Dependent Variable…………………………………………………………………………….17**](#_Toc452457339)

**Criteria For Cleaning the Missing values…….………………………….…………………………17**

**Total Population=0 in the dataset**[**……………………………………………………………...17**](#_Toc452457339)

**Table 3:Missing values in the entire rows**[**……………………………………………………..17**](#_Toc452457339)

**Filter out those values where total population**[**………………………………………………..17**](#_Toc452457339)

**Exploring the columns where the values are in total**[**…………………………………………17**](#_Toc452457339)

**Table 4: The columns are in total**[**……………………………………………………………...18**](#_Toc452457339)

**Convert the total value's columns in percentage**[**……………………………………………...18**](#_Toc452457339)

**Drop the unnecessary attributes**[**……………………………………………………………….18**](#_Toc452457339)

**Check the duplicate values**[**……………………………………………………………………..18**](#_Toc452457339)

**Table 5:Correlation between target variable and other attributes**[**………………………….18**](#_Toc452457339)

Exploratory Data Analysis [………………………………..…………………………….…20](#_Toc452457337)

**Figure 1: Correlation with Income**[**………………………………………………………,...…20**](#_Toc452457339)

**Figure 2: Selective features correlation with income ….**[**.........................................................21**](#_Toc452457339)

**Figure 3: Correlation matrix of income using Pearson method……………………………..[21](#_Toc452457339)**

**Figure 4: Average income of each method**[**……………………………………………………22**](#_Toc452457339)

**Figure 5: Histogram of State and Income**[**…………………………………………………….22**](#_Toc452457339)

**Figure 6: Sum of income of each state**[**………………………………………………………...23**](#_Toc452457339)

**Figure 7: Pie Chart of income** [**23**](#_Toc452457339)

**Figure 8 : Heat map for correlation of race population** [**24**](#_Toc452457339)

**Figure 9: Bar Graph of Average GDP of states**  [**24**](#_Toc452457339)

**Figure 10: Bar Graph of poverty level with white density** [**24**](#_Toc452457339)

**Figure 11: Bar Graph of poverty level with Black density**  [**24**](#_Toc452457339)

**Figure 12: Bar Graph poverty level with Asian density**  [**24**](#_Toc452457339)

**Figure 13: Bar Graph poverty level with Hispanic density**  [**24**](#_Toc452457339)

**Figure 14: Bar Graph poverty level with white density**  [**24**](#_Toc452457339)

**Figure 15: Total population Density in States of US** [**24**](#_Toc452457339)

**Figure 16: Bar Graph of race population** [**24**](#_Toc452457339)

**Figure 17: Race population Density in States of US** [**24**](#_Toc452457339)

**Experimental Design…………….……………………………………………….………..…………25**

**Normalize the dataset** [**25**](#_Toc452457339)

**Min-Max Scaler for normalization** [**25**](#_Toc452457339)

**Prediction and test set**  [**25**](#_Toc452457339)

**Train and test split approach** [**25**](#_Toc452457339)

**Model implementation and Evaluation….………………………………………….…………………………………………………………….…………26**

**Decision Tree Regression** [**26**](#_Toc452457339)

**Linear Regression** [**26**](#_Toc452457339)

**Support Vector Regression** [**27**](#_Toc452457339)

**Random Forest Regression Model** [**28**](#_Toc452457339)

**Table 7: Accuracy Table** [**28**](#_Toc452457339)

**Conclusion…..……………………………………………………………………………….29**

**Source of References...………….….……………………………………………………….30**

**Abstract**

The dataset acs2015 census tract data is based on 5-year estimates from the 2015 American Community Survey and depicts a survey of the people counted in the United States. The census data can be combined with other data analysis techniques to forecast things like state or income, poverty, unemployment, employment, total population in US states.

The United States Census Bureau (USCB), also known as the Bureau of the Census, is a key component of the United States' Federal Statistical System, and is in charge of gathering data about the American people and economy. The Census Bureau is part of the United States Department of Commerce, and the President of the United States appoints its director. The objective of the survey is showing the actual goal and scope of national government and local government in the different area. The US census bureau data always update the data related to every field for taking wise decisions in upcoming future.

The main aim of this dataset to estimates the income of men and women , total population of men and women , analyse the income growth by State , poverty level of race , estimate the employment and unemployment growth in every states.

# 

**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.

There are two types of surveys conducted by the Bureau: demographic and economic surveys. Demographic surveys include the Decennial Census of Population and Housing, the American Community Survey (ACS), the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the American Housing Survey (AHS).

Between 2015and 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.

**Review of the Literature**

The United States Census Bureau (USCB), also known as the Census Bureau, is a major agency of the United States Federal Statistics System that is responsible for producing data on the country's population and economy.

The American Community Survey (ACS), the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the American Housing Survey are all examples of demographic surveys (AHS).

The American Community Survey (ACS) gathers data on a wide range of economic, social, and demographic variables, as well as housing characteristics, from people. Every month, a random sample of housing unit addresses is collected for the survey, for a total of 3.5 million housing units contacted each year. The huge sample size ensures that the entire country is covered and that statistically reliable estimates for small and large geographic areas may be made. The American Community Survey (ACS) is a programme run by the United States Census Bureau that collects data on thousands of towns around the country on a variety of topics. Respondent engagement is required to guarantee that all communities are sufficiently represented. Criminal and civil consequences for unauthorised disclosure preserve the confidentiality of responses.

The previous section discussed how the ACS assists the federal government in making educated judgments about a variety of programmes. Many of these choices have a direct impact on local communities. Many people rely on federal funds that are allocated using ACS data. Based on ACS data, more than $360 billion in Federal monies were awarded directly to states in 2008, $72 billion to local regions, $48 billion to counties, and $10 billion to school districts, according to an analysis of ACS-related data used to distribute Federal funds.

The ACS is also an important source of data for state and local governments. It is typically the sole credible source of data for tiny and rural communities' demographic and economic characteristics. These communities must first understand the underlying demographic and economic aspects of their people in order to target scarce resources more effectively and better serve its residents. Thousands of towns would be compelled to collect data on their own or make choices without access to timely local data if the ACS did not exist. Furthermore, because ACS data uses uniform methodology and ideas, it allows for objective comparisons across communities. In community planning, state and municipal governments use ACS data to guide choices about transportation, housing, disaster management, and health care planning and allocation. They also use the ACS for economic growth, such as attracting new enterprises and encouraging expansion of existing ones.

State and local governments must plan comprehensively. Comprehensive plans are developed by cities, townships, and counties across the United States to guide future growth and development. These serve as a road plan for land use and include transportation, sewer and water systems, educational and recreational facilities, natural resources, and air and water quality control programmes, among other things. The all-encompassing scope of these plans typically offers voters with a wide set of community goals as well as the justification for their local leaders' decision-making processes. Local governments all around the country rely on the ACS for data to help them plan comprehensively. These local communities can use ACS data to learn more about their inhabitants' demographic and economic traits, such as educational attainment and languages spoken, resident migration, and work commuting patterns.

**Costs and burdens of community surveys in the United States**

To offer statistically reliable data on both a sufficient number of completed questionnaires is required for both small and big locations and groups, and this influences the survey's price. The ACS is expensive includes both the direct and indirect expenses to taxpayers, as seen in the ACS budget, as well as the time spent by individuals responding to the questionnaire in 2013, the ACS had a budget of $242 million per year. Based on average hourly earnings in the private sector of roughly $24 in 2013,6 the estimated value of respondent time spent on the ACS in 2013 was approximately $58 million.

Furthermore, the person MOHAMMAD ALKHALIL is already done working on the same dataset one year ago on Kaggle. His worked done on the income of people in US by using Python programming.

As I told that one person already worked on the same dataset, but this is a large dataset with missing values so there is so much work pending to do in same dataset. There are 37 different attributes in US census data 2015 , I will be work on income rate in each state and race population so on.

# 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.

**Keywords**

Python, Decision Tree, linear regression , random forest and support vector.SK learn, and machine learning.

**Tools**

All formulation and data visualization will be done by using Python.

**Detail of the US Census Demographic Data -2015**

US Census dataset contains the figures for 74001 separate records and 37 attributes for 52 states are included in the data collection. ‘Census\_Tract\_Id’, ‘State’, ‘County’, ‘Men’, ‘Women’, ‘TotalPop’, ‘Hispanic', 'White', 'Black', 'Native', 'Asian', 'Pacific', 'Citizen', 'Income', 'Poverty', 'Child Poverty', 'Professional', 'Service', 'Office', 'Construction', 'Production', 'Drive', 'Carpool', 'Transit', 'Walk', 'Other Transp', 'Work At Home', 'Mean\_Commute', 'Private\_Work', 'Public\_Work', 'Self\_Employed', 'Family\_Work', ’Employed’, 'Unemployment'. Interestingly, we have missing values in the entire rows so we cannot fill the missing values in the dataset. At the end, we decided to drop the rows which have missing values in the dataset.

**Detailed Data Dictionary**

## **Table 1: Summary of Attributes**

Total Attributes (Detailed data dictionary)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Serial No. | Attributes | Data Type | Description | Data Values |
|  | CensusTract | Character | Census id | Length – 74001, Na values- 0 |
|  | State | Character | States name | Length – 74002, NA values -0 |
|  | County | Character | County name | Length – 1928, NA Values- 0 |
|  | TotalPop | Numeric | Total Population | Length – 74001, NA Values – 0 |
|  | Men | Numeric | Total men | Length – 74001, NA Values – 0 |
|  | Women | Numeric | Total Women | Length – 74001, NA Values – 0 |
|  | Hispanic | Numeric | Person belong to Spanish’s Culture but living in US | Length \_ 73311, NA Values – 0690 |
|  | White | Numeric | Number of White people | Length – 73311, NA values – 0 |
|  | Black | Numeric | Number of Black people | Length \_ 73311, NA values – 0 |
|  | Native | Numeric | Number of Citizen | Length \_ 73311, NA values – 0 |
|  | Asian | Numeric | Number of Asian People | Length – 73311, NA values – 0 |
|  | Pacific | Numeric | Number of people belong to pacific region | Length \_ 73311, NA values – 0 |
|  | Citizen | Numeric | Number of citizen | Length \_ 74001, NA values- 0 |
|  | Income | Numeric | Median household income | Length \_ 72901, NA values – 1100 |
|  | Income Err | Numeric | Median Household income error | Length \_ 72901, NA values – 1100 |
|  | IncomePerCap | Numeric | Income per capita | Length – 73261, NA values – 740 |
|  | IncomePerCapErr | Numeric | Income per capita error | Length – 73261, NA values – 740 |
|  | Poverty | Numeric | Percent under poverty level | Length\_73166, NA values – 835 |
|  | ChildPoverty | Numeric | Percent of Child under poverty level | Length\_72883, NA values – 1118 |
|  | Professional | Numeric | Percent employed in business, management, science and art. | Length – 73194, NA values- 807 |
|  | Service | Numeric | Percent employed in service job | Length – 73194, NA values- 807 |
|  | Office | Numeric | Percent employed in sales and office jobs | Length – 73194, NA values- 807 |
|  | Construction | Numeric | Percent employed in construction work | Length – 73194, NA values- 807 |
|  | Production | Numeric | Percent employed in production, transportation and material movement | Length – 73194, NA values- 807 |
|  | Drive | Numeric | Percent commuting car alone | Length – 73204, NA values – 797 |
|  | Carpool | Numeric | Percent carpooling in car, van or truck | Length – 73204, NA values – 797 |
|  | Transit | Numeric | Percent commuting public on transport | Length – 73204, NA values – 797 |
|  | Walk | Numeric | Percent walking to work | Length – 73204, NA values – 797 |
|  | Other Transport | Numeric | Percent transport via other means | Length – 73204, NA values – 797 |
|  | WorkAtHome | Numeric | Percent working at home | Length – 73204, NA values – 797 |
|  | MeanCommute | Numeric | Mean commute time | Length \_ 73052, NA values – 949 |
|  | Employed | Numeric | Percent employed | Length \_ 74001, NA values – 0 |
|  | PrivateWork | Numeric | Percent employed at private company | Length – 73194, NA values- 807 |
|  | PublicWork | Numeric | Percent employed in public work | Length – 73194, NA values- 807 |
|  | SelfEmployed | Numeric | Percent self- employed | Length – 73194, NA values- 807 |
|  | FamilyWork | Numeric | Percent in unpaid family | Length – 73194, NA values- 807 |
|  | Unemployment | Numeric | Unemployment rate | Length – 73199, NA values – 802 |

**Research Questions**

1. **Which state have highest income ?**

We will analyse the data that based on the gender to describe the highest income as well as we will predict the poverty level of children in US states. So, government of US can take essential steps to tackle the low-income rate and poverty rate.

1. **Which occupation have highest employment rate?**

Here, we will going to show the highest employment rate in different occupations according to different category such as professional, self-service, public work , private so on.

1. **Which population have highest density in US?**

We will find the race density in the US country which race population category most live in the US country.

1. **Estimate the highest poverty rate in race population?**

This information can be used to assess the number of people living under the poverty line. This will assist government authorities in providing critical services to people, such as health care, education, and food, as they are the county's future.

# Methodology:

The workflow to build a models can be summarised as below:

Data Pre-processing

# Part 1: Data Preprocessing:

Cleaning the data with certain pre-processing procedures is required before processing the US Census Dataset.

1. **Import the libraries**

We will import some libraries to support our python code, The import keyword in Python is used to make code from one module available in another. Imports are essential for successfully organising your Python code. Imports can help you be more productive by allowing you to reuse code while yet keeping your projects manageable.

1. **Load the file and data preparation**

Use the built in pd.read\_csv() function to load the file in pythonand we can understand the shape of data by using the shape()function. There are 74001 observations and 37 attributes.

1. **Dataset shape**

The US census dataset 2015 have 74001 observations and 37 attributes.

1. **Descriptive statistics**

The describe function in Python - pandas can be used to get descriptive or summary statistics (). Describe The mean, standard deviation, and interquartile range (IQR) values are returned by this function.

**Table 1: Descriptive Statistics**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No | CensusTract | TotalPop | Men | Women | Hispanic | White | Black | Native |
| count | 73311 | 73311 | 73311 | 73311 | 73311 | 73311 | 73311 | 73311 |
| mean | 2.84E+10 | 4366 | 2148 | 2219 | 17 | 62 | 13 | 1 |
| std | 1.64E+10 | 2097 | 1057 | 1080 | 23 | 31 | 22 | 4 |
| min | 1E+09 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 1.3E+10 | 2926 | 1426 | 1479 | 2 | 39 | 1 | 0 |
| 50% | 2.8E+10 | 4085 | 1997 | 2078 | 7 | 71 | 4 | 0 |
| 75% | 4.2E+10 | 5458 | 2682 | 2783 | 20 | 88 | 14 | 0 |
| max | 7.22E+10 | 53812 | 27962 | 27250 | 100 | 100 | 100 | 100 |

Table 1: Descriptive Statistics

1. **Check the missing values**

To get how many missing values are in each column we use **sum()** along with **isnull()** function. There are 23569 missing values in the dataset.

**Missing Values:** The dataset contains a number of missing values in the numeric columns Hispanic', 'White', 'Black', 'Native', 'Asian', 'Pacific', 'Citizen', 'Income', 'Poverty', 'Child Poverty', 'Professional', 'Service', 'Office', 'Construction', 'Production', 'Drive', 'Carpool', 'Transit', 'Walk', 'Other Transp', 'Work At Home', 'Mean Commute', 'PrivateWork', 'PublicWork', 'SelfEmployed', 'FamilyWork','Unemployment'. Interestingly, we have missing values in the entire rows so we cannot fill the missing values in the dataset. At the end, we decided to drop the rows which have missing values in the dataset.

**Table 2: Missing Data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Columns | CensusTractId | State | County | TotalPop | Men | Women | Citizen |
| 43 | 1E+09 | Alabama | Baldwin | 0 | 0 | 0 | 0 |
| 107 | 1.02E+09 | Alabama | Calhoun | 0 | 0 | 0 | 0 |
| 108 | 1.02E+09 | Alabama | Calhoun | 0 | 0 | 0 | 0 |
| 868 | 1.1E+09 | Alabama | Mobile | 0 | 0 | 0 | 0 |
| 1063 | 1.12E+09 | Alabama | Shelby | 0 | 0 | 0 | 0 |
| 1460 | 4.01E+09 | Arizona | La Paz | 0 | 0 | 0 | 0 |

Table 2: Missing Data

we can analyse from the above table that we have missing values in the entire rows so we can delete those rows which have missing values.

1. **Dependent Variable**

The target variable is “income” representing the total income of different occupations.

## **Criteria for cleaning the missing values**

**Total Population is = 0 in the dataset**

**Table 3: Missing Values in entire rows**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Columns | CensusTractId | State | County | TotalPop | Men | Women | Citizen |
| 43 | 1E+09 | Alabama | Baldwin | 0 | 0 | 0 | 0 |
| 107 | 1.02E+09 | Alabama | Calhoun | 0 | 0 | 0 | 0 |
| 108 | 1.02E+09 | Alabama | Calhoun | 0 | 0 | 0 | 0 |
| 868 | 1.1E+09 | Alabama | Mobile | 0 | 0 | 0 | 0 |
| 1063 | 1.12E+09 | Alabama | Shelby | 0 | 0 | 0 | 0 |
| 1460 | 4.01E+09 | Arizona | La Paz | 0 | 0 | 0 | 0 |

Table 3: Missing Values in entire rows

1. **Filter out those values where Total population > 0**

We selected only those values from the dataset where is total population > 0 because as I already told that we do not have missing values in the form of random pattern. We have entire rows missing in the dataset. We removed the entire rows from our dataset.

1. **Exploring the columns where the values are in total**

Here, I explored only those attributes which have values in total like **Total Pop, Men, Women, Citizen and Employed**. So we have to convert these values in the percentage of total population.

**Table 4 : The columns are in total**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Serial No. | TotalPop | Men | Women | Citizen | Employed |
| 0 | 1948 | 940 | 1008 | 1503 | 943 |
| 1 | 2156 | 1059 | 1097 | 1662 | 753 |
| 2 | 2968 | 1364 | 1604 | 2335 | 1373 |
| 3 | 4423 | 2172 | 2251 | 3306 | 1782 |
| 4 | 10763 | 4922 | 5841 | 7666 | 5037 |

Table 4 : The columns are in total

**Covert the total value’s columns in Percentage**

We converted the above values of attributes(TotalPop, Men,Women , Citizen and Employed) into percentage for better modelling which were in total because other attribute’s values are already in the percentage.

1. **Drop the unnecessary attributes**

I have drop those attributes which are not useful in my dataset such as IncomeErr, IncomePerCap, IncomePerCapErr because we have already ‘INCOME’ attributes as our target variable. So, we do not need any unnecessary columns which does not provide any useful information.

Furthermore, drop the ‘TotalPop’, and ‘Employed’ attributes as well from our dataset because total population is always sum of MEN+WOMEN population and in the case of ‘Employed’ attribute, we have detail of different occupations in the different attributes such as 'Professional', 'Service', 'Office', 'Construction', 'Production', 'Drive', 'Carpool', 'Transit', 'Walk', 'Other Transp', 'Work At Home', 'Mean Commute', 'PrivateWork','PublicWork', 'SelfEmployed', 'FamilyWork'. So we dropped the Employed attributes from our dataset.

1. **Check the duplicate values**

Luckily, we don’t have any duplicate values in our dataset.

1. **Table 5: Correlation between target variable and the other Attributes**

|  |  |
| --- | --- |
| Features | Income |
| Income | 1 |
| Professional | 0.733445 |
| Poverty | 0.703583 |
| ChildPoverty | 0.663878 |
| Service | 0.588114 |
| Production | 0.499943 |
| Unemployment | 0.485206 |
| WorkAtHome | 0.378687 |
| Construction | 0.332996 |
| Employed | 0.317922 |
| White | 0.314785 |
| Black | 0.310531 |
| Carpool | 0.287733 |
| Asian | 0.282894 |
| MeanCommute | 0.229634 |
| Hispanic | 0.228052 |
| Citizen | 0.20462 |
| Men | 0.176304 |
| TotalPop | 0.1742 |
| Women | 0.166493 |
| Walk | 0.147192 |
| OtherTransp | 0.105428 |
| SelfEmployed | 0.089954 |
| Drive | 0.084967 |
| Native | 0.072084 |
| CensusTract | 0.068449 |
| Office | 0.067585 |
| PrivateWork | 0.036137 |
| Transit | 0.009625 |
| Pacific | 0.007838 |
| PublicWork | 0.007253 |
| FamilyWork | 0.004328 |

Table 5: Correlation between target variable and the other Attributes

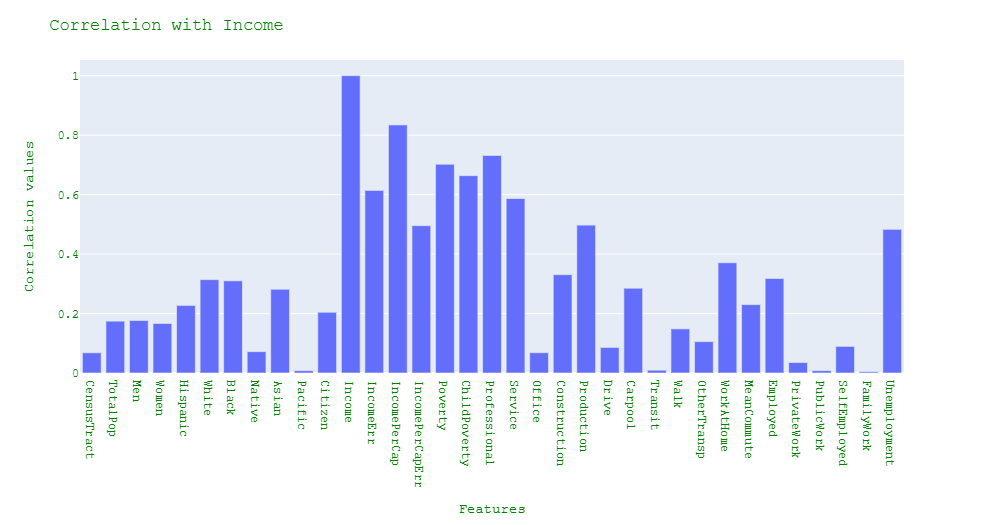
# Its shows that professional, poverty, child poverty, service, production and Unemployment gives the satisfying correlation with INCOME.

# Part 2:Exploratory Data Analysis:

## **Graphical Representation of ‘Income’ and ‘State’**

All features are visualised using Histograms and Box plot, Bar graph and pie chart to show the output in different manner.

**Figure 1: Correlation with Income (Bar Graph)**

****

**Figure 1: Correlation with Income**

From the above bar graph , we analyse the correlation of features with our target variable. Here, we can see that some of our features are making the good correlation with target variable ‘Income’ where as some features are showing bad correlation with target.

**Features selection**

We selected only those features which were making the nice correlation with target variable so we will again see the correlation of selective features through visualization process.

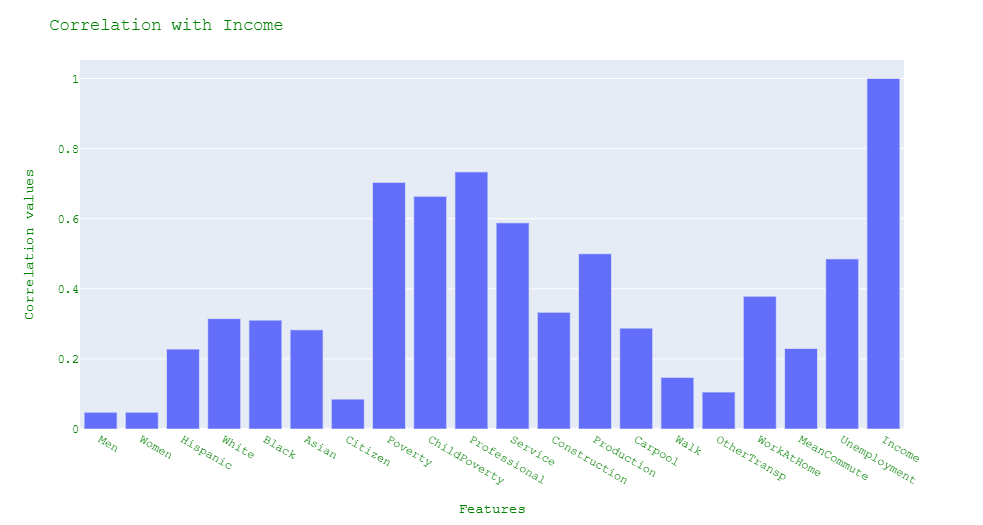
**Figure 2 : Selective Features correlation with income**

Figure 2 : Selective Features correlation with income

from the above chart, we selected only Men, Women, Hispanic, White, Black, Asian, Citizen, Poverty, Child poverty, Professional, Service, Construction, Production, Carpool, Walk, Other Transp, Work at Home, Mean Commute, Unemployment and Income which made the good correlation with target variable.

**Figure 3: correlation matrix of income using Pearson method**

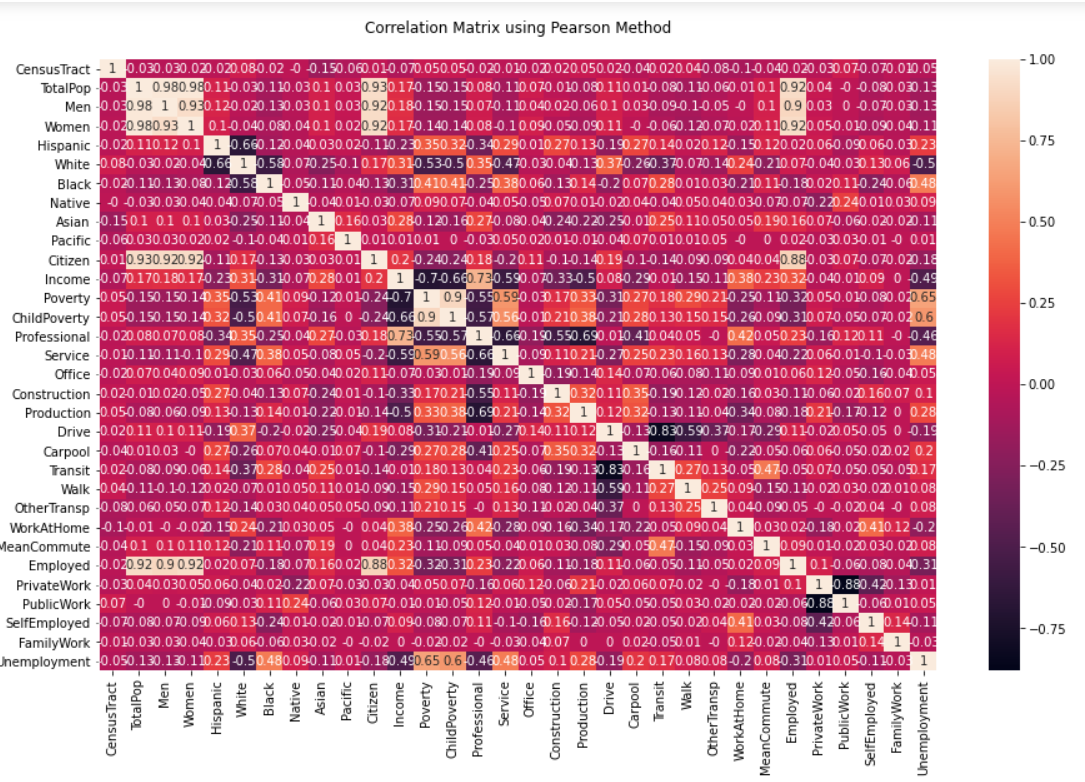
****

Figure 3: correlation matrix of income using Pearson method

In this Pearson matrix, we can see the correlation matrix of income with features. The values shown between -0.75 and 1.00 on the side right corner.

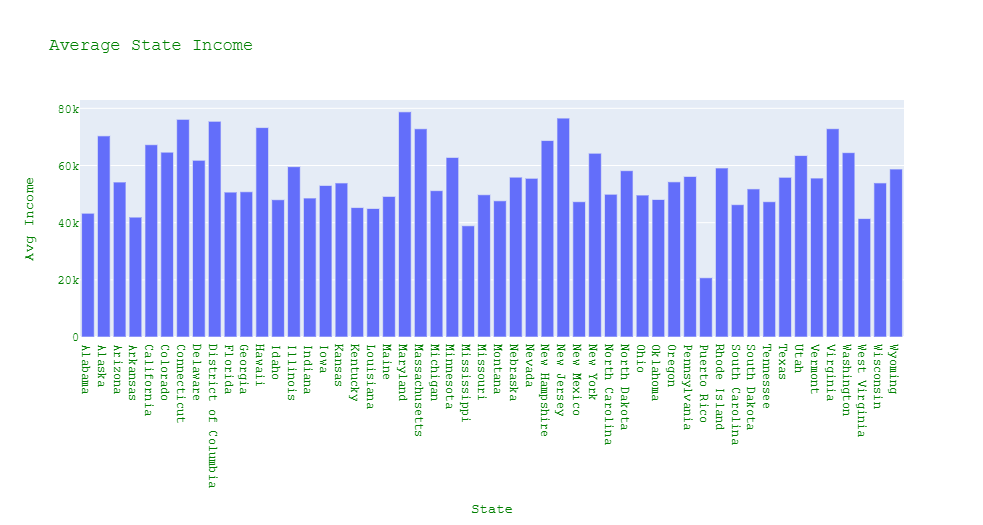
**Figure 4: Average income of each state**

Figure 4: Average income of each state

In the above bar graph, this bar graph shows the average income of each state. It is clearly seen from the graph that most of state high average income except Puerto Rico.

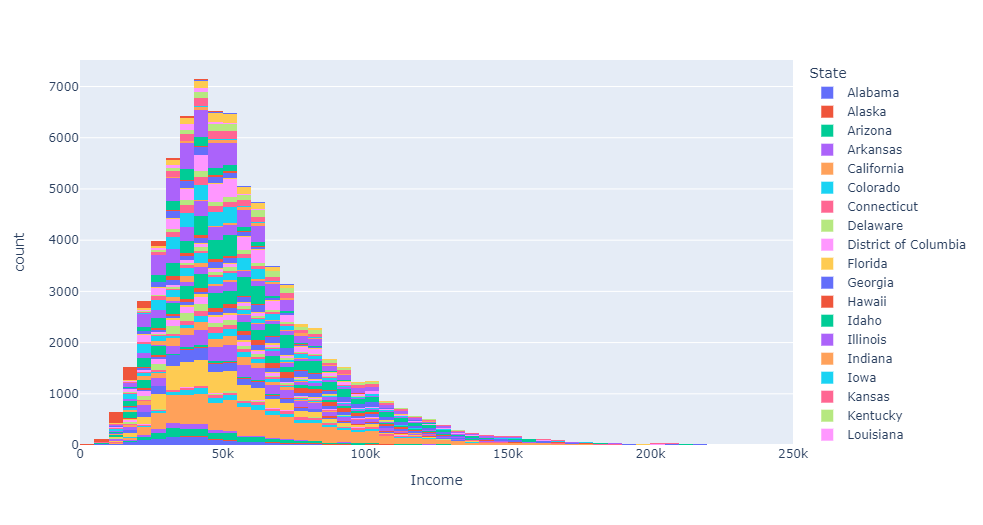
**Figure 5: Histogram of Sate and income (count)**

Figure 5: Histogram of Sate and income (count)

In the histogram, we count the income within each state , how many people are earning on which level income. The income level is from 50k to 250k in this dataset.

**Figure 6: Sum of Income with each State**

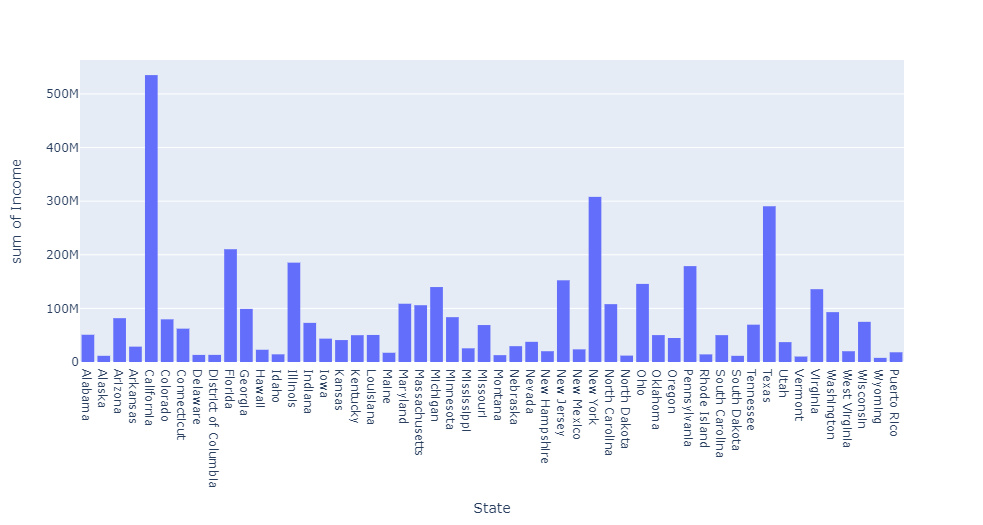
****

Figure 6: Sum of Income with each State

In this histogram, we found that top 5 states have highest income such as **California(535.1586M), Florida(210.3136M), Illinois(185.3157M), New York(308.017M), and Texas(490.4916M).**

**Figure 7 : Pie chart of income**

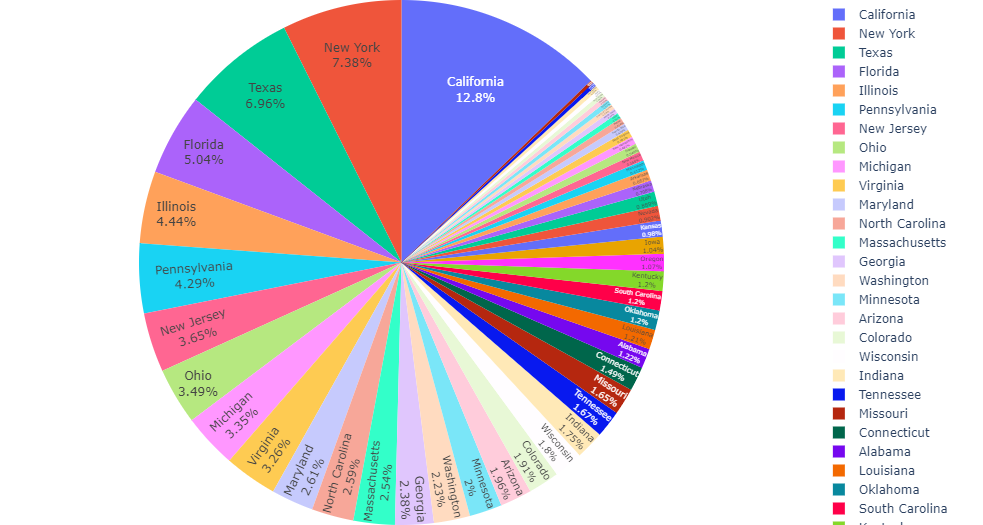
****

Figure 7 : Pie chart of income

This pie chart also provides the same information like as above graph representation California has highest income (12.8%), follow up the New York(7.38%) and Texas(6.96%) and Florida(5.04%) and Illinois(4.44%) respectively.

**Figure 8: Heat map for correlation of race population**

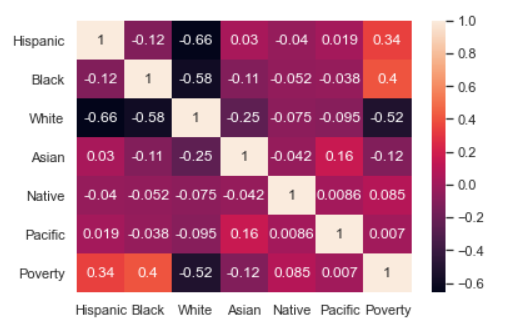
****

Figure 8: Heat map for correlation of race population

We can see the correlation of race population with poverty. There is building the satisfying negative correlation with race features.

**Figure 9: Bar graph of race population**

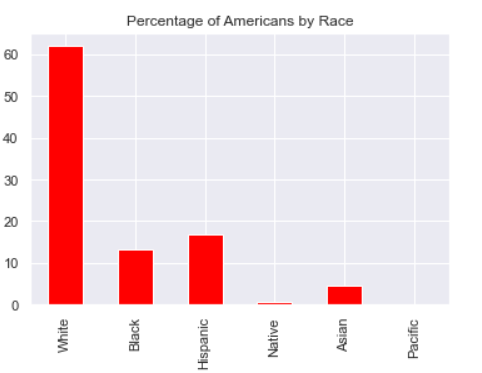
****

Figure 9: Bar graph of race population

We can observed from above graph that white people have highest density in US as compared to other race features.

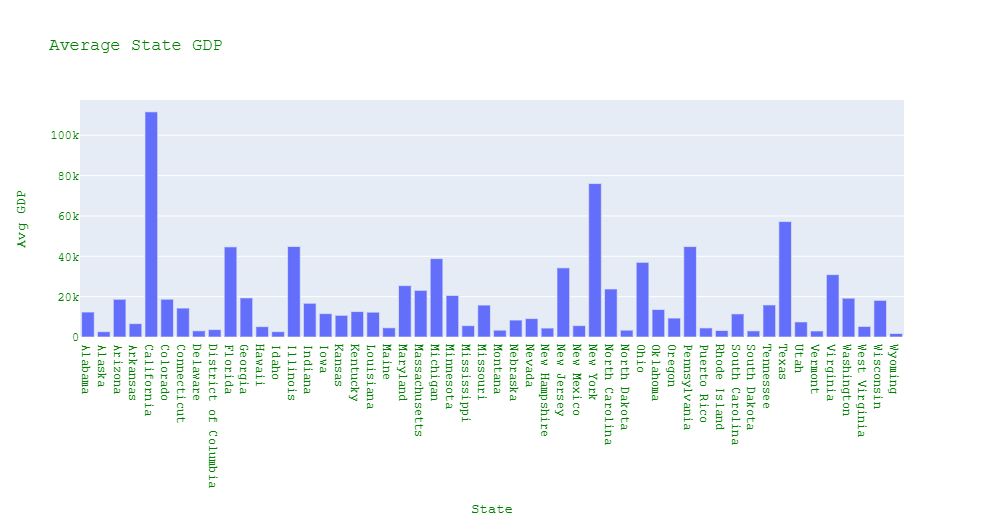
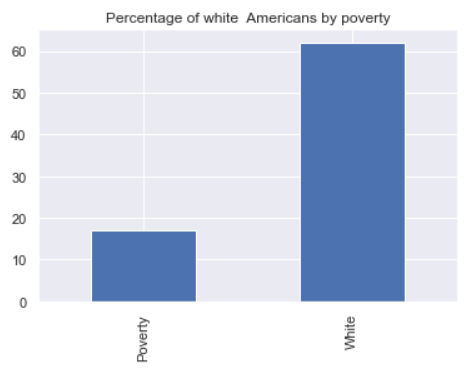
**Figure 10: Bar graph of Average GDP of States**

Figure 10: Bar graph of Average GDP of States

From the above bar graph, this is clearly show the Average GDP Growth of states in US. Furthermore, we can say that California on the top among from other states, New York is occupied the second position in GDP category. At the end, there are also many states which contributing a lot in the GDP growth of US, but there are many states also which have less GDP growth.

**Figure 11: Poverty level with white people**

# 

Figure 11: Poverty level with white people

# Above bar graph shows that density of white people has less poverty level in the US. There are more than 60% white people accounted around 17-18% poverty level.

# Figure 13: Poverty level with Black people

# 

# Figure 13: Poverty level with Black people

# So we can see that there are poverty level high as compared to density level of Black population.

# Figure 14: Poverty level with Asian people

# 

# Figure 14: Poverty level with Asian people

# Above bar graph that Asian people have low density in US but poverty level is very much high in the US states. We can say that all Asian people are living under the poverty line.

# Figure 15: Poverty level with Hispanic people

# 

# Figure 15: Poverty level with Hispanic people

# Sadly, we can say that Hispanic people in the US are facing the unemployment problems and live under the poverty level. At the end, we observed from the above graphs that only white people density live, enjoy the life and earn with good income.

# Figure 16: Total Population Density in States of US

# 

# Figure 16: Total Population Density in States of US

# Here, bar graph shows the population level of men and women according to states in the US.

# Figure 17: Race Population Density in States of US

# 

# Figure 17: Race Population Density in States of US

# As we already seen from the previous information that white people are more and as well, they are living above from the poverty level in the US states as compared to other population.

# Experimental Design

## **Normalize the data sets:**

Normalization is the process of changing data into a normal form, which entails using a mathematical approach to achieve the balance in the data. It is the process of converting several types of data into a comparable scale.

**Min-Max Scaler for Normalization**

One of the most prevalent methods of data normalisation is min-max normalisation. The minimum value of each feature is converted to a 0, the highest value is converted to a 1, and all other values are converted to a decimal between 0 and 1.

# 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.

**Train and test split approach**

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.

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**

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.

Output

**The shape of X\_train = (58181, 19)**

**The shape of y\_train= (58181,)**

**The shape of X\_test= (14546, 19)**

**The shape of y\_test= (14546,)**

**Models Implementation and Evaluation**

**Decision Tree Regression**

The Decision Tree algorithm is part of the supervised learning algorithms family. The decision tree approach, unlike other supervised learning algorithms, may also be utilised to solve regression and classification issues.

By learning simple decision rules inferred from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data).

We start from the root of the tree when using Decision Trees to forecast a class label for a record. The values of the root attribute and the record's attribute are compared. We follow the branch that corresponds to that value and jump to the next node based on the comparison.

Types of Decision Trees

The many types of decision trees we have are determined by the type of target variable we have. There are two types of it:

**Categorical Variable Decision Tree**: A Categorical Variable Decision Tree is a decision tree with a categorical target variable.

**Continuous Variable Decision Tree:** A Continuous Variable Decision Tree is one that has a continuous target variable.

|  |  |
| --- | --- |
| Output |  |
| R2 score | 0.6802 |
| Explained Variance score | 0.6802 |
| Mean absolute error | 11051.3634 |
| Mean squared error | 259812460.3719 |
| Root Mean squared error | 16118.6991 |

**Linear regression model**

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 men and total women are related to the dependent variable (total population). According to the model, if the number of total men or total women is 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.

|  |
| --- |
| Output |
| R2 score = 0.7297 |
| Explained Variance score= 0.7297 |
| Mean absolute error = 10484.2429 |
| Mean squared error = 219650701.6112 |
| Root Mean squared error = 14820.6175 |

**Support Vector Regression(SVR)**

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

|  |
| --- |
| Output |
| R2 score = 0.8264 |
| Explained Variance score= 0.8385 |
| Mean absolute error = 8870.0395 |
| Mean squared error = 141049190.7392 |
| Root Mean squared error = 11876.4132 |

# Random Forest Regression Model

# 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.

We used the Random Forest Regression Model on our dataset to get the more accuracy of dataset. We got the accuracy through random forest is 0.84% higher than other model.

|  |
| --- |
| Output |
| R2 score = 0.8416 |
| Explained Variance score= 0.8418 |
| Mean absolute error = 7748.7486 |
| Mean squared error = 128674338.5799 |
| Root Mean squared error = 11343.4712 |

# Accuracy

## **Table 7: Accuracy Table**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | Model | Train\_test\_split | | 1 | Decision Tree | 0.66 | | 2 | linear Regression | 0.73 | | 3 | SVR | 0.83 | | 4 | Random Forest=50 | 0.85 | | 5 | Random Forest=100 | 0.85 | |

# After training and testing the model of the dataset, it is very important to check the accuracy of the models. In this tarining and testing models, the accuracy to be calculated is 0.84.

**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, decision tree, Support Vector Regression 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.

Based on all the above model evaluation results that we have created with different attribute sets, we applied the Decision Tree , Linear Regression, Support Vector Based Regression Model and Random Forest Based Regression Model on our dataset for better accuracy.

* Obtained model accuracy score of about 83% with support vector regression.
* Obtained model accuracy score of about 85% with Random Forest based Regression model.
* At the end, we recommend the Random Forest Based regression model which provide the good accuracy 85% as compared to another model. So I choose the Random Forest based regression model is best fit model for my project.

# 

# Source of Reference

1. Bureau, U., 2022. Narrative Profiles. Census.gov.

Available at: <https://www.census.gov/acs/www/data/data-tables-and-tools/narrative-profiles/2017/> [Accessed 14 April 2022].

1. Census.gov. 2022.Available at: <https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs\_researchers\_handbook\_2020.pdf> [Accessed 18 April 2022].
2. Bureau, U., 2022. American Community Survey 2016-2020 5-Year Data Release. Census.gov.Available at: <https://www.census.gov/newsroom/press-kits/2021/acs-5-year.html> [Accessed 14 April 2022].
3. En.wikipedia.org. 2022. American Community Survey - Wikipedia. Available at: <https://en.wikipedia.org/wiki/American\_Community\_Survey> [Accessed 18 April 2022].
4. Bureau, U., 2022. American Community Survey 5-Year Data (2009-2020). Census.gov. Available at: <https://www.census.gov/data/developers/data-sets/acs-5year.html> [Accessed 18 April 2022].
5. Www2.census.gov. 2022. Available at: <https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2015nct-race-ethnicity-analysis.pdf> [Accessed 18 April 2022].

**GitHub account Info**

[**https://github.com/Jashan0087/Capstone-Project-**](https://github.com/Jashan0087/Capstone-Project-)