Project : Predict the Income level of population of the U.S. by using Machine Learning Algorithim

Name : Taranjeet Kaur

Student Id : 0776290

DAB402 : Capstone Project

St. Clair College, Mississauga

Table of Contents

ABSTRACT

Income is one of the most important source of well-being, but it is difficult to measure accurately of each person. In the United States, income data are available from surveys, tax records, and government programs, but each of these sources has important power and major limitations when used alone. This data set contains to the census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau.

The aim of my this project is to apply several machine learning algorithms to accurately model individuals' income, whether he makes more than $50,000 or not, using data collected from the 1994 U.S. Census.

I have choose to this dataset, probably the most interesting topics for census data users are income and poverty. I want to know how many people live in a place and also want to know something about how well those people are living. Income is generally used as a measure of the economic well-being of individuals and communities. So, according to this, my purpose here is to address these questions by presenting descriptions and evaluations of the most commonly used measures of income and poverty in the census. My attention will be on if, when and why you should consider using each. The dataset that will be used is the **Census income dataset**, which was and contains about 199523 records and 42 features. My prediction task is to determine whether a person makes his income over 50k a year.

How Does it help

- ❖ Real Estate Demands
- ❖ Basic Amenities
- ❖ Fulfilling Infrastructure Demands

Keywords— Logistic Regression, Random Forest, Naïve bayes classifier, Confusion Matrix, KNN.

Research Questions

The project present some questions about this experiment or survey:

1. Does an individual make more than 50k income or not?

   Answer: After visualization with the distribution of each attribute and it achieve the possibility of earning more than $50,000 per year. Due to this study, I culminate that age', 'education', 'hours per week', and 'sex', these attributes play a crucial rule for predict the income of individual in U.S. Finally, after applying the four models and also using cross validation technique, I have chosen to The Random Forest Classifier, which provides the highly accuracy to known about the income level.

2. What are the most important features that help to define the income of an individual?

   Answer: During analysis, I find that 'age', 'education', 'hours per week' and 'sex' play an important role to determine the income level of individuals in United State.

3. Which model more helpful to know about the income level of each person?

   Answer: I have applied to four models on my dataset to predict the income level of individuals in U.S. For instance, I have used to KNN, Logistic Regression, Random Forest, Naïve Bayes classifier. After applied to these models on my dataset, I get 99.99% accuracy by Random forest classification model. As a result, Random forest is the best fit model to predict the income level of individuals.

4. Which kind of people affected by their income level including: job type such as Private Job, Government Job, Self-employed – not incorporated?

   Answer: By visualization, People who was working in Private sector earned more income as compared to those who were working in Government, and run their own business.

TOOLS:-

I will done my all visualization and calculation with the help of Python.

GITHUB SOURCE:

https://github.com/TaranjeetKaur99/TaranjeetKaur99.git%20

Introduction:

Over the last two decades, humans experience grown a lot of dependence on data and information in community and with this origin growth, technologies have advance for their storage, analysis and processing on a large scale. The fields of Data Mining and Machine Learning have not only make the most of them for knowledge and discovery but also to explore certain hidden patterns and concepts which relate to the prediction of future events, actually this is not easy to obtain. The problem of variation in income has to become great concern in the recent years. Making the poor better off does not seem to be the unique criteria to be in search for eliminate this issue. The folks of the United States believe that the approach of diversity is unacceptable and demands a fair share of wealth in the society. My project actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals. My paper has been structured as an introduction, literature review. The data for my study was accessed from the University of California Irvine (UCI) Machine Learning Repository. It was actually extracted by Barry Becker using the 1994 census database. The binomial label in the data set is the income level which predicts whether a person earns more than 50 Thousand Dollars per year or not based on the given set of attributes.

Literature Review:

Certain efforts using machine learning models have been made in the past by researchers for predicting income levels. The informative power of integrated data sets is improved when demographic attributes are combined with geographic ones. Certainly, a large source of geographic attribution data is the U.S. Census Bureau, which publishes the results of each 10-year census. This data details all sorts of demographic information about geographic regions as small as a census tract (on the order of a few thousand people). Users should learn what is known about the quality of census data, in comparison with other sources, and determine which source is best suited to their use.

Census data are collected once every 10 years and made available 1 to 3 years after collection, which may affect analyses for areas experiencing rapid change. Population estimates update the basic demographic information for small areas; The **American Community Survey** ACS is intended to provide updated long-form-type information if it is implemented. The census misses some people, duplicates others, and puts others in the wrong location. Although the census is the best source of population counts (it has better coverage than household surveys), recognition of coverage problems is important when using the data.

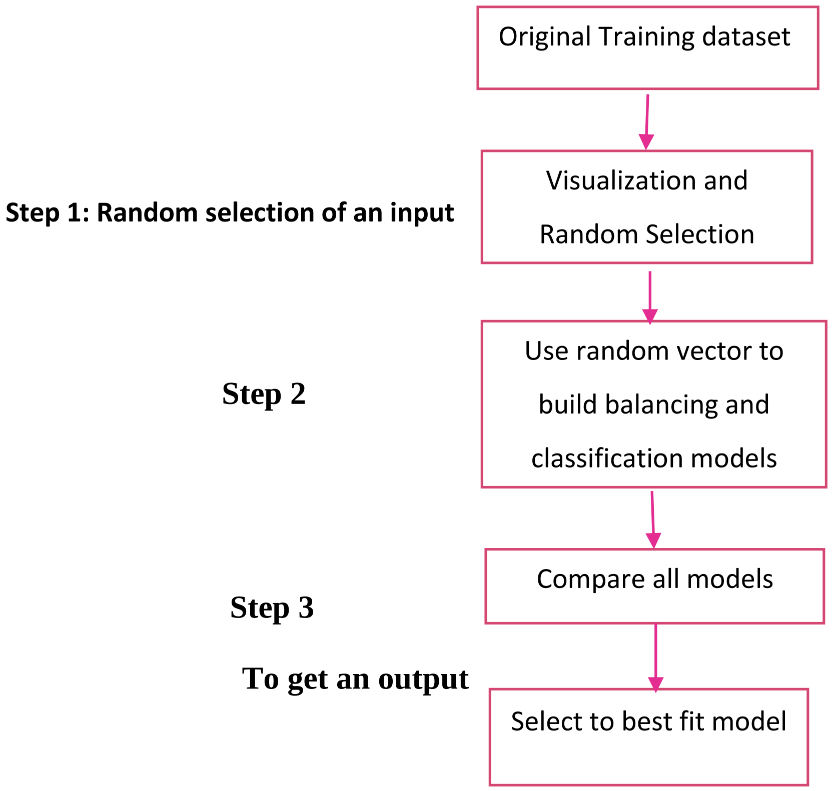
The census strives to obtain complete data for everyone but, even after follow-up, obtains no or limited information for some households. Here they are used to some methods which are helps to recognize the problem, Imputation methods use data from nearby households to supply records for whole-household no offenders and to fill in individual missing items on person and housing unit records. Missing data rates are high for some long-form items, such as income and housing finances, and for residents of group quarters. Imputation introduces variability in estimates and may bias estimates if bystanders differ from prisoners in ways that the imputation procedures do not reflect.

People may underreport or overreport items such as age, income, and utility costs, which can introduce bias in agenda if under- and overreports do not cancel each other out. People may also have different interpretations of questions such as family. Long-form estimates are based on large samples, but they can have substantial uncertainty for small areas or small groups. Sampling variability for the ACS, if implemented, will be higher than for the long form. For analyses of more than one census, changes in definitions or processing features can affect comparability across time.

Chockalingam et. Al. explored and analysed the Adult Dataset and used several Machine Learning Models like Logistic Regression, Stepwise Logistic Regression, Naive Bayes, Decision Trees, Extra Trees, k-Nearest Neighbor, SVM, Gradient Boosting and 6 configurations of Activated Neural Network. They also drew a comparative analysis of their predictive performances. (Bekena) implemented the Random Forest Classifier algorithm to predict income levels of individuals. (Topiwalla) made the usage of complex algorithms like XGBOOST, Random Forest and stacking of models for prediction tasks including Logistic Stack on XGBOOST and SVM Stack on Logistic for scaling up the accuracy. (Lazar) implemented Principal Component Analysis (PCA) and Support Vector Machine methods to generate and evaluate income prediction data based on the Current Population Survey provided by the U.S. Census Bureau. (Deepajothi et. al.) tried to replicate Bayesian Networks, Decision Tree Induction, Lazy Classifier and Rule Based Learning Techniques for the Adult Dataset and presented a comparative analysis of the predictive performances. Another one is (Lemon et. al.) attempted to identify the important features in the data that could help to optimize the complexity of different machine learning models used in classification tasks.

I am now proceed to an important part of my process — data modeling. The first step that I have done to check the information about my dataset. Based on the analysis above, I will fill the missing values in to data, and group certain categories logically, to allow my project to learn better. I also will try to proceed with visualization in my dataset, which allows to fit my data in a symmetric distribution, which further allows the model to learn better. I would like to check the correlation between the numeric columns. Afterthat, I will proceed to the main step of machine learning, fitting the model and predicting the outputs. I will fit the data into multiple classification models to compare the performance of all models and select the best model. After all process, I will try to observe that Random Forest Classifier over fits to the train set, and obtain a 99% accuracy on train data. I will be moved step by step, analyzing, cleaning and modeling the data, and applied various machine learning models to achieve the desired predictions. I also tuned the model to improve the accuracy, and were able to achieve a model with quite a good accuracy. This become my final output of my project which is relate to predict income level of individuals of the Census Dataset of U.S. Information from the census helps the government and local authorities to plan and fund local services, such as education, doctors' surgeries and roads.

Methodology


Original Training dataset
Step 1: Random selection of an input
Visualization and Random Selection
Step 2
Use random vector to build balancing and classification models
Step 3
Compare all models
To get an output
Select to best fit model

Detailed data dictionary:-

Features Names	Description	Type	No. of levels	Categories
Age	Age of the worker	Numeric	91	Count: 6 , 91
class_worker	Class of worker	String	9	Categorical: Private, Government, Self – Employed
det_ind_code	Industry code	Numeric	52	Number : 0 to 52
det_occ_code	Occupation code	Numeric	47	Number : 0 to 47
education	Level of Education	String	17	Categorical: High school, 10th grade, Children, Degree,
wage_per_hour	Wage per hour	Numeric	1240	Number 0 to 1240
hs_college	Enrolled in educational institution last week	String	3	Categorical: High school, College or University, Not in Universe
marital_stat	Marital status	String	7	Categorical: Single, Married, Divorced, Widowed
major_ind_code	Major industry code	String	24	Categorical: Construction, Finance, Entertainment, Business, Personal service

Features Names	Description	Type	No. of levels	Categories
major_occ_code	Major occupation code	String	15	Categorical: Professional specialty, Executive admin, Machine operator, Precision Production craft
race	Race	String	5	Categorical : Asian, Asian indian aleut, White, Balck, Other
hisp_origin	Hispanic origin	String	10	Categorical: Mexican, South American,All other,
Sex	Sex	String	2	Categorical: Male, Female
union union_member	Member of a labor	String	3	Categorical: Yes, No, Not in universe
unemp_reason	Reason for unemployment	String	6	Categorical: Job loser, Job leaver, Re-entrant, Other
full_or_part_emp	Full- or part-time employment status	String	8	Categorical: Armed forced, Not in labour force, Full time schedules, Unemployed full time
capital_gains	Capital gains	Numeric	132	Number: 0 to 132
capital_losses	Capital losses	Numeric	113	Number 0 to 113
stock_dividends	Dividends from stocks	Numeric	1478	Number 0 to 1478

Features Names	Description	Type	No. of levels	Categories
tax_filer_stat	Tax filer status	String	6	Categorical: Head of household, Nonfiler, Joint both under 65, Single
region_prev_res	Region of previous residence	String	6	Categorical: South, West, Midwest, Northeast, Abroad
state_prev_res	State of previous residence	String	50	Categorical: Arkansas, Utah, Alaska, Not in universe
det_hh_fam_stat	Detailed household and family status	String	38	Categorical: Householder, Secondary individuals, Spouse of householder, Nonfamily householder, others
det_hh_summ	Detailed household summary in household	String	8	Categorical: Chid 18 or older, Householder, Nonrelative
mig_chg_msa	Migration code - change in MSA	String	9	Categorical: MSA to MSA, Nonmover, NonMSA to NonMSA
mig_chg_reg	Migration code - change in region	String	8	Categorical: Same country, Different region, Nonmover

Features Names	Description	Type	No. of levels	Categories
mig_move_reg	Migration code - move within region	String	9	Categorical: Different state in South, Same country, Different region, Nonmover
mig_same	Live in this house one year ago	String	3	Categorical: Yes, No, Not in universe
mig_prev_sunbel	Migration - previous residence in sunbelt	String	3	Categorical: Yes, No, Not in universe
num_emp	Number of persons that worked for employer	Numeric	7	Number 0 to 7
fam_under_18	Family members under 18	String	5	Categorical: Both parents presents, Mother Only present, father Only present, Neither parent present
country_father	Country of birth father	String	42	Categorical: United states, Vietnum, Columbia, Germany, Mexico
country_mother	Country of birth mother	String	42	Categorical: United states, Vietnum, Columbia, Germany, Mexico
country_self	Country of birth	String	42	Categorical: United states, Vietnum, Columbia,

Features Names	Description	Type	No. of levels	Categories
				Germany, Mexico
Citizenship	Citizenship	String	5	Categorical: Native, Foreign, Native- Born in Puerto Rico or U.S. Outlying, Foreign born- U.S. citizen by naturalization
own_or_self	Own business or self-employed?	Numeric	3	Numer 0 to 2
vet_question	Fill included questionnaire for Veterans Admin	String	3	Categorical: Yes, No, Not in universe
vet_benefits	Veterans benefits	Numeric	3	Number 0 to 2
weeks_worked	Weeks worked in the year	Numeric	3	Number: 30, 52, 49
Year	Year of survey	Numeric	2	Year: 94 and 95
income_50k	Income less than or greater than 50,000	String	2	<=50k, >=50k

The descriptive statistics of the dataset:

	count	mean	std	min	25%	50%	75%	max
age	199523	34	22	0	15	33	50	90
det_ind_code	199523	15	18	0	0	0	33	51
det_occ_code	199523	11	14	0	0	0	26	46
wage_per_hour	199523	55	275	0	0	0	0	9999
capital_gains	199523	435	4698	0	0	0	0	99999
capital_losses	199523	37	272	0	0	0	0	4608
stock_dividends	199523	198	1984	0	0	0	0	99999

	count	mean	std	min	25%	50%	75%	max
unknown	199523	1740	994	38	1062	1618	2189	18656
num_emp	199523	2	2	0	0	1	4	6
own_or_self	199523	0	1	0	0	0	0	2
vet_benefits	199523	2	1	0	2	2	2	2
weeks_worked	199523	23	24	0	0	8	52	52
year	199523	94	1	94	94	94	95	95

As per above sort description we have seen in ‘count’ there is no null values in the dataset. And, we can also see ‘mean’ it shows variation among the features and values are on different scales so we have to scale the features in similar scale.

Data Cleaning

To commence with, the data seems to already be pre-processed, since missing values are consistently denoted by a question mark (i.e. “?”) and there are no null values in any of the columns.

Missing Values: Missing values are represented by “?” in this dataset. I have checked how many of those question marks each column has.

I have find, the total of my missing values of result is 415717 in my dataset, after that I replace it with NAN values in each cell.

Feature name	Description	Missing Values(?) [Count]
state_prev_res	State of previous residence	708
mig_chg_msa	Migration code - change in MSA	99696
mig_chg_reg	Migration code - change in region	99696

Feature name	Description	Missing Values(?) [Count]
mig_move_reg	Migration code - move within region	99696
country_father	Country of birth father	6713
country_mother	Country of birth mother	6119
country_self	Country of birth	3393

Variance:

Variance () function takes a sequence or an iterator as the parameter which containing the sample data and returns the sample variance. Its underlying idea is that if a feature of dataset is constant (e.g.: it has zero variance), then it cannot be used for any interesting patterns, and can be dropped from the dataset. A model with high variance may represent the data set accurately but could lead to over fitting.

Features	variance
age	498
industrycode	326
occupationcode	209
wages	75568
capital_gains	22066800
capital_losses	73928
stock_dividends	3936905
others	987575
num_emp	6
own_or_self	0
vet_benefits	1
weeks_worked	596
year	0

Visualization

In this task, I performed a number of visualization tasks to get some information about the data. Visualization is a key component of exploration. We can choose to use either Matplotlib or

Seaborn for plotting, I did it with both matplotlib and Seaborn also. I use to Seaborn because it has a variety of styles.

Visualization with pie chart:

The figure1 contains to original dataset which represent the distribution of 6.21% entries labeled with 50000+ and 93.8% entries labeled with -50000. I split the dataset into training and test sets while maintaining the above distribution. The following graphs to the original dataset. Which shows the income level.

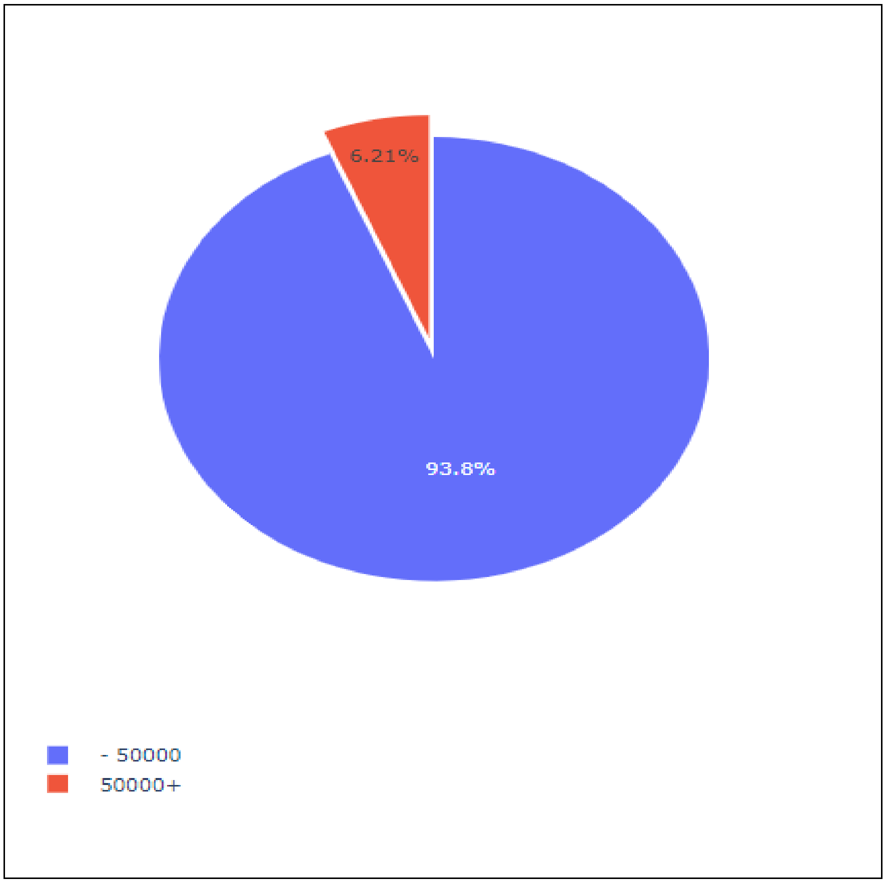


Figure 1 : Distribution level of Income

In the figue 2, I find the the majority of the individuals work in the private sector, if we avoid the not in universe, which is relate the not information . This is one concerning statistic is the number of individuals with not in universe work class. The probabilities of making above $50,000 are similar among the work classes except for self-employed incorporated and federal government. Federal government is seen as the most elite in the public sector, which most likely explains the higher chance of earning more than $50,000. Self-employed incorporated implies that the individual owns their own company, which is a category with an almost infinite ceiling when it comes to earnings.

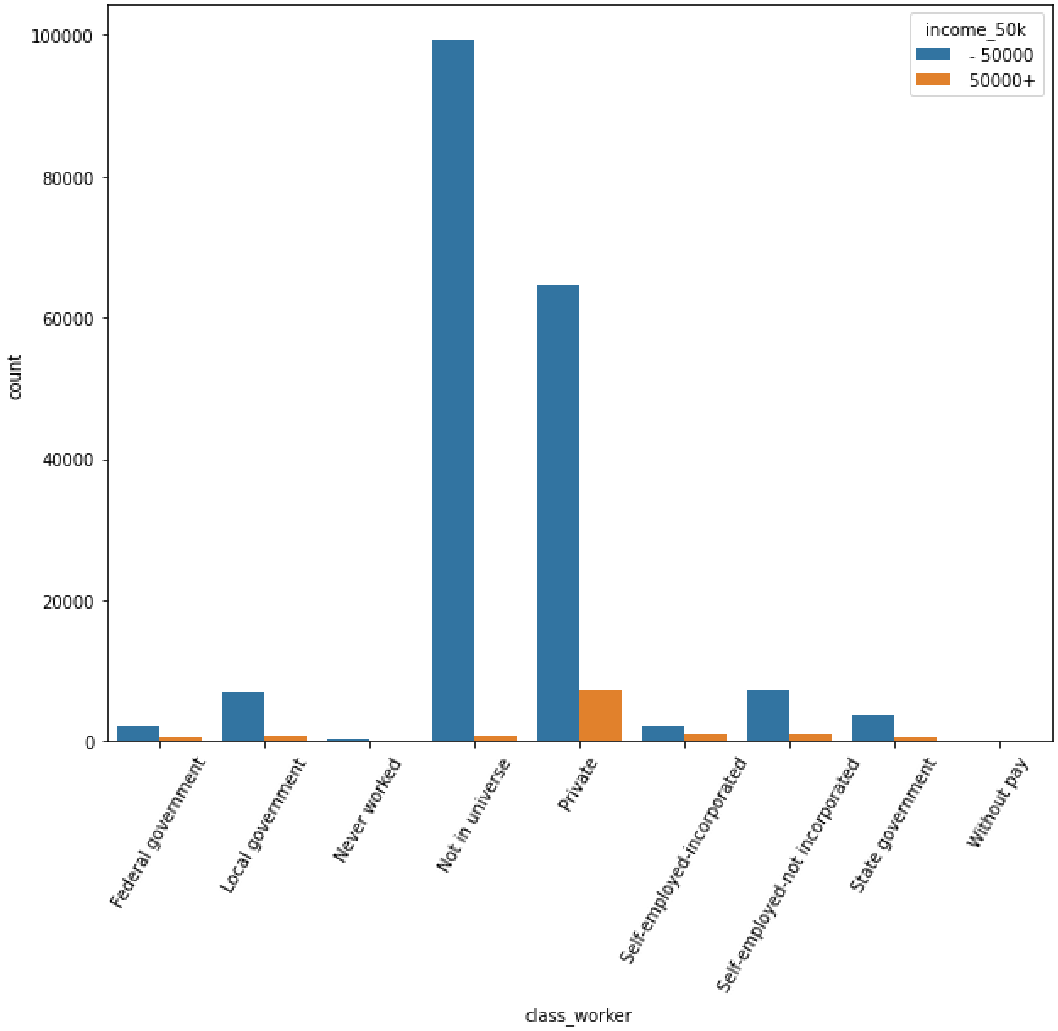


Figure 2 : Figure out class worker versus income level status

In the below figure 3, we can see the education feature describes the highest level of education of each individual in the dataset. It shows the distribution of the different levels of education among individuals in the dataset. The Other group represents Preschool through 12th grade. Most of the individuals in the dataset have at most a high school education while only a small portion have a doctorate. I think this is a equitable representation.

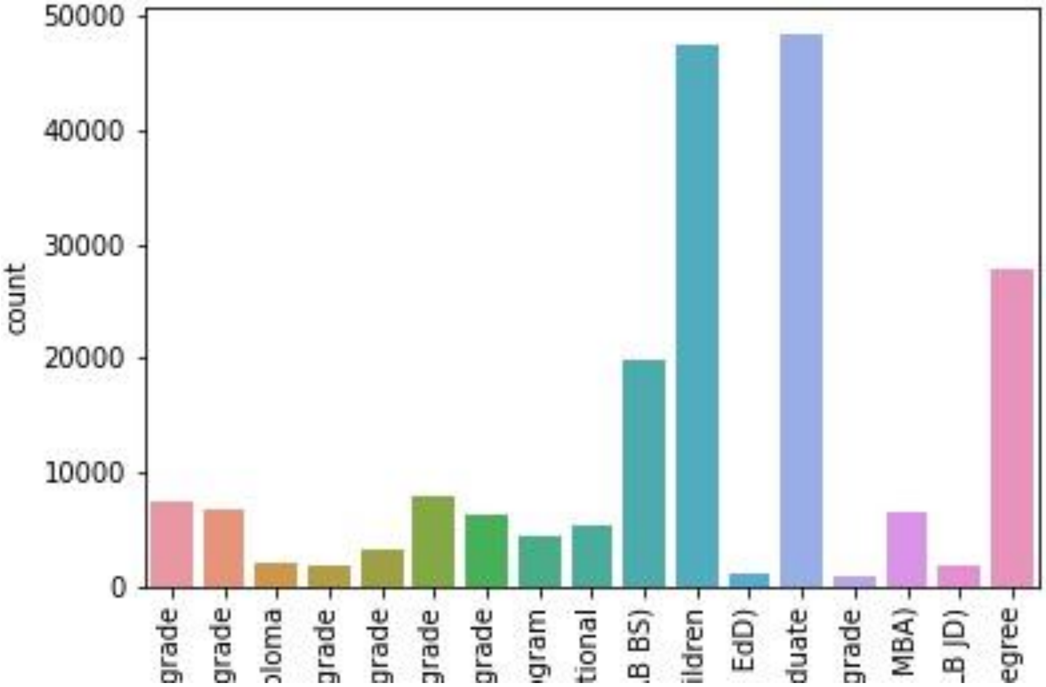


Figure 3 : highest Level Education Distribution

In the below figure 4, we can see the distribution of marital status, which is related the information about individual regarding their matrimonial status. We can figure out majority of folk who are unmarried and, those people which are stay their spouse, make highest amount of income.

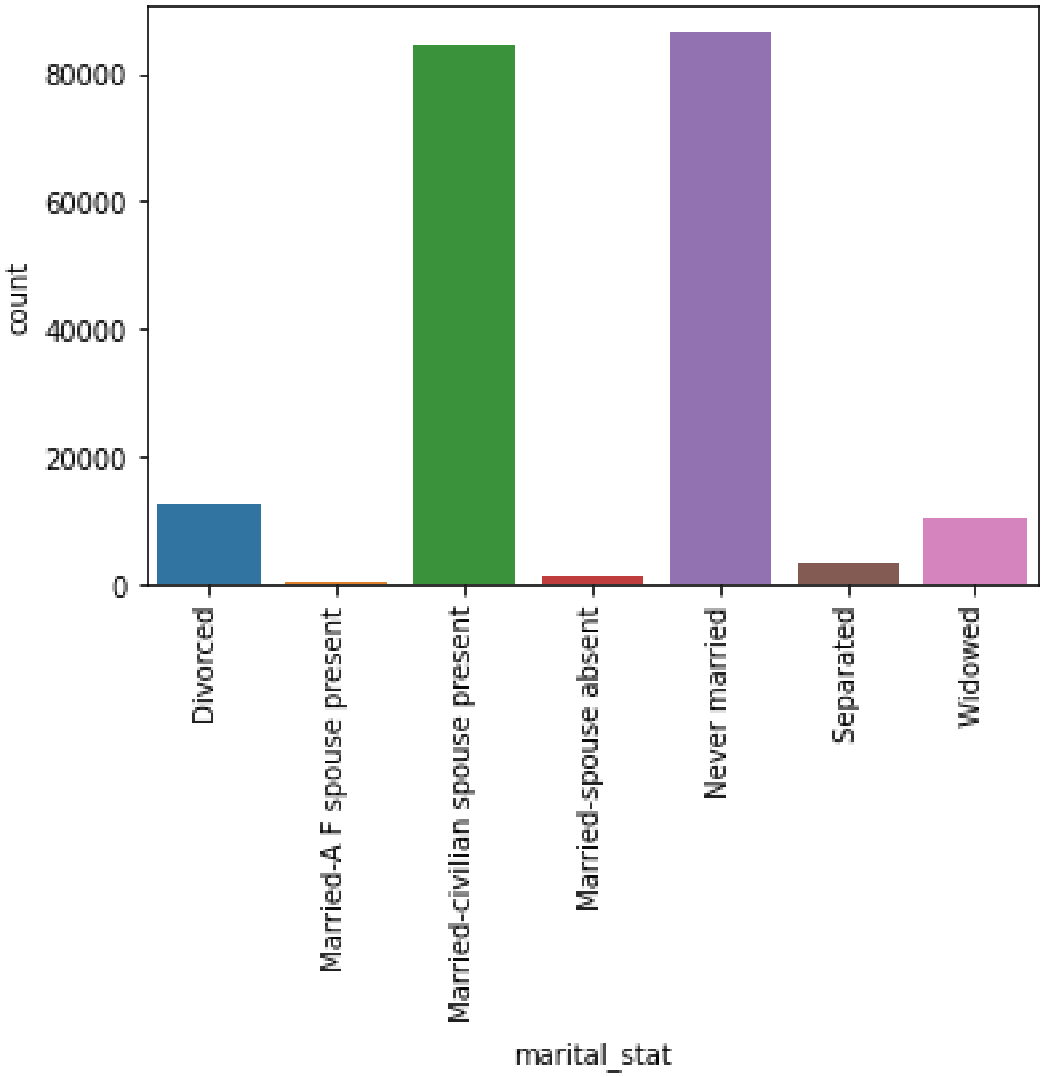


Figure 4 :   Maritial Status

In Figure 5, we can see that there is almost double the sample size of males in comparison to females in the dataset. While this may not affect our predictions too much, as well as distribution of income. Also the percentage of males who make greater than $50,000 is much greater than the percentage of females that make the same amount of income. This will certainly be a noteworthy component, and should be a feature considered in our prediction model.

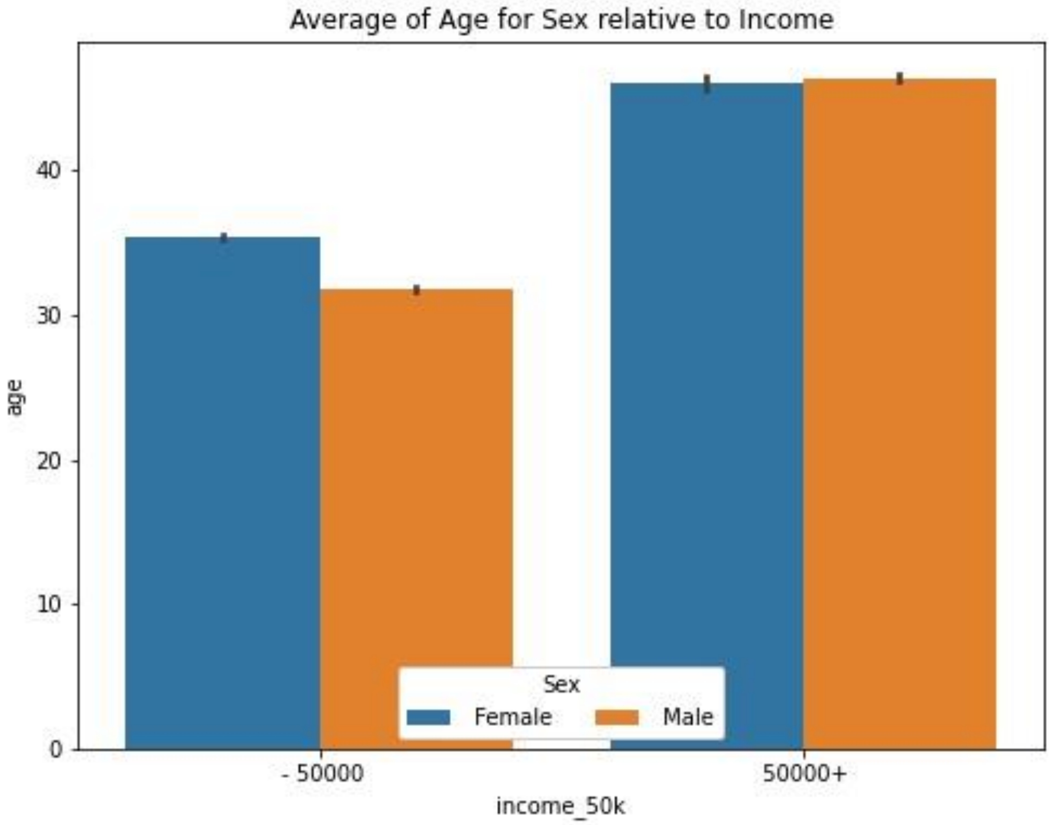


Figure 5 Age versus Income level

Looking forward at the distribution in Figure 6, the vast majority of individuals are working full time which is expected as the morals. Regardless of the non-uniform distribution, that the percentage of individuals making over $50,000 drastically decreases when less than 40 hours per week, and increases significantly when greater than 40 hours per week.

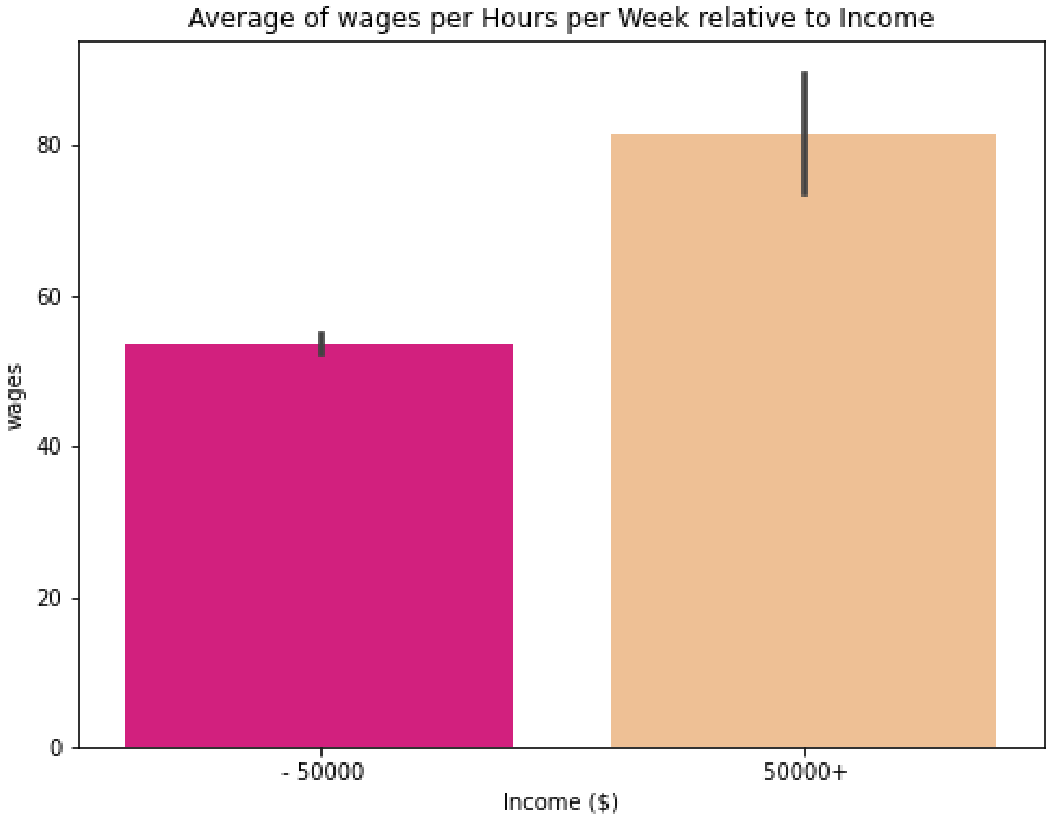


Figure 6 Wages versus Income

In the below figure, the age feature describes the age of the individual. Figure 7 shows the age distribution among the entries in our dataset. The ages range from 17 to 90 years old with the majority of entries between the ages of 25 and 50 years. Because there are so many ages being represented, we can see closely to the entries into age groups with intervals of ten years to present the data more concisely.

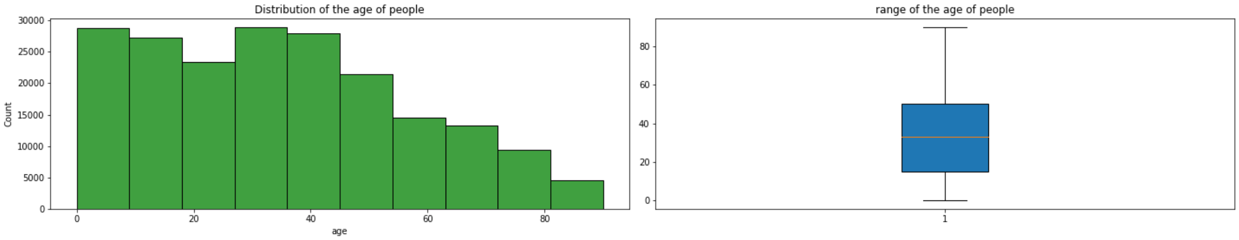


Figure 7: Distribution of age of people

In the figure 8, I have shown to Correlation which is a statistical term and probably in common usage it refers to how close two variables are to having a linear relationship with each other. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have shown the highest correlation, after that we can drop one of the two features.

In my dataset, we can see here is no high correlation between any two features, I think, I have not need to drop any feature.

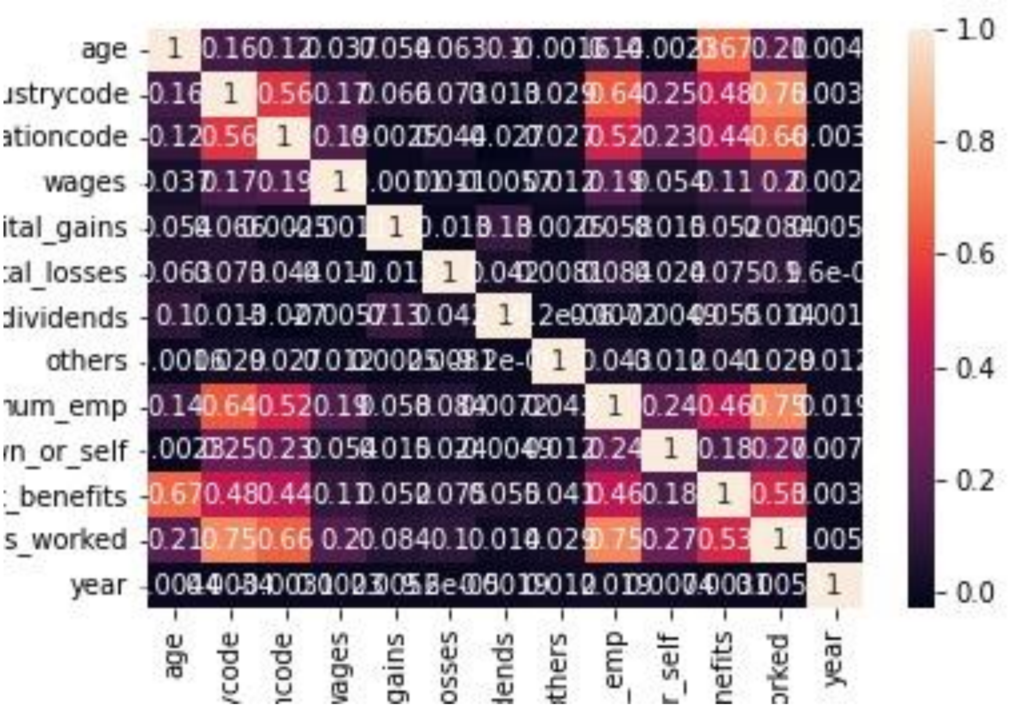


Figure 8 Heat-Map showing Feature-to-Feature and Feature-to-Label's Pearson Correlation Coefficients

DATA PREPROCESSING

Min Max Scaling:

Min Max Scalar drop off the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution. So, I choose to this for my Age and Income attribute to get data range between 0 and 1 corresponding to detail of these attributes. It is always seen that, a higher level of education is correlated to a higher percentage of individuals with the label who makes greater than 50,000. One interesting thing to note is the ratio of individuals labeled 50000+

to -50000 is almost the same between those that have a doctorate and those that went to a professional school.

Imblearn Technique:

My dataset has consists of one class is in the majority and if the class has above 90% weightage and another class has less than 10% weightage then the dataset is imbalanced. The imbalanced dataset is highly biased towards one class, it creates a problem to train our machine learning model. Machine learning is not able to identify minority class correctly.

As a result, first I have balance my dataset by using SMOTE, it is refers to Synthetic Minority Oversampling Technique. SMOTE selects the data points of the minority class in feature space to draw a line between those points and generate new points along with the line. Thus this technique synthesizes new data points for minority class and oversample that class.

Train and Test Split Approach:

Training and testing is a method, which helps to compute the accuracy of our model. It is called Train/Test because of we split the the data set into two sets: a training set and a testing set. I have used 70% for training, and 30% for testing. I train the model using the training set and, afterthat test the model using the testing set.

Data Modeling:

To predict something useful from the datasets, we need to implement machine learning algorithms. Since, there are many types of algorithms like Logistic Regression, Random Forest, Naïve bayes classifier, Confusion Matrix, KNN, which I have applied these all on my dataset to predict the accuracy and with the help of accuracy, know about best fit model for my dataset.

Logistic Regression :

In my dataset, the dependent variable is categorical in this case. My target variable is Income, which relates to two labels such as 50000+ and -50000. So, that is only reason, I have applied to logistic regression on my dataset to predict the accuracy. When the outcome is ‘0’ or ‘1’, it indicates success/failure. This model is used to find the probability of binary output based on the predictor variable. Although it says ‘regression’, this is actually a classification algorithm. I have used explainer instance for logistic regression has a method named fit() and score() which tells us fit train data to model and calculate model accuracy on the test dataset.

Naïve Bayes Classifier :

Naive Bayes is a classification method which is based on Bayes’ theorem. This assumes independence between predictors. A Naive Bayes classifier will assume that a feature in a class is unrelated to any other. A Naive Bayes classifier will say these characteristics independently contribute to the probability of the attribute. This is even if features depend on each other. This model is easy to build a Naive Bayesian model. Not only is this model very simple, it performs better than many highly sophisticated classification methods to predict the accuracy.

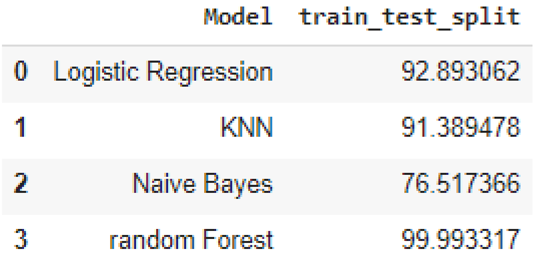
KNN - (K - Nearest- Neighbors) :

This is a Machine Learning algorithms for classification and regression- mostly for classification. This is a supervised learning algorithm that considers different centroids and uses a usually Euclidean function to compare distance. Then, it analyzes the results and classifies each point to the group to optimize it to place with all closest points to it. It classifies new instance using a majority values of k of its neighbors. The instance it assigns to a class is the one most common among its K nearest neighbors. For it’s, this is use to a distance function to predict accuracy.

Random Forest Classifier:

It is perhaps the most popular and widely used machine learning algorithm given its good or excellent performance across a wide range of classification and regression predictive modeling problems. In other words random forest is a meta estimator like effective way to achieve goal that fits a number of classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

#comparing accuracy



	Model	train_test_split
0	Logistic Regression	92.893062
1	KNN	91.389478
2	Naive Bayes	76.517366
3	random Forest	99.993317

By seeing the above output, I observe that Random Forest Classifier over fits the train set, and gives a 99% accuracy on train data.

Cross Validation:

Cross Validation is a technique which involves preserve a specific test of a dataset on which we do not train the model. After that, we test our model on this specific test before finalizing it.

Here are the steps involved in cross validation:

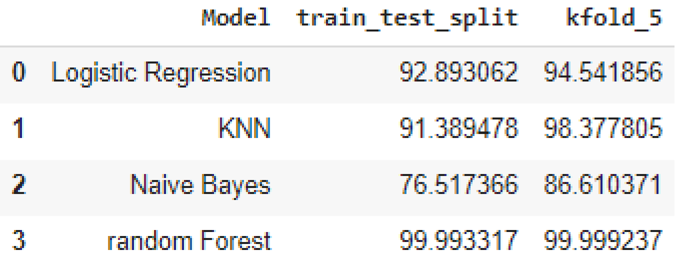
- First, we reserve a sample or specific test of data set.
- Train the model using the remaining part of the dataset.

- In the final, we use the reserve sample of the test (validation) set. This will help in measure the effectiveness of our model's performance. If our model delivers a positive result on validation data, then go ahead with the current model.

K- Fold Cross Validation :

In this method, the dataset is split into 'k' number of subsets, k-1 subsets then are used to train the model and the last subset is kept as a validation set to test the model. Then the score of the model on each fold is averaged to evaluate the performance of the model. I have implemented 5 fold cross-validation with the help of sklearn.model_selection module which provides with K-Fold class (I have used four classes with K-fold), and this helps to make it easier to implement cross-validation.

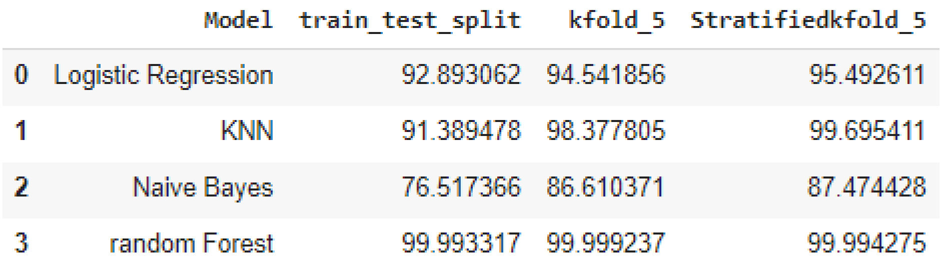
So in this way, my K-Fold classes used to split method which requires a dataset to perform cross-validation on as an input arguments.



	Model	train_test_split	kfold_5
0	Logistic Regression	92.893062	94.541856
1	KNN	91.389478	98.377805
2	Naive Bayes	76.517366	86.610371
3	random Forest	99.993317	99.999237

Stratified K-Fold:

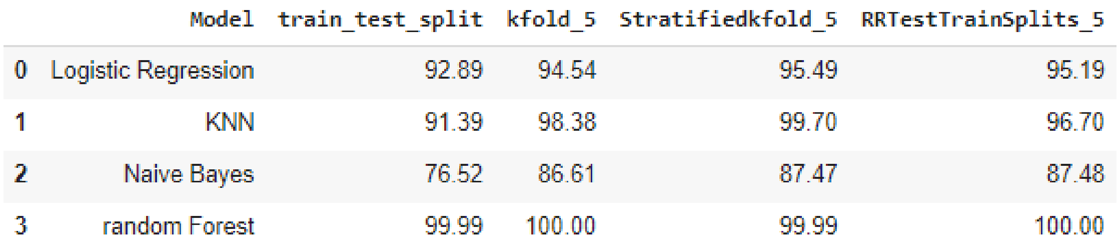
Stratified K fold cross-validation object is a variation of KFold that returns stratified folds. The folds are made by preserving the percentage of samples for each class. It provides train/test indices to split data in train/test sets.



	Model	train_test_split	kfold_5	Stratifiedkfold_5
0	Logistic Regression	92.893062	94.541856	95.492611
1	KNN	91.389478	98.377805	99.695411
2	Naive Bayes	76.517366	86.610371	87.474428
3	random Forest	99.993317	99.999237	99.994275

Shuffle Split k-fold:

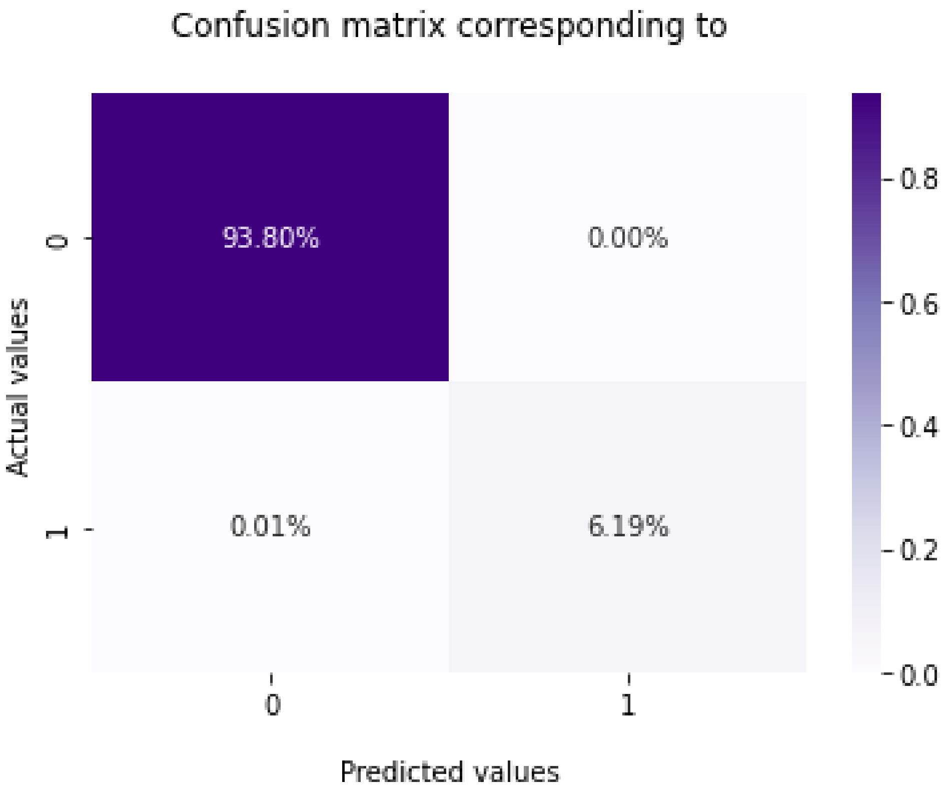
This is a very flexible strategy of cross-validation. In this technique, the datasets get randomly partitioned into training and validation sets.



	Model	train_test_split	kfold_5	Stratifiedkfold_5	RRTestTrainSplits_5
0	Logistic Regression	92.89	94.54	95.49	95.19
1	KNN	91.39	98.38	99.70	96.70
2	Naive Bayes	76.52	86.61	87.47	87.48
3	random Forest	99.99	100.00	99.99	100.00

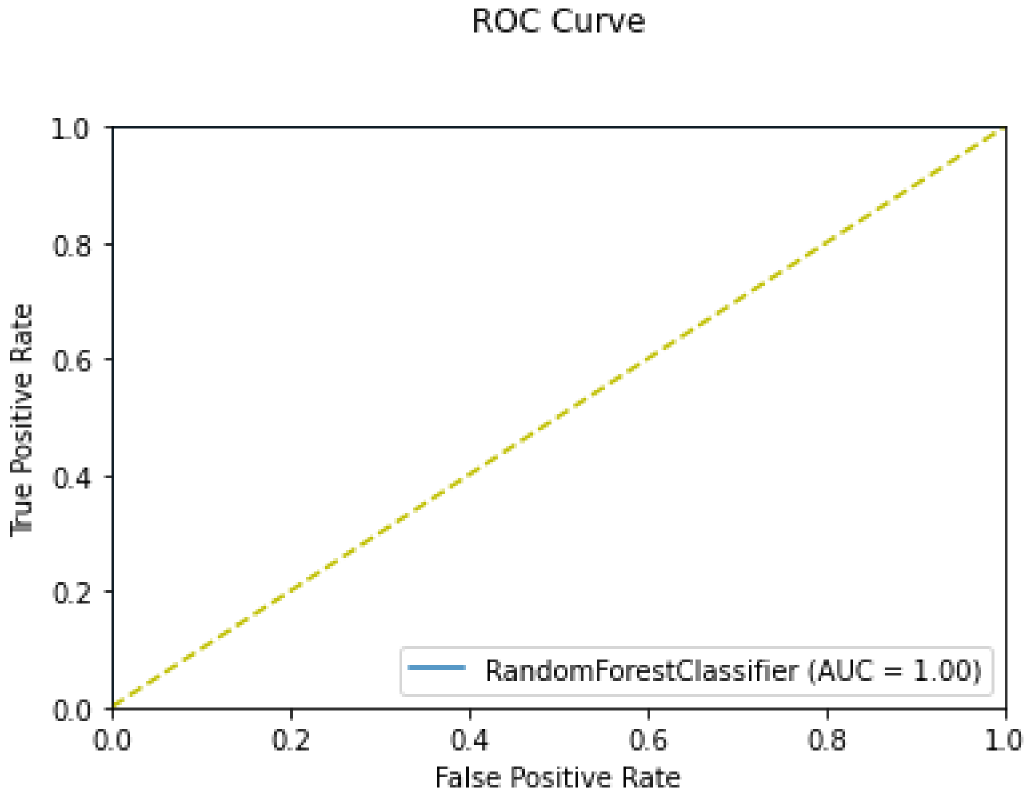
Confusion matrix :

A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier. A confusion matrix is such that C (i , j) is equal to the number of observations known to be in group and predicted to be in group .Thus in binary classification, the count of true negatives is C (0 , 0), false negatives is C (1 , 0) , true positives is C (1 , 1) and false positives is C (0 , 1).This confusion matrix compares the actual target values with those predicted by the machine learning model.



ROC Curve :

The Receiver Operator Characteristic Curve (ROC) class will be used for creating an ROC AUC curve for our model. It accepts prediction function and feature names for creating explainer instance. We'll then call the show () method passing it this explanation instance to generate the ROC curve on given data.

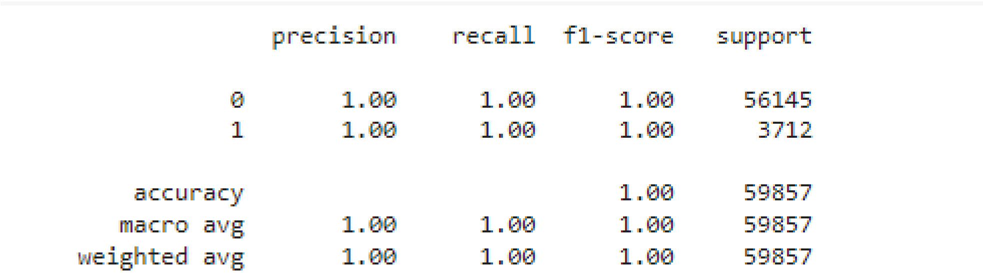


According to the obtained Training and Validation accuracy, it can be seen that my model is a good fit. The Area Under the Curve (AUC) shown in the above figure, which is 1.00, which is represent a better performance of the model because this is perfectly distinguish between all the Positive and the Negative class points correctly.

Classification Report:

It seems to be precision as the proportion of times that when we predict its positive which is actually turns out to be positive. Whereas recall can a accuracy over just the positives, so, all in all it is the proportion of times which we have labeled with positive correctly over the amount of times which was actually positive.

I have received precision, recall and f1-score (weighted average of precision and recall) are same result with 1.00 that means my algorithm has classified as equal amount of values.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56145
           1       1.00      1.00      1.00      3712

    accuracy                           1.00     59857
   macro avg       1.00      1.00      1.00     59857
weighted avg       1.00      1.00      1.00     59857
```

Conclusion:

The first step, I took was to visualize the distribution of each attribute and its effect on the likelihood of earning more than $50,000 per year. From this analysis, I concluded that the most useful features for prediction were ‘age’, ‘education’, ‘hours per week’, and ‘sex’. Finally, after applying the four model and also using cross validation technique, I have chosen to The Random Forest Classifier, due to the fact that for achieving very high results in terms of evaluation metrics. It has 99.99% accuracy score, also the same result for F1-score, it is reached at 1.00. So, this means, that people earned more than $50,000 per year behalf of the features and accuracy.

References:

- Vidya Chockalingam, Sejal Shah and Ronit Shaw: “Income Classification using Adult Census Data”.
- Mohammed Topiwalla: “Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting”, University of SP Jain School of Global Management.
- Chet Lemon, Chris Zelazo and Kesav Mulakaluri: “Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”.

- SMOTE: Synthetic Minority Over-sampling Technique –Nitesh V. Chawla – Journal of Artificial Intelligence Research 16(2002)321-357.
- Ron Kohavi, "Scaling Up the Accuracy of Naïve-Bayes Classifier: a Decision- Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.
- Haojun Zhu: "Predicting earning Potential using the Adult Dataeset", https://archive.ics.uci.edu/ml/datasets/Adult.