**Data -** **ipums\_la\_98-small**

**What the dataset is about?**

* IPUMS means Integrated Public Use Micro data Series.
* IPUMS is the world's largest individual-level population database.
* IPUMS consists of micro data samples from United States (IPUMS-USA) and international (IPUMS-International) census records, as well as data from U.S. and international surveys.
* IPUMS-International distributes integrated micro data of individuals and households only by agreement of collaborating national statistical offices and under the strictest of confidence.

**Rules of data:**

* Before data may be distributed to an individual researcher, an electronic license agreement must be signed and approved.
* To gain access to the data, a researcher must agree to the following:

1. Implement security measures to prevent unauthorized access to census micro data. Under IPUMS-International agreements with collaborating agencies, redistribution of the data to third parties is prohibited.
2. Use the micro data for the exclusive purposes of scholarly research and education. Researchers must explicitly agree to not use micro data acquired for any commercial or income-generating venture.
3. Maintain the confidentiality of persons, households, and other entities. Any attempt to ascertain the identity of persons or households from the micro data is prohibited. Alleging that a person or household has been identified is also prohibited.
4. Report all publications based on these data to IPUMS-International, which will in turn pass the information on to the relevant national statistical agencies.

**Exploratory**

* The data is about the International Household Survey.it deals with many features of household.
* My data set has **61 features column** (year,gq,gqtypeg,farm,ownershg,value,rent,ftotinc,nfams,ncouples,nmothers,nfathers,momloc,stepmom,momrule,poploc,steppop,poprule,sploc,sprule,famsize,nchild,nchlt5,famunit,eldch,yngch,nsibs,relateg,age,sex,raceg,marst,chborn,bplg,school,educrec,schltype,empstatg,labforce,occ1950,occscore,sei,ind1950,classwkg,wkswork2,hrswork2,yrlastwk,workedyr,inctot,incwage,incbus,incfarm,incss,incwelfr,incother,poverty,migrat5g,migplac5,movedin,vetstat,tranwork) and **1 target column** (Movedin).
* Checked the missing values.
* Cleaned the data without any null values.
* All my data variables consists of numerical and categorical data and the target column has 0 and 1.
* Normalized the numerical data into 0 - 1.
* Did label-encoding to all the categorical data to convert them as 0, 1, 2, 3…. Depends on different categories in each column.
* Did one-hot-encoding to convert the categorical data into variance of 0 - 1
* Converted al the normalized and one-hot-encoding data into data frame and combined the entire data
* Data is spited into training and test data.

**TRIED DIFFERENT CLASSIFICATION METHODS TO GET “90% ACCURACY”**

**Logistic Regression:**

* Logistic Regression Model is trained on the training data set and tested on test data
* The trained data is applied on the test data.
* Now the accuracy of the model is-> “**77.8%”**
* Analyse the confusion matrix

**Neural Network -1 (3 hidden layers of size 2000, 1000 & 800respectively)**

* Neural NetworkModel is trained on the training data set and tested on test data
* Now the accuracy of the model is -> **“77.1%”**
* Analyse the confusion matrix

**Neural Network -2 (3 hidden layers of size 1500,900 and 700respectively)**

* Neural NetworkModel is trained on the training data set, and applied on the test data
* Now the accuracy of the model is found to be-> **“77.3%”**
* Analyse the confusion matrix

## Decision tree:

* Decision tree Model is trained on the training data set, and applied on the test data
* Now the accuracy of the model is-> “**77.6%”**
* Analysed the confusion matrix

**Support vector machine:**

* Support vector machine Model is trained on the training data set, and applied on the test data
* Now the accuracy of the model -> “**76.8%”**
* Analysed the confusion matrix
* Then the confusion matrix is plotted, wherein ***False positive*** is **1.11%** and ***False Negative*** is **90%**

**Support vector machine: (LINEAR FUNCTION)**

* In Support vector machine we implement linear function model is trained on the training data set, and applied on the test data
* Now the accuracy of the model is found to be-> “**88.6%”**
* Analysed the confusion matrix

**Random Forest:**

* Random Forest Model is trained on the training data set, and applied on the test data
* Now the accuracy of the model is found to be-> “**100%”**
* Analysed the confusion matrix

**Comparison:**

* From all the different classification methods there observed a slight increase of accuracy at every method.
* The major change of accuracy in **Support vector machine-(LINEAR FUNCTION) which is 88.6%** and in **Random Forest which is 100% accuracy.**
* In remaining method, the accuracy of **Logistic Regression, Neural Networks** and **Decision Tree** and **SVM** is **77%** only with an increase in decimal points.