**Income Qualification**

**Project Report**

**Aparna Nair**

**aparna.nair.vijay@gmail.com**

**Business Scenario**

Identify the level of income qualification needed for the families in Latin America.

Problem Statement Scenario:

Many social programs have a hard time ensuring that the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of the population can’t provide the necessary income and expense records to prove that they qualify.

In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to classify them and predict their level of need.

While this is an improvement, accuracy remains a problem as the region’s population grows and poverty declines.

The Inter-American Development Bank (IDB)believes that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT’s performance.

**Objectives**

**Analysis Tasks to be perfomed**

1. Identify the output variable.
2. Understand the type of data.
3. Check if there are any biases in your dataset.
4. Check whether all members of the house have the same poverty level.
5. Check if there is a house without a family head.
6. Set poverty level of the members and the head of the house within a family.
7. Count how many null values are existing in columns.
8. Remove null value rows of the target variable.
9. Predict the accuracy using random forest classifier.
10. Check the accuracy using random forest with cross validation.

**DATASET DESCRIPTION**

Train and Test Dataset having 9557 observations of 143 variables and 23586 observations of 142 variables

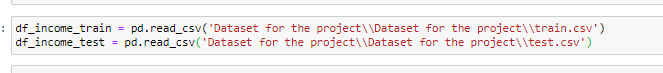
**Statistical algorithm execution – Python code and outputs**

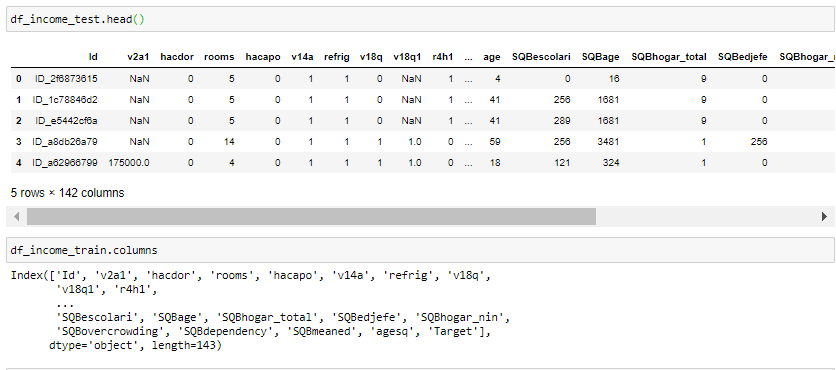
1. Objective 1) Identify the output variable.

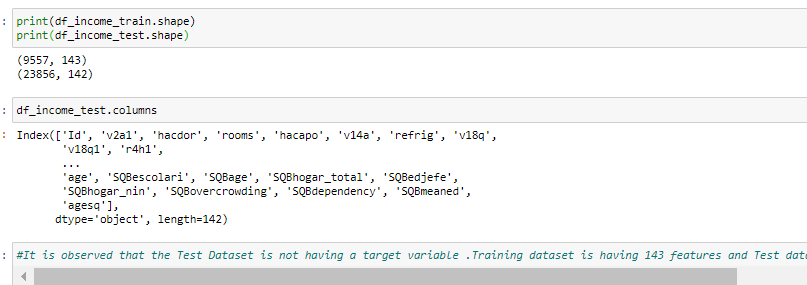
Importing the libraries



Importing the datasets

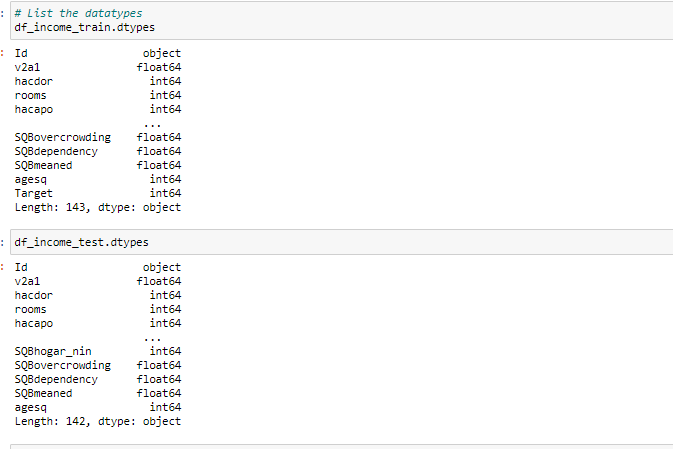


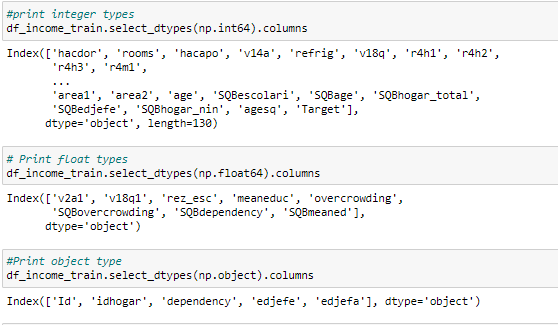




It is observed that the Test Dataset is not having a target variable .Training dataset is having 143 features and Test data set is having 142

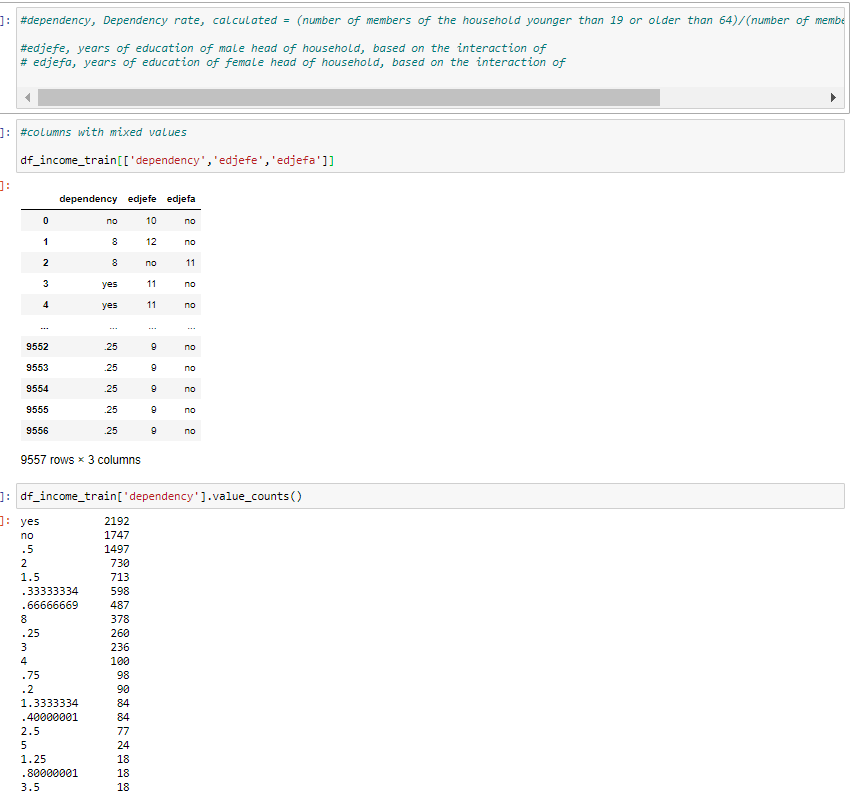
Objective 2 ) Understand the type of data.



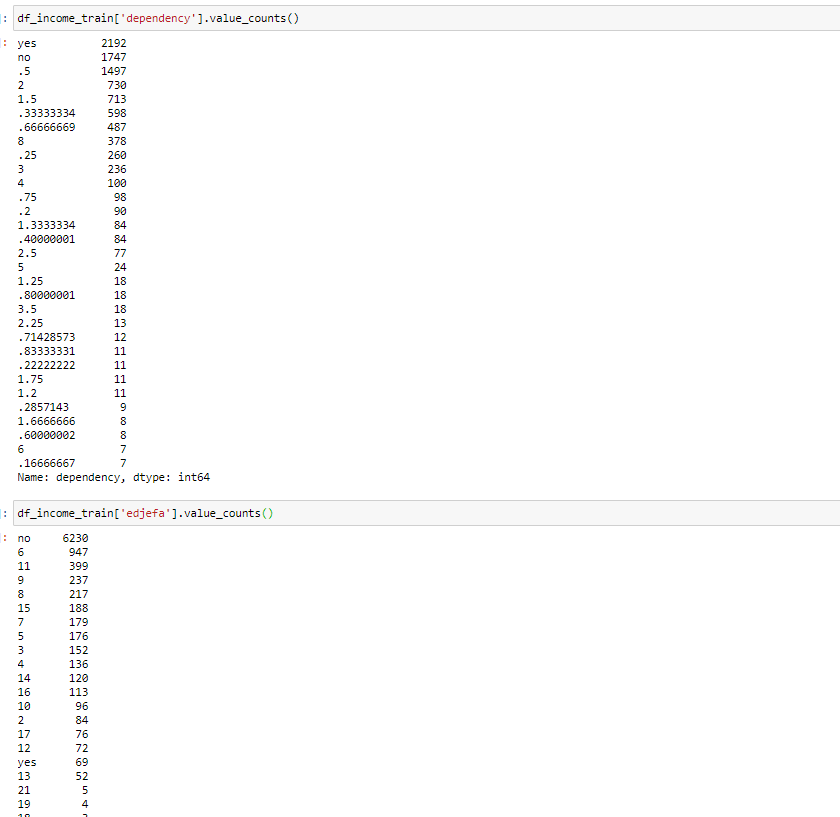


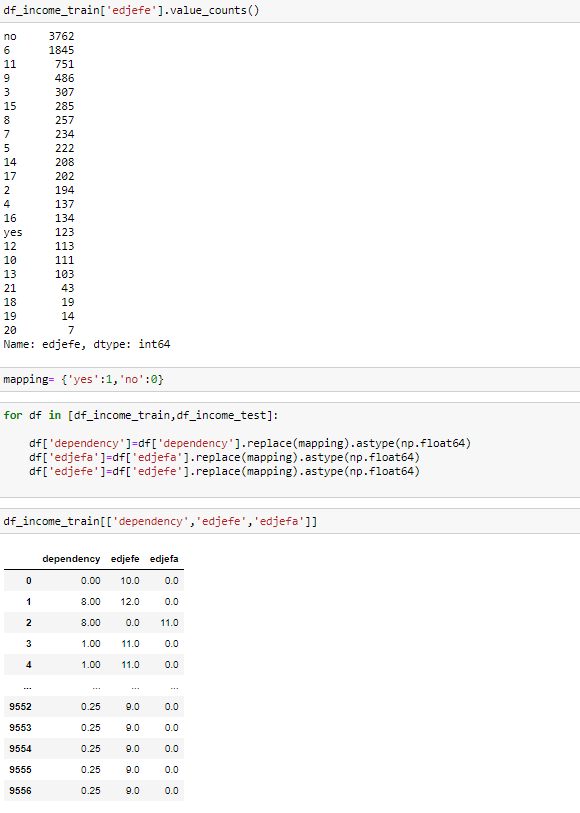
Datatypes in the dataset – int64,float64 and object

Data Cleaning



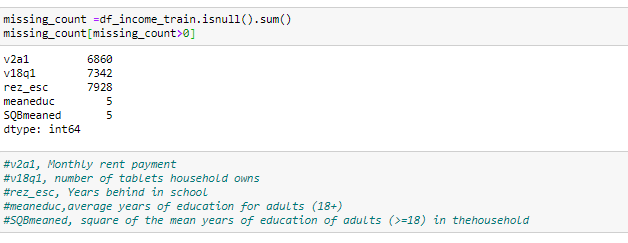
# Mixed values coming for dependency,edjefe,edjefa





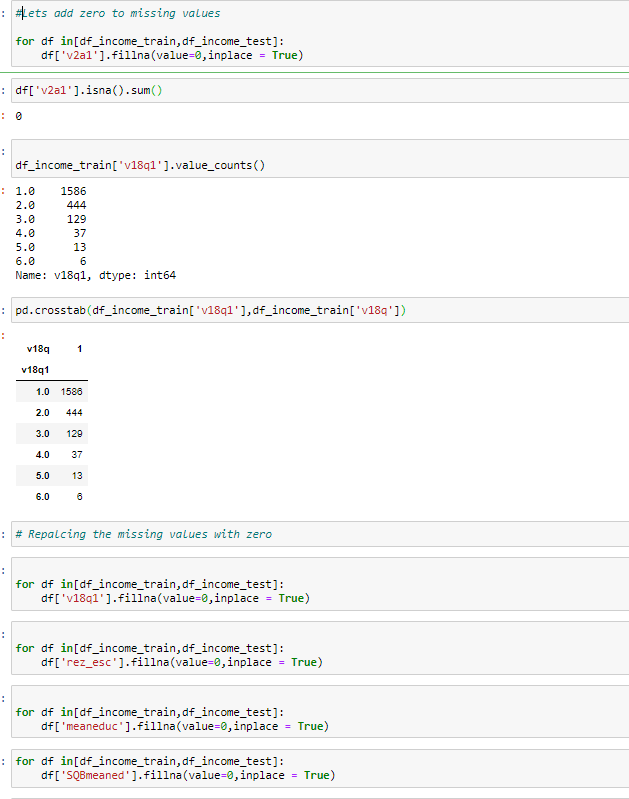
Replaced yes and no values with 1 and 0 respectively.

Objective 7) Count how many null values are existing in columns.



Missing values present for v2a1,v18q1,rez\_esc,meaneduc,SQBmeaned



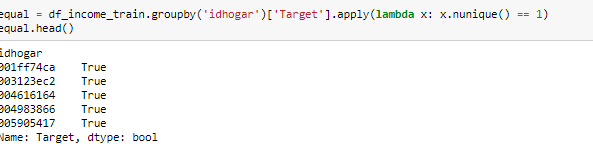


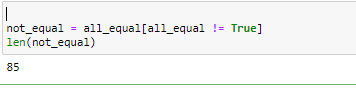
Objective 8) Remove null value rows of the Target variable.



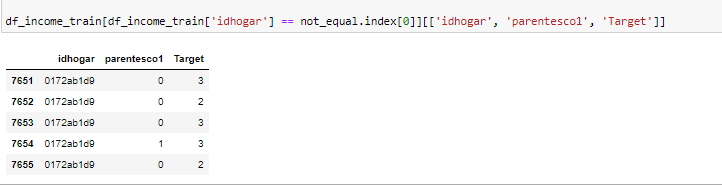
No missing values present for Target variable

Objective 4)Check whether all members of the house have the same poverty level



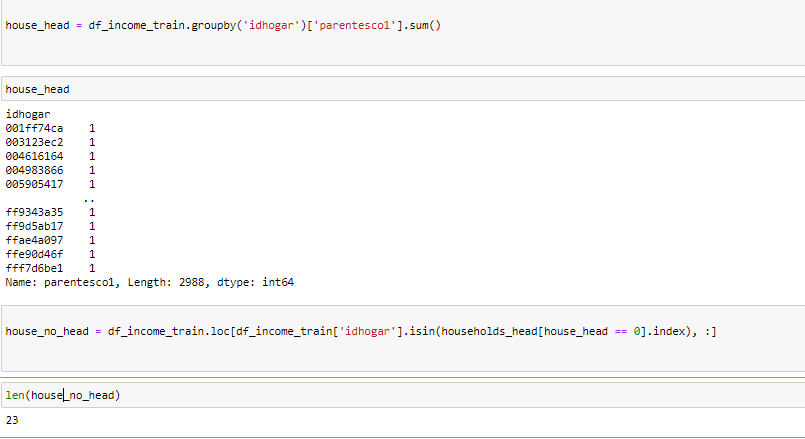


There are 85 houses where all members dont have the same poverty level



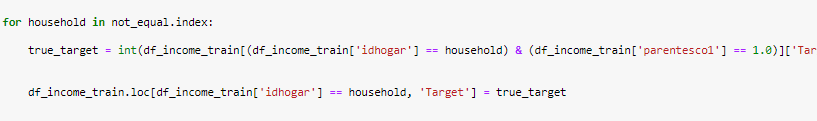
Poverty level are different for the members in a house

Objective 5)Check if there is a house without a family head



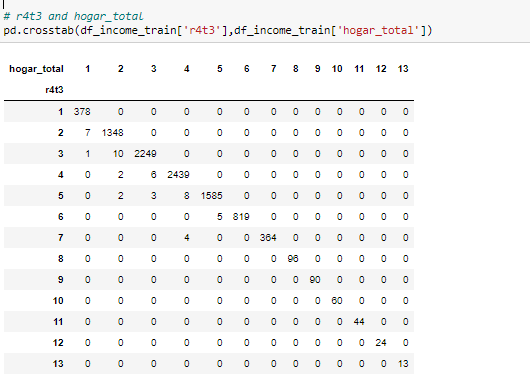
1. houses without family head

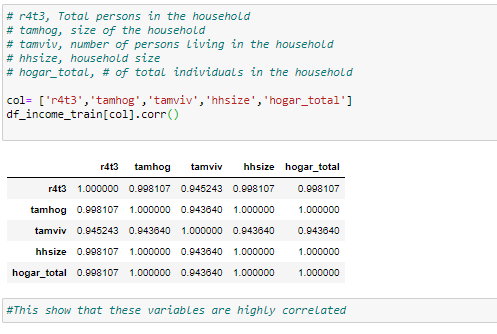
Objective 6)Set poverty level of the members and the head of the house within a family.



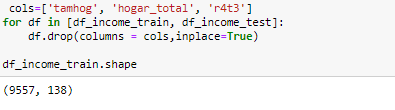
Objective 3)Check if there are any biases in your dataset

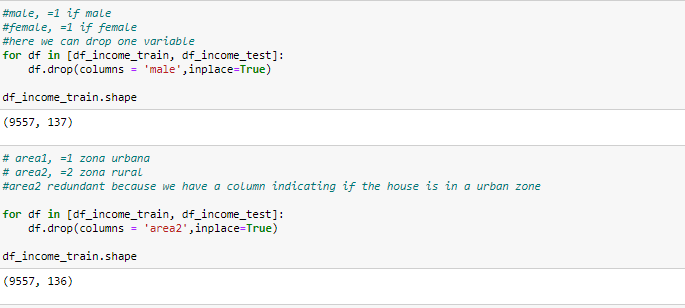






Variables are highly correlated which indicates bias in dataset

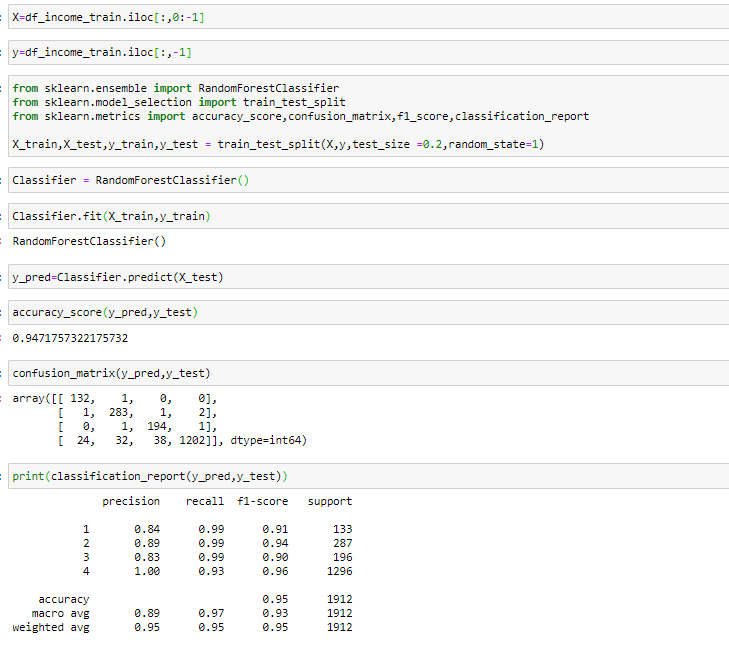




Dropping the squared variables



Objective 9 )Predict the accuracy using random forest classifier.



Model has an accuracy of 94.71 %

Objective 10)Check the accuracy using random forest with cross validation.



Model accuracy for Random Forest Classifier with cross validation is 94.6 %

Result

Analysis Task

* Objective 1 - Identify the output variable.

It is observed that the Test Dataset is not having a target variable.Only Training Dataset is having Target variable (Refer Page 4)

* Objective 2 - Understand the type of data.

Datatypes in the dataset – int64,float64 and object (Refer page 5)

* Objective 3 - Check if there are any biases in your dataset There is bias present in the dataset with some variables highly correlated (Refer pages 13-15)
* Objective 4 - Check whether all members of the house have the same poverty level.

There are 85 houses where all members dont have the same poverty level. Poverty level are different for the members in a house (Refer page 12)

* Objective 5 - Check if there is a house without a family head

There are 23 houses without a family head. ( Refer page 13)

* Objective 6 -Set poverty level of the members and the head of the house within a family.

The poverty level of the members and the head of the house within a family was set (Refer page 13)

* Objective 7 - Count how many null values are existing in columns.

Missing values present for v2a1,v18q1,rez\_esc,meaneduc,SQBmeaned (Refer pages 9-11)

* Objective 8 -Remove null value rows of the target variable.

No missing values present for Target variable ( Refer page 12)

* Objective 9 - Predict the accuracy using random forest classifier.

Model has an accuracy of 94.71 %( Refer page 16)

* Objective 10 - Check the accuracy using random forest with cross validation

Model accuracy for Random Forest Classifier with cross validation is 94.6 %(Refer page 17)