# A Statistical approach to adult census income level prediction

Probability and Statistics

# Flow of the Presentation

1
Introduction

2
Distributions and Observations

3
Statistical Data Analysis

4
ML Models Implementation

5
References

# Introduction

The prominent inequality of wealth and income prevalent in the United States is a huge concern for the government. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improve the economic stability of a nation.. This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem.

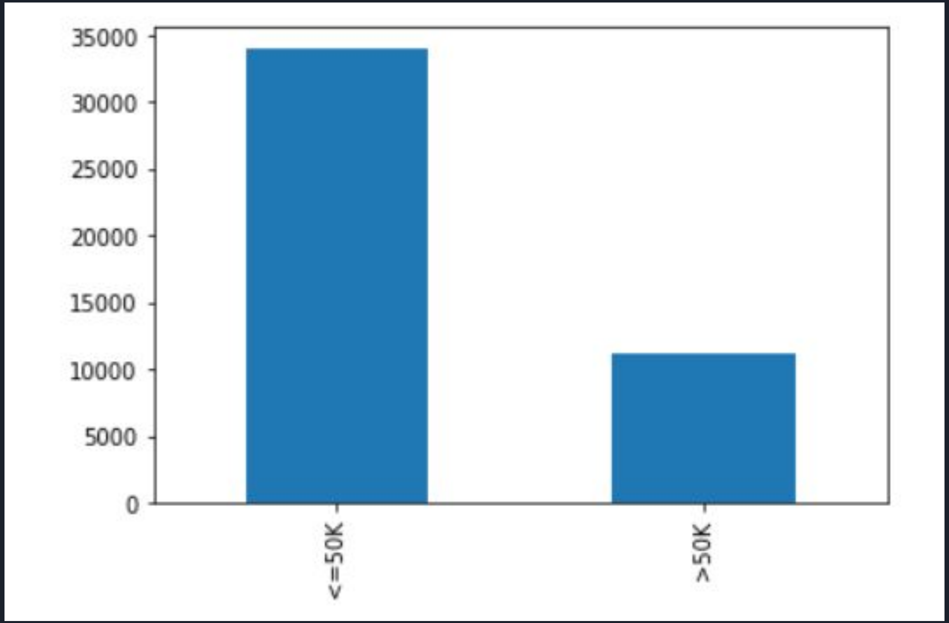
# Insights

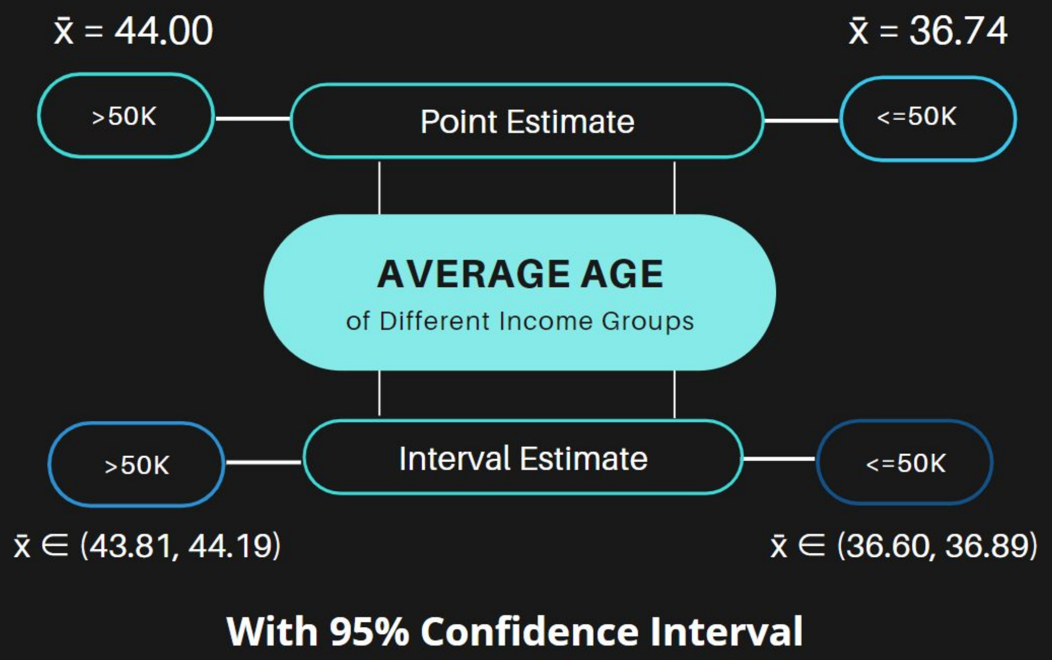
The UCI Adult Dataset has been used for the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K Dollars or less equal to 50K Dollars category based on a certain set of attributes

# Attributes included

| | age | workclass | fnlwgt | education | educational-num | marital-status | occupation | relationship | race | gender | capital-gain | capital-loss | hours-per-week | native-country | income |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 25 | Private | 226802 | 11th | 7 | Never-married | Machine-op-inspct | Own-child | Black | Male | 0 | 0 | 40 | United-States | <=50K |
| **1** | 38 | Private | 89814 | HS-grad | 9 | Married-civ-spouse | Farming-fishing | Husband | White | Male | 0 | 0 | 50 | United-States | <=50K |
| **2** | 28 | Local-gov | 336951 | Assoc-acdm | 12 | Married-civ-spouse | Protective-serv | Husband | White | Male | 0 | 0 | 40 | United-States | >50K |
| **3** | 44 | Private | 160323 | Some-college | 10 | Married-civ-spouse | Machine-op-inspct | Husband | Black | Male | 7688 | 0 | 40 | United-States | >50K |
| **4** | 18 | ? | 103497 | Some-college | 10 | Never-married | ? | Own-child | White | Female | 0 | 0 | 30 | United-States | <=50K |

# Distribution of people's income




$\bar{x} = 44.00$
$\bar{x} = 36.74$
>50K
Point Estimate
<=50K
AVERAGE AGE
of Different Income Groups
>50K
Interval Estimate
<=50K
$\bar{x} \in (43.81, 44.19)$
$\bar{x} \in (36.60, 36.89)$
With 95% Confidence Interval

# Income V/S Other Categorical Features

Two columns are independent

H0

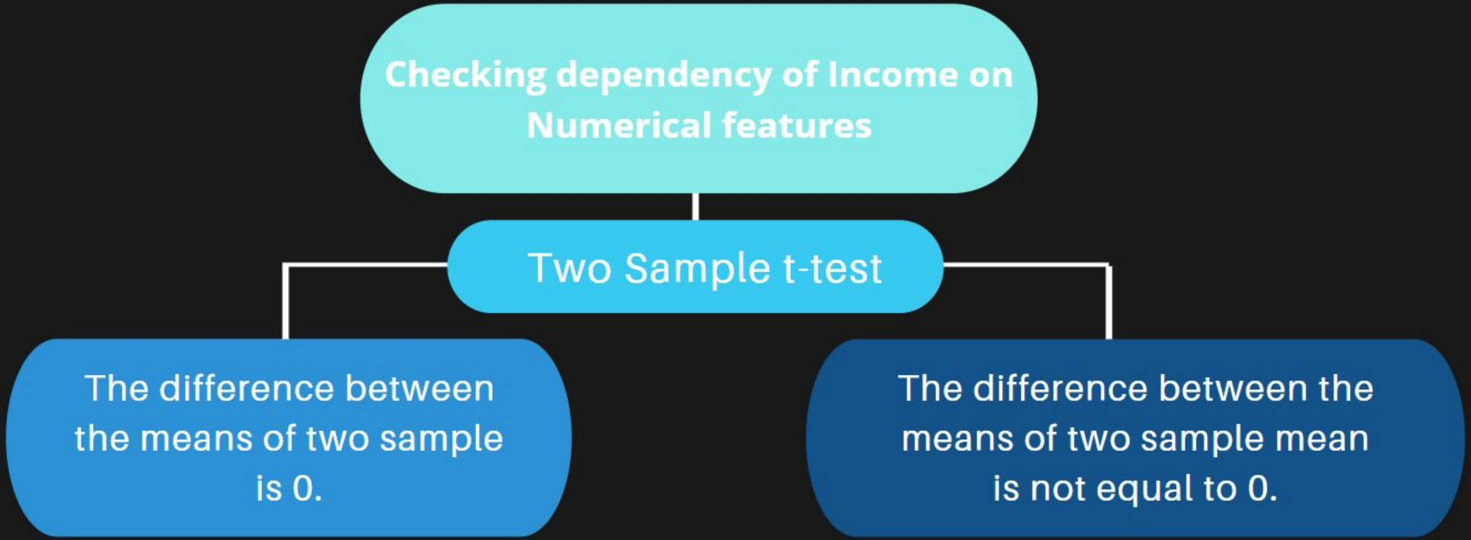
Chi Square

Test

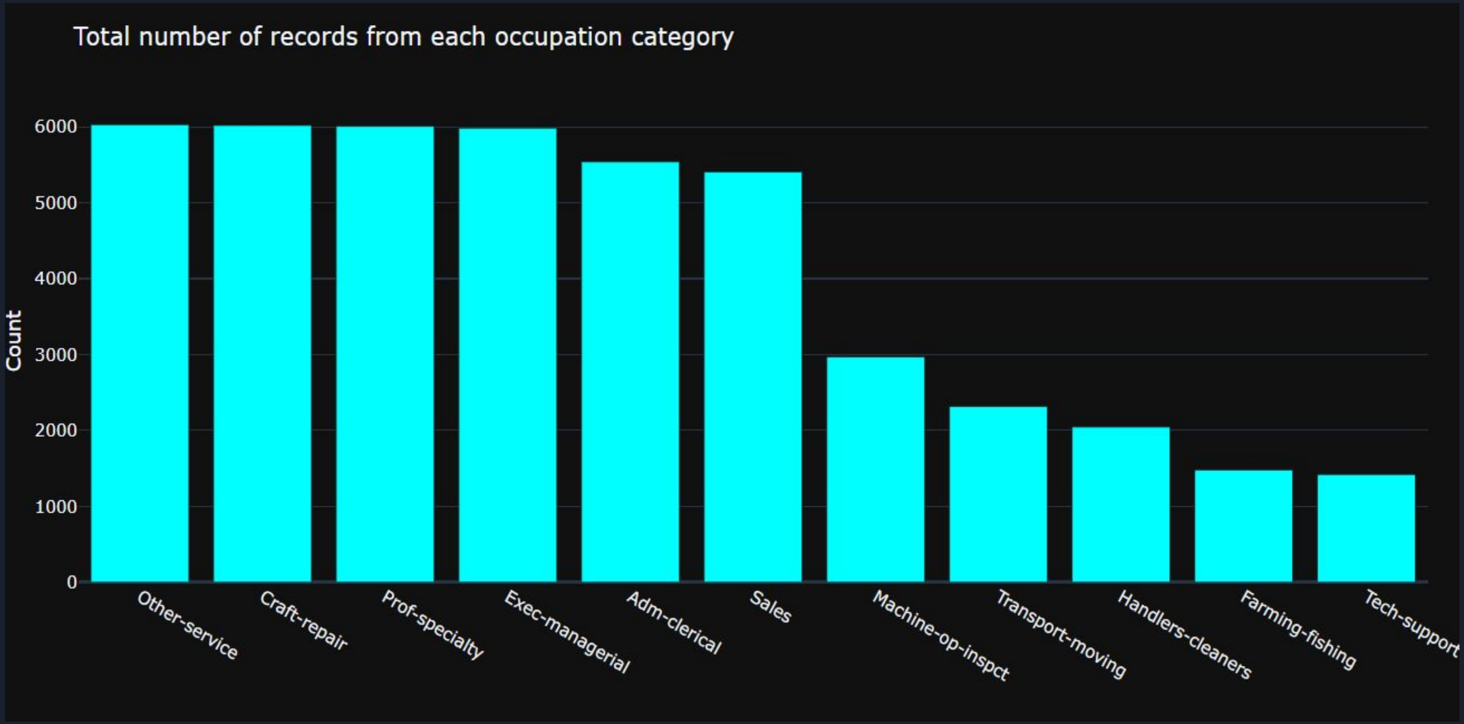
Two columns are dependent

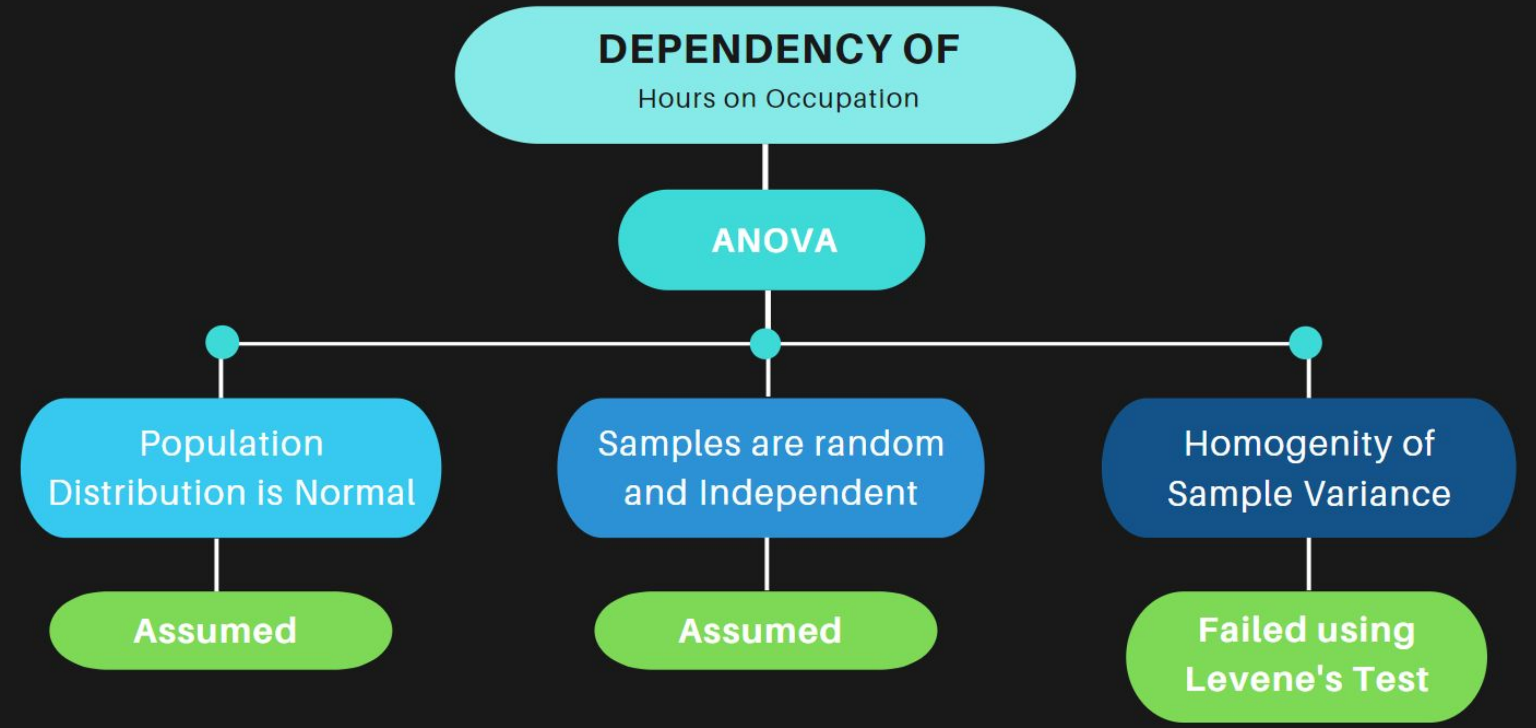
H1

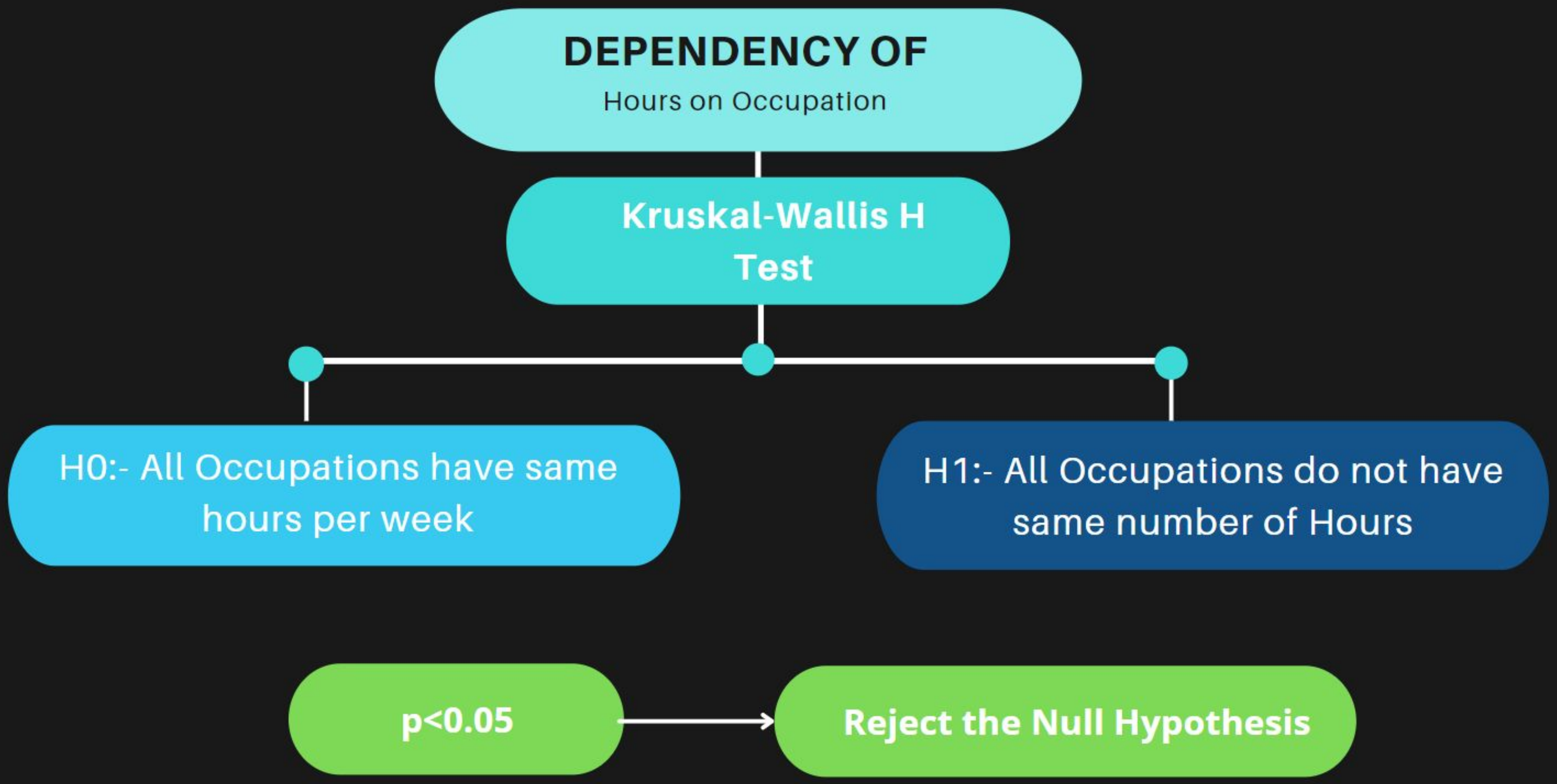
```
{'workclass': [7.502170429307795e-145, 'Reject H0'],
 'education': [0.0, 'Reject H0'],
 'marital-status': [0.0, 'Reject H0'],
 'occupation': [0.0, 'Reject H0'],
 'relationship': [0.0, 'Reject H0'],
 'race': [1.6578176146497785e-84, 'Reject H0'],
 'gender': [0.0, 'Reject H0'],
 'native-country': [1.5099071570211737e-16, 'Reject H0'],
 'income': [0.0, 'Reject H0']}
```




```
{'age': [0.0, 'Reject H0'],
 'fnlwgt': [0.1224230703562435, 'Fails to Reject H0'],
 'educational-num': [0.0, 'Reject H0'],
 'capital-gain': [0.0, 'Reject H0'],
 'capital-loss': [7.594015921975305e-222, 'Reject H0'],
 'hours-per-week': [0.0, 'Reject H0']}
```


Total number of records from each occupation category
Count
6000
5000
4000
3000
2000
1000
0
Other-service
Craft-repair
Prof-specialty
Exec-managerial
Adm-clerical
Sales
Machine-op-inspct
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support


DEPENDENCY OF
Hours on Occupation
ANOVA
Population Distribution is Normal
Samples are random and Independent
Homogenity of Sample Variance
Assumed
Assumed
Failed using Levene's Test


DEPENDENCY OF
Hours on Occupation
Kruskal-Wallis H Test
H0:- All Occupations have same hours per week
H1:- All Occupations do not have same number of Hours
p<0.05
Reject the Null Hypothesis

# Accuracy and F1 score

Following are the f1 scores we got for each one of them:

XGBoost: 0.6943

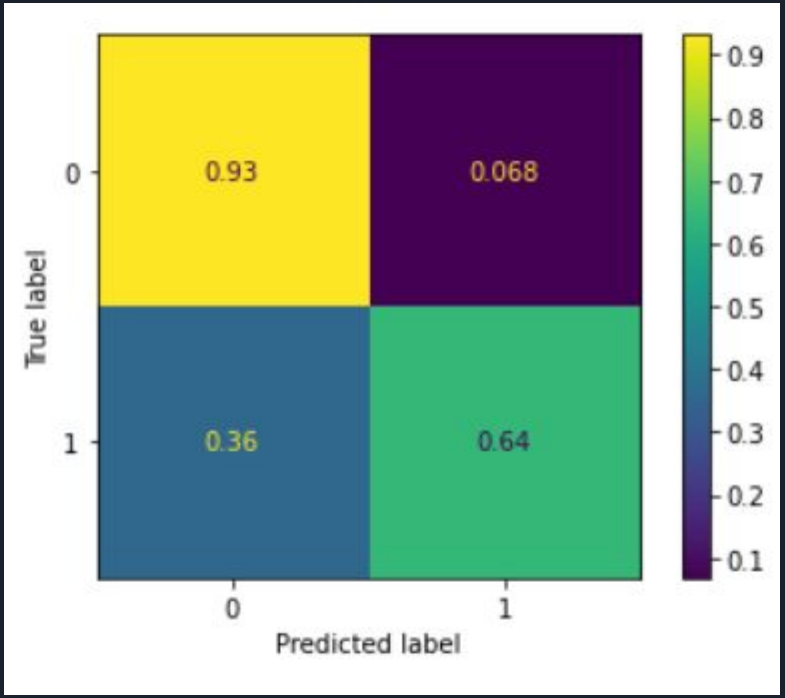
Light GBM: 0.7011

Cat Boost: 0.6894

Adaboost:- 0.7042

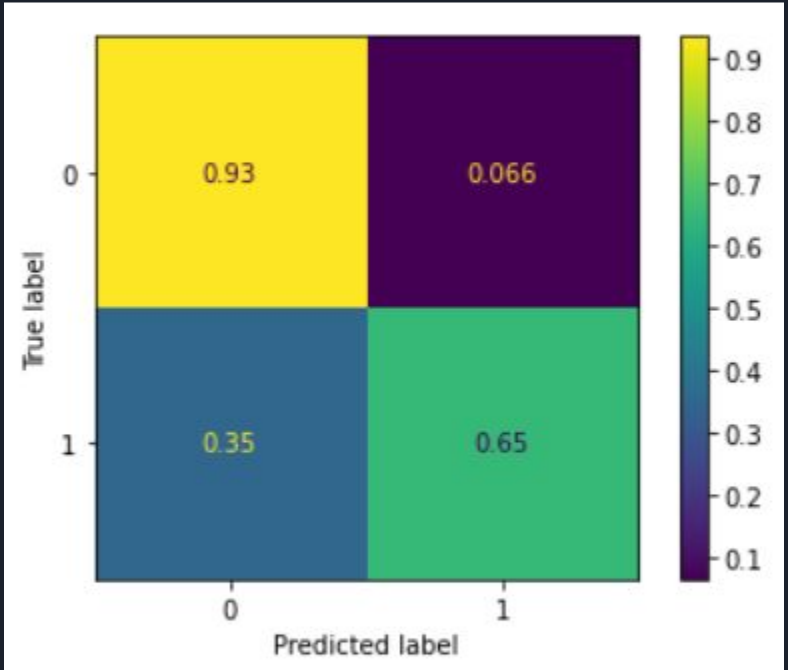
Stacking classifier: 0.6988

# XGBoost

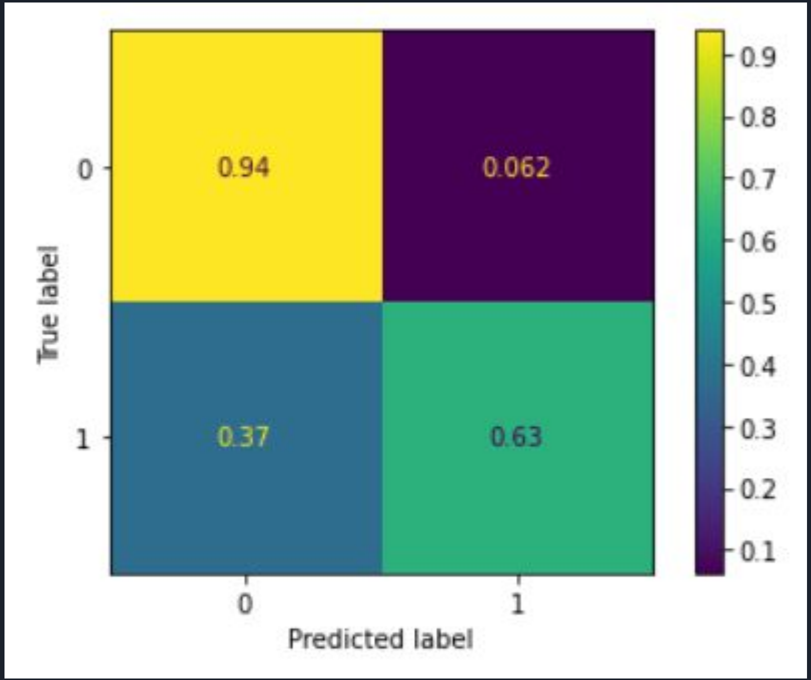


0.93
0.068
0.36
0.64
True label
Predicted label
0
1
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

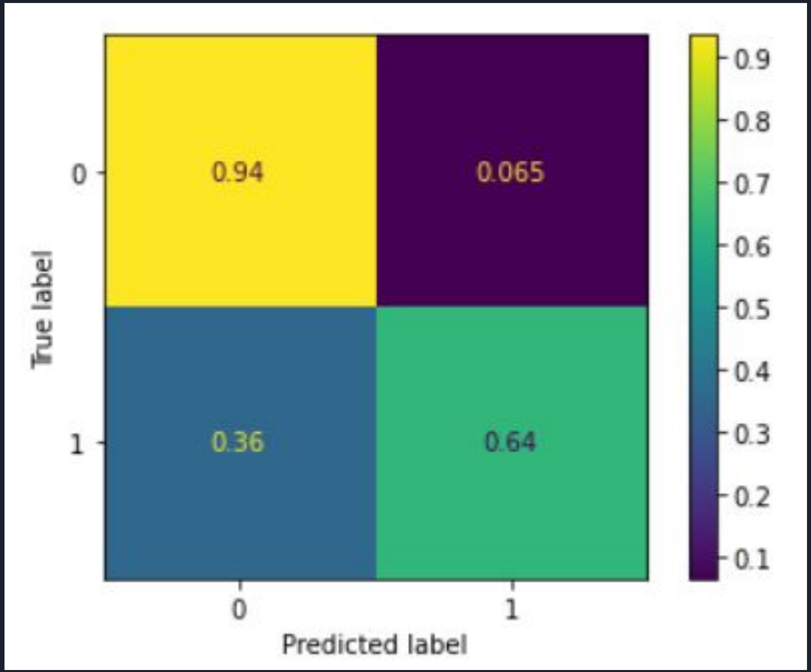
# Light GBM


0.93
0.066
0.35
0.65
0
1
True label
0
1
Predicted label
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

# Catboost


0.94
0.062
0.37
0.63
0
1
True label
Predicted label
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

# Stacking classifier



# Conclusion

Here, we have used several ensemble methods to increase the efficiency of the model, and got the highest efficiency of approximately 86.2% for the stacking classifier, which is very close to what is achieved in the original paper.

# References


- A Statistical Approach to Adult Census Income Level Prediction
- A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set
- XGBoost: A Scalable Tree Boosting System
- CatBoost: gradient boosting with categorical features support
- LightGBM: A Highly Efficient Gradient Boosting Decision Tree
- Dataset Source

# Thank you

This project is presented by:

Amulya Tiwari (2K19/ME/032)

Gaurav Kumar (2K19/ME/090)