Institute for Advanced Computing and Software Development

CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING

DEPARTMENT OF e-DBDA

Academic Year MAY 2021

Project Topic

A GENDER PAY GAP ANALYSIS

Project Member:
1311 Bhagyashri Gawas
1341 Prachi Pande

Mr. Prashant Karhale
Center coordinator

Mr. Akshay Tilekar
Project Guide

Contents

Introduction

- The gender pay gap is a indicator of inequality between women and men.
- The gender pay gap is a global topic.
- The difference in income between the gender with respect occupation and education.
- In our project we are analyzing the wages gap in between gender.

• DATASET INFORMATION

1. Gender pay gap data collected from vincentarelbundock
2. Features are

i. year

ii. Income

iii. age

iv. occcode

v. occupation

vi. prestg10

vii. childs

viii. wrkstat

ix. education

x. maritalstatus

```
[ ]  1 gender_pay.head()
```

	year	realrinc	age	occ10	occrecode	prestg10	childs	wrkstat	gender	educcat	maritalcat
0	1974	4935.0	21.0	5620.0	Office and Administrative Support	25.0	0.0	School	Male	High School	Married
1	1974	43178.0	41.0	2040.0	Professional	66.0	3.0	Full-Time	Male	Bachelor	Married
2	1974	NaN	83.0	NaN	NaN	NaN	2.0	Housekeeper	Female	Less Than High School	Widowed
3	1974	NaN	69.0	NaN	NaN	NaN	2.0	Housekeeper	Female	Less Than High School	Widowed
4	1974	18505.0	58.0	5820.0	Office and Administrative Support	37.0	0.0	Full-Time	Female	High School	Never Married

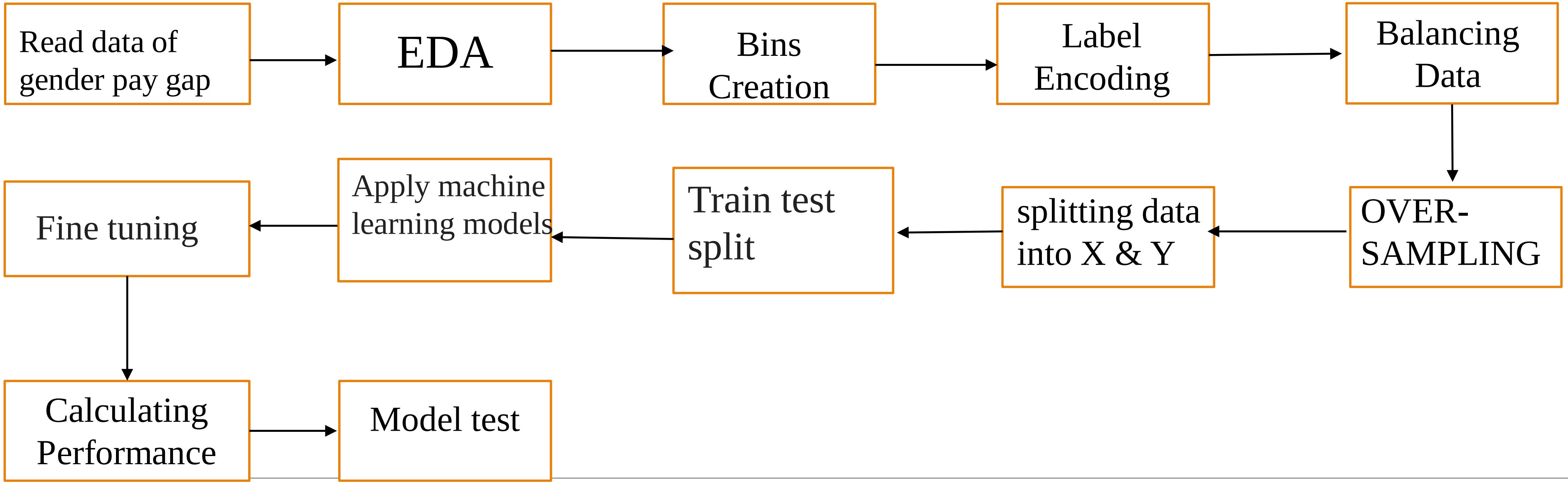
Technologies Used:

- Python
- Machine learning algorithms

Models used in the system:

- Random Forest model
- Adaboost model
- Xgboost model

FLOWCHART OF SYSTEM:



EDA

- Renamed columns name

```
[ ]  1 gender_pay=gender_pay.rename(columns={"realrinc": "Income", "occ10": "occCode", "occrecode": "occupation", "educcat": "education","maritalcat": "maritalstat" })
```

- Removed null values from target column

```
[ ]  1 gender_pay = gender_pay.dropna(subset=['Income'])
```

- Removed null values from age columns with respect to childs and gender columns with median of it.

```
[ ]  1 df['age'].fillna(df.groupby(['childs','gender'])['age']. transform ('median'),inplace=True)
```

- Removed null values from child columns with respect to age and gender columns with median of it.

```
[ ]  1 df['childs'].fillna(df.groupby(['age','gender'])['childs']. transform ('median'),inplace=True)
```

- Changed the datatypes of features

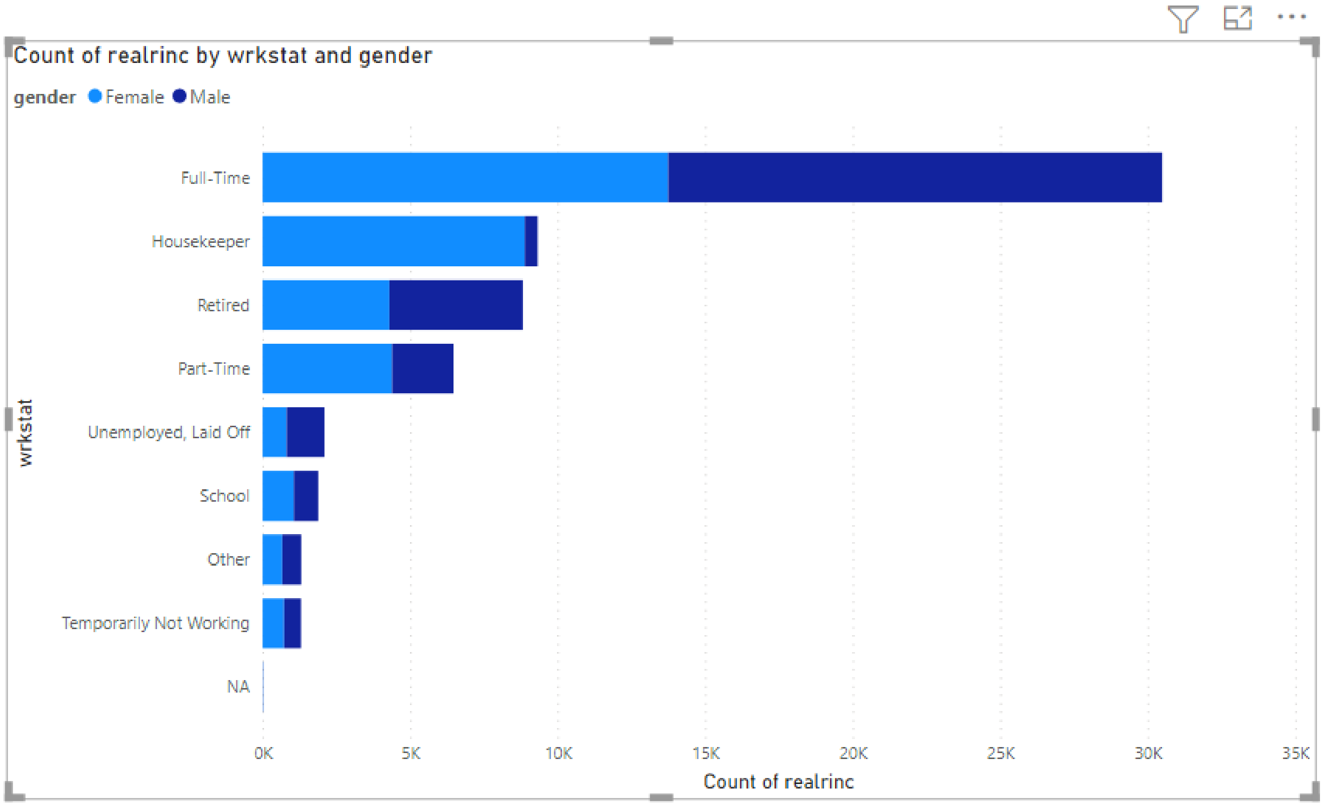
```python
convert_dict = {'age': int,
                'occCode': int,
                'prestg10': int,
                'childs':int

                }

df =df.astype(convert_dict)
```

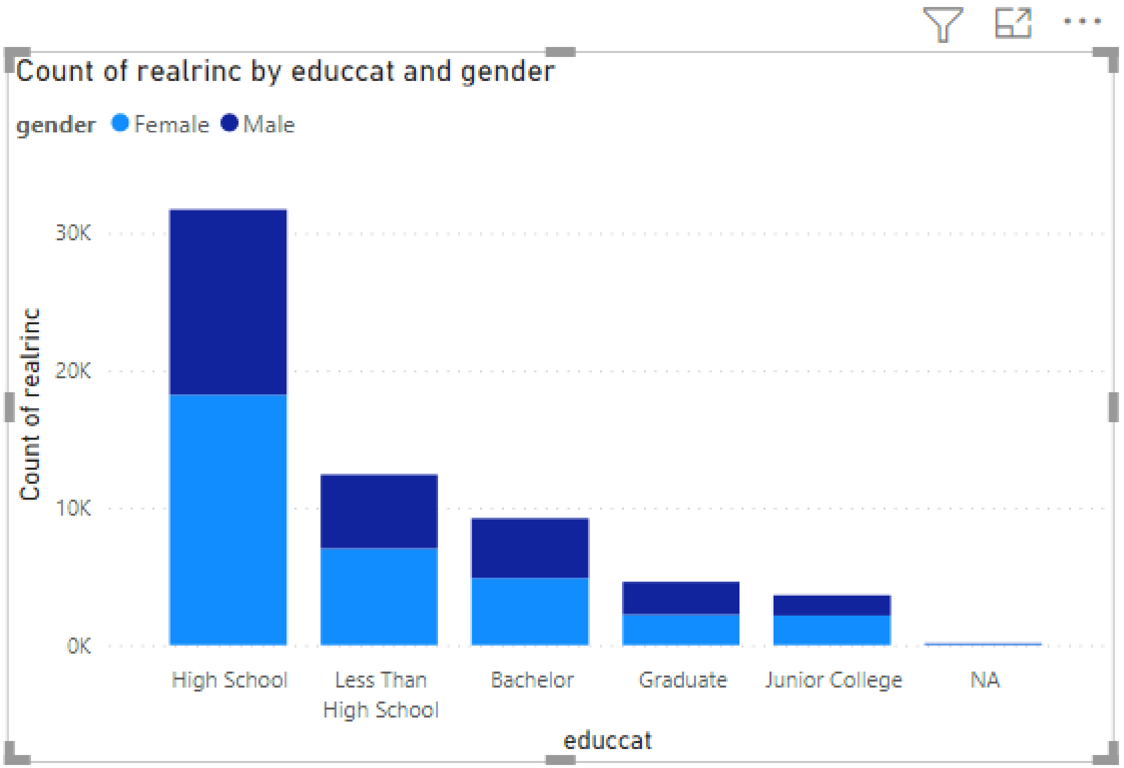
DATA VISUALIZATION

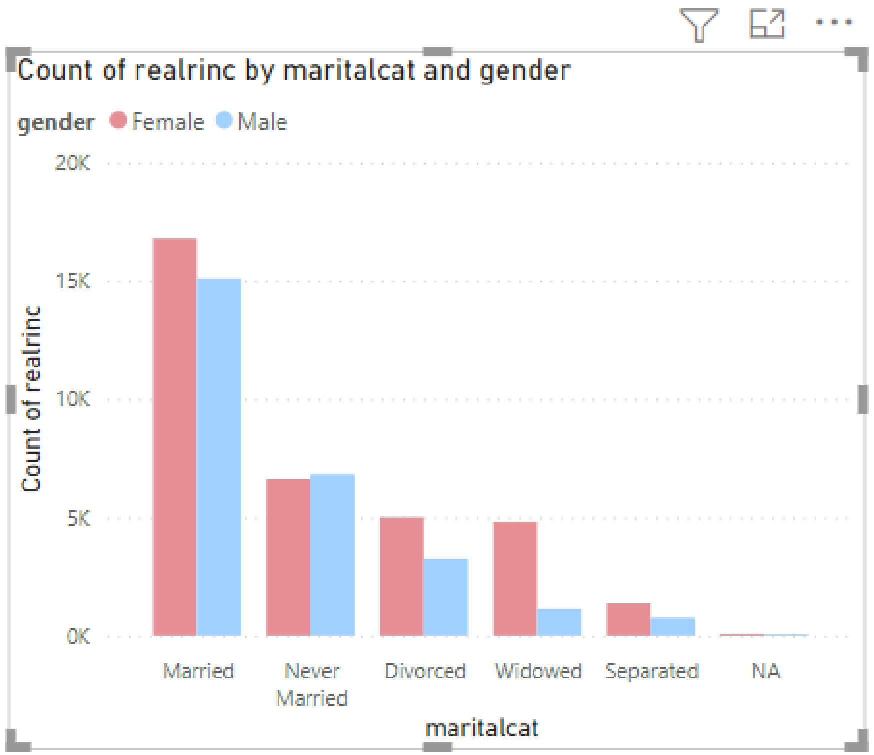
Gender wise distribution of Income vs work status



Genderwise income distribution with education

Marital status vs Income





DATA PROCESSING

- Bins Creation
- Label Encoding
- Split X & Y
- Oversampling
- Train test split

Model Building

- We applied three ensemble models.
- Random forest, Adaboost, Xgboost.

Random Forest: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Adaboost : It is boosting technique. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

Xgboost: Xgboost stands for Extreme Gradient Boosting.It implements machine learning algorithums under the gradient boosting framework.

- apply fine tuning for increasing accuracy of model.

Model results

Random Forest

```
<class 'sklearn.ensemble._forest.RandomForestClassifier'>
Confusion Matrix
 [[2241  780  269  147   26]
 [ 812 1543  732  341   55]
 [ 260  641 1778  614  110]
 [ 105  287  593 2168  279]
 [  12   14   43   92 3337]]
Accuracy :  0.640488454192951
```

XGBoost and AdaBoost

```
<class 'sklearn.ensemble._gb.GradientBoostingClassifier'>
Confusion Matrix
 [[2272  778  236  106   71]
 [ 773 1598  692  305  115]
 [ 257  751 1407  708  280]
 [ 101  295  795 1466  775]
 [  26   58  166  328 2920]]
Accuracy :  0.559233751953238
<class 'sklearn.ensemble._weight_boosting.AdaBoostClassifier'>
Confusion Matrix
 [[2143  845  274  124   77]
 [ 791 1454  756  334  148]
 [ 295  711 1229  847  321]
 [ 149  301  791 1360  831]
 [  39   58  154  581 2666]]
Accuracy :  0.5122981654030905
```

Testing Result

```
Sample = (1974,21,5620,5,25,0,5,0,1,1)
Test = np.array(Sample)
Test=Test.reshape(1,-1)
Result= rfc.predict(Test)
Result
```

```
(10,)
(1, 10)
array([0])
```

Conclusion

After passing the data elements we got the predicted value as similar to the original test dataframe

Application

- To compare the wages gap between the occupation with respect to education of male and female.
- To compare the wages gap between the age with respect to work status of male and female.
- Year wise changes of income between male and female

Conclusion

- We have seen that there is gender pay gap in data of United States from 1974 to 2018.
- This is due to the gender difference in education, as well as the substantial reduction in the Income.
- There is a gender pay gap with respect to education and occupation in Income.

Future Scope

- This project can be used to determine gender pay gap and how it varies by year-to-year.
- As gender pay gap is the global issue so it will help to determine what features are highly making impact and what can be done to overcome that features.
- It will help to provide equality for genders income with respect to all working sectors

THANK YOU