|  |  |
| --- | --- |
| Internship Project Title | RIO-125: HR Salary Dashboard - Train the Dataset and Predict Salary |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 10/04/2022 | 09/05/2022 | 130 | Jupiter Notebook | Python,Numpy,Pandas,plotly,streamlit |

**TABLE OF CONTENTS**

* Acknowledgements
* Objective
* Introduction
* Internship Activities
* Methodology
* Assumptions
* Process of preparing model
* Dataset collection
* Basic Understanding of the dataset
* Exploratory data analysis
* Preprocessing
* Modelling
* Logistic Regression
* Random Forest
* Decision Tree
* Gradient Boosting
* XgBoost
* Hyperparameter tunning
* Conclusion
* Limitations
* Link to code and executable file

**Acknowledgements**

I express my deep sense of gratitude to all my mentors from ICT academy for their continues support and guidance throughout this journey. Thanks to [TCS iON](https://www.linkedin.com/company/tcs-ion/) for this wonderful opportunity and support throughout the process. I am grateful to my industrial mentor from TCS iON for his expert advice and suggestions during this journey. I am grateful to my parents, well-wishers and friends for their moral support and help.

**OBJECTIVE**

The main aim of this project is finding the right future salary of an applicant based on some parameter of a particular domain.

**INTRODUCTION**

It is imperative to provide an unbiased salary for an employee which he/she truly deserves, and also has to be appropriate to the market demands. An individual’s annual income results from various factors. This project put an emphasis on predicting the employee’s new salary based on several factors such as age, work class, education, experience, previous occupation, past income, working hours-per-week etc. The dataset contains 15 columns. The target column is income which is divided into two classes: <=50K and >50K.Based on the above dataset parameters, salary can be predicted accurately. We can also run few machine learning algorithms(supervised) to predict the income of several group of people selected for study and notice which factors contribute the most towards the predictions.

**INTERNSHIP ACTIVITIES**

1.Understanding the problem statement and finding the dataset

2.Studied the project reference materials and watched webinars.

3.Daily activity report submission

4.Visited digital discussion room for updates and discussions.

5.Performed necessary actions (data understanding, EDA, preprocessing, modelling and hyper parameter tunning) for getting desired output from dataset.

6.Tried to create a simple dashboard using streamlit

**METHODOLOGY**

The methodology follows the best practices in the literature and in the industry, including different phases:

**1.Data collection**

In this project the supervised algorithm is chosen and a dataset consisting of about 48842 records and 15 attributes are downloaded from Kaggle. This dataset was chosen, as the number of records, attributes are large enough to build an efficient model and the prediction has a great scope. It is in CSV format.

**2. Dataset Understanding**

Statistical tools and simple exploratory data analysis techniques are used in order to provide a preliminary and compact description of the data.

**3. Data cleaning**

The missing values are removed and possible conflicts in the data format (e.g. text encoding) are fixed.

**4. Manual feature engineering**

Irrelevant features are discarded and others are standardized (e.g. converted into numerical features) by exploiting the domain knowledge.

**5.Data Modelling**

Machine Learning classifying models are examined by the performance measures which assess how well the algorithms are performing within a given situation.

**6. Model selection**

A grid search is performed to find the optimal hyper-parameters for a set of well-known machine learning models.

**7. Model training and validation**

The selected models are trained and cross-validated in order to find the classifiers that best describe the data and are able to predict the output variable with the highest scores.

**8. Model comparison**

**E**ach model is compared to the others with respect to standard scores and with the classification accuracy, the F1 score, the Precision-Recall etc.

**ASSUMPTIONS**

The features were collected through a random survey where the participants answered the questionnaire honestly and truthfully. Another assumption about the data is that the dataset is free from selection bias.

**PROCESS OF PREPARING MODEL**

The process of predictive analysis begins with importing the necessary Python libraries, and then performing data mining over the dataset which is to be trained using different ML algorithms. Once it is trained, comparison using Classification Report is executed and if a user gives a new input the model has to predict the salary either >50K or <=50K.

**1.DATASET SELECTION**

In this project the supervised algorithm is chosen and a dataset consisting of about 48842 records and 15 attributes are downloaded from Kaggle. This dataset was chosen, as the number of records, attributes are large enough to build an efficient model and the prediction has a great scope. It is in CSV format.

**2.BASIC UNDERSTANDING OF THE DATASET**

The dataset has 48842 observations and 15 categories with a mix of integer and objective type data. There are 52 duplicate values. In three columns there exist the presence of a special character (?) and that is replaced with nan. This missing value columns should be either dropped for replace with suitable values for further analysis. The statistical summary is taken for both integer and objective datatypes and unique values in different columns are identified. When checked for income distribution gender wise its seen that males are earning more than females.Also self employed people is making more money.

**3.DATA VISUALIZATION**

Visualization of data is the mechanism to represent the data in terms of graphs, and charts. The best representation of integer data is via histogram which represents distribution of data.

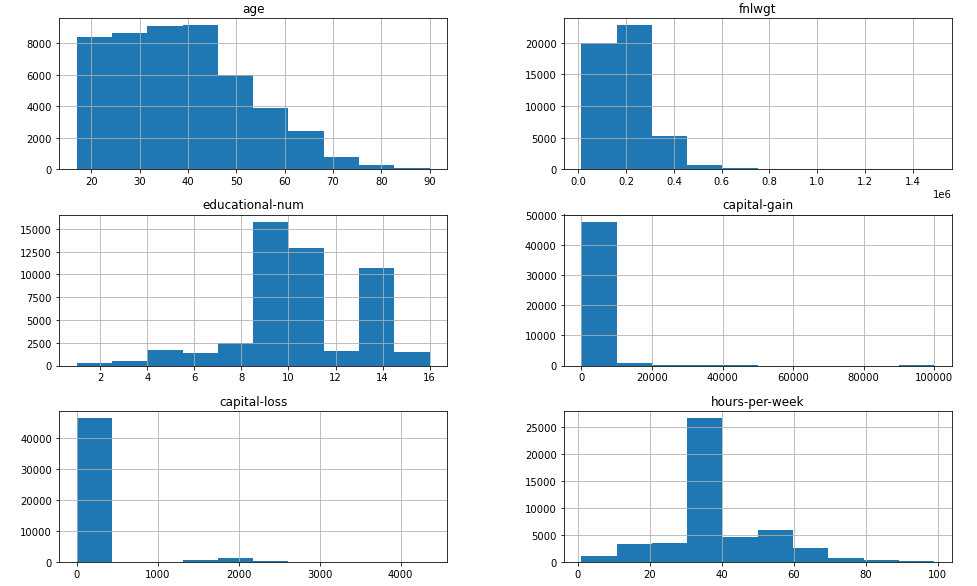


Fig-1 Histogram for integer columns

From histogram it is clear that that most of the people are from age between 17 and 50 and working hours can be seen more between 30 to 40.

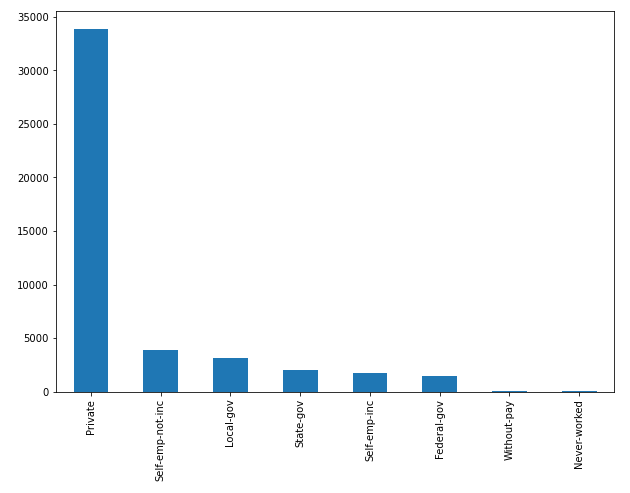
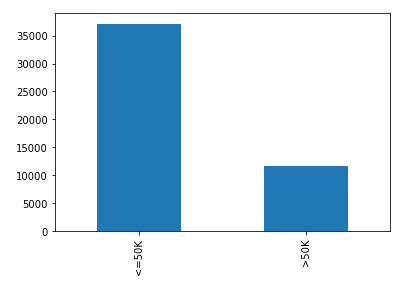
 

Fig-2 Bar chart of working class and income

The parameter work class and income are shown via bar- chart with high accuracy and most of the people is working in private sector is selected for the study and income is less than 50k for most of the people.

The educational background is represented using pie chart and people with HS grad takes maximum percentage.

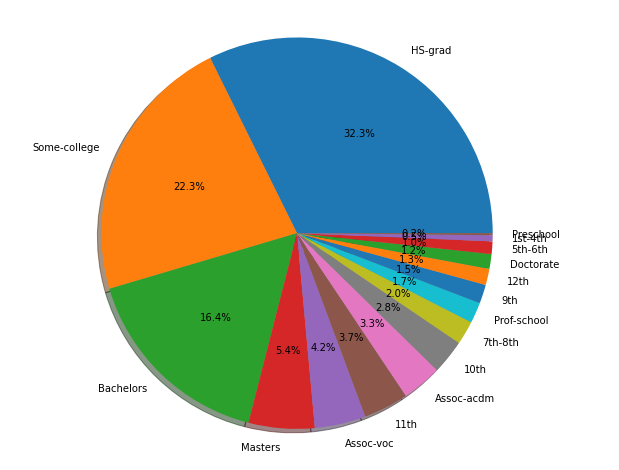


Fig-3 Educational background

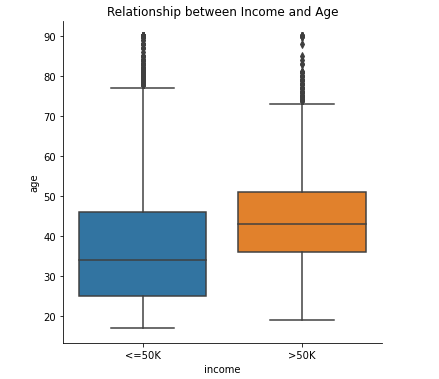


Fig-4 Relationship between income and age

Figure 4 shows the relationship between income and age. Age of people with income less than 50K is between 25 and 45. Age of people with income more than 50K is between 35 and 50 and salary increment is happening with age.

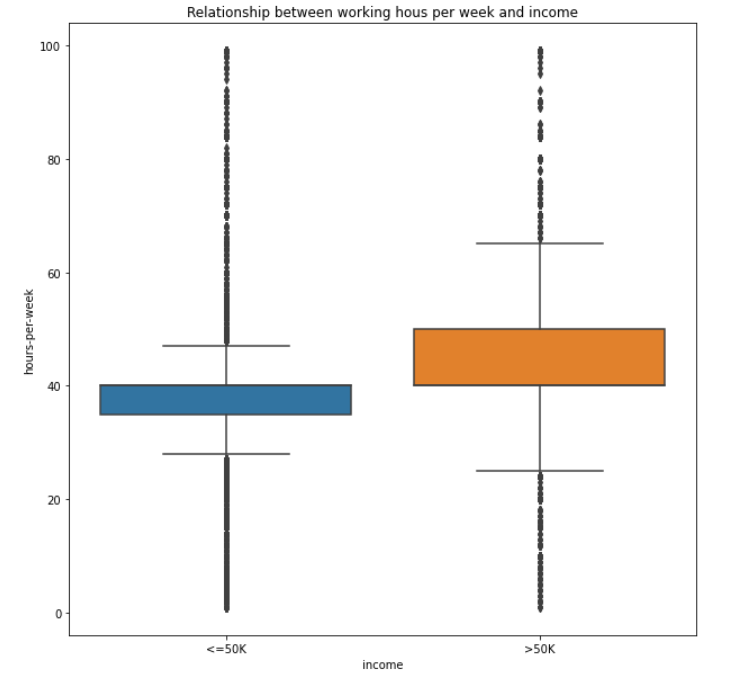


Fig-5 Relationship between hours-per week and income

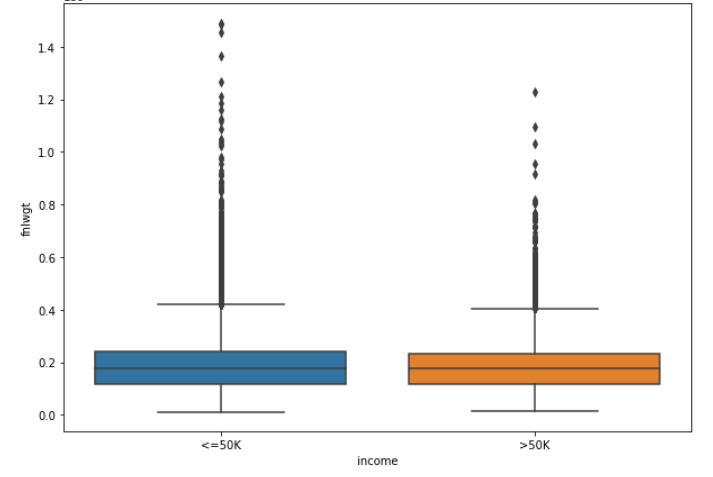


Fig-6 Relationship between final weight and income

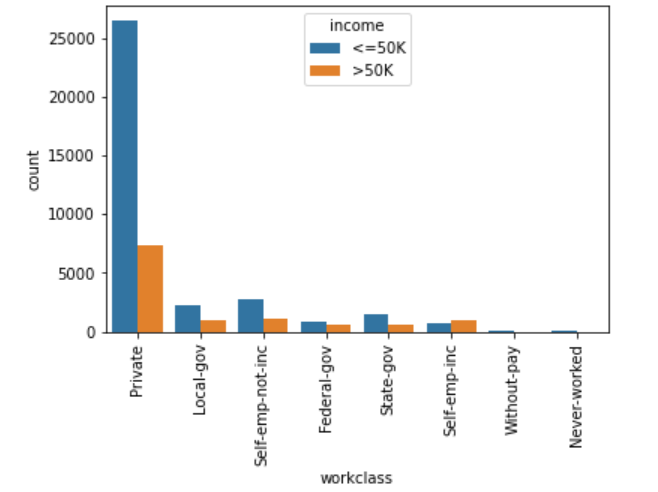


Fig-7 Relationship between working class and income

For all working-class number of people earning less than 50k are more than those earning more than 50k.Data mostly consist of employees from private sector and they are earning more

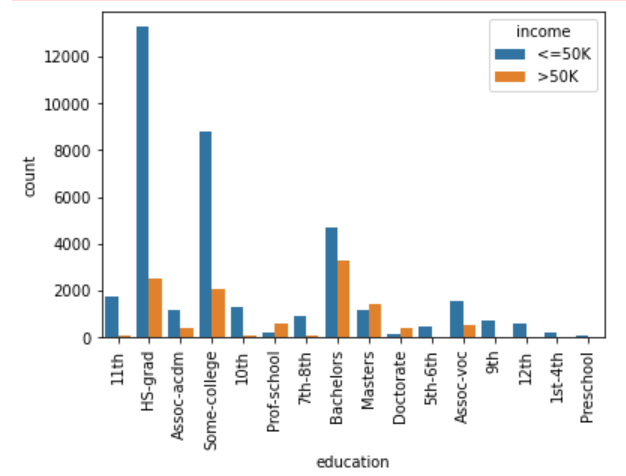


Fig-8 Relationship between education and income

Data mostly consist of people who has education as HS-grad. People with bachelor’s degree is earning more.

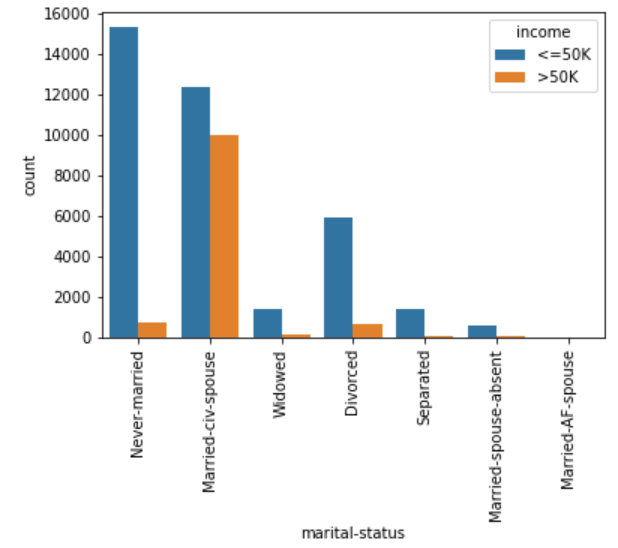


Fig-9 Relationship between marital status and income

The people with marital status as Married-civ-spouse has highest people with income more than 50k.

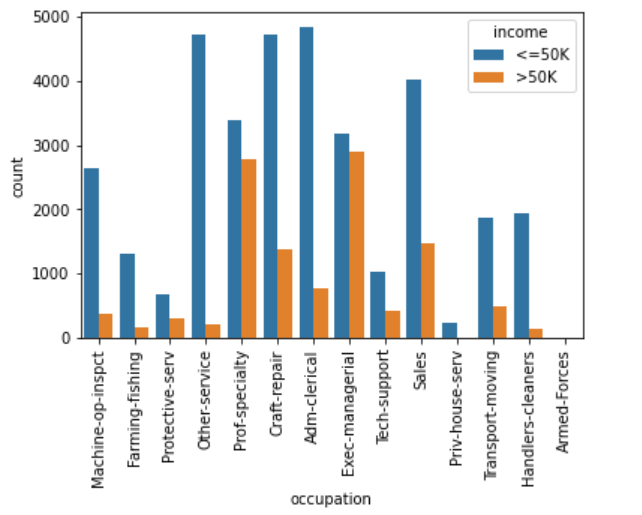


Fig-10 Relationship between occupation and income

Most of the people who have income more than 50k either have prof-speciality or exec-managerial as occupation.

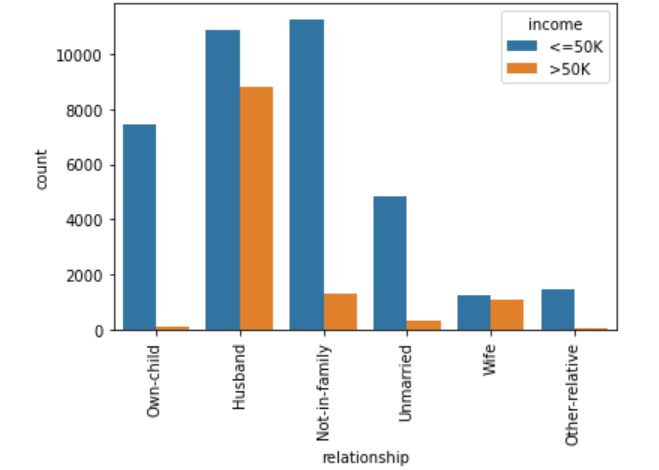


Fig-11 Relationship and income

Mostly a person with relation as husband in a family has most count of people with more than 50k income

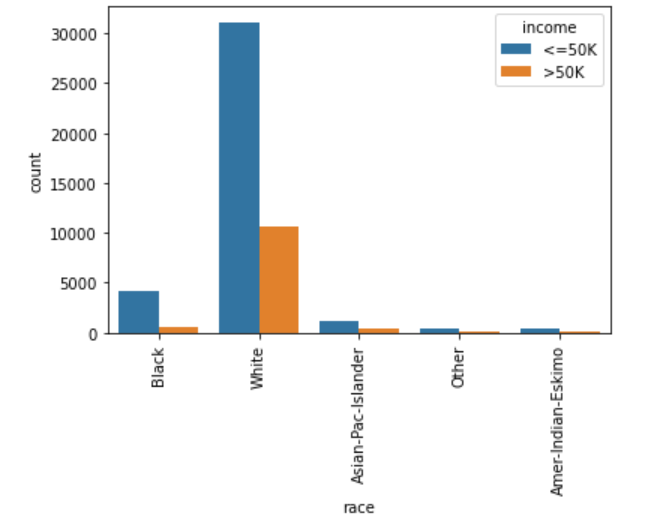


Fig-12 Race and income

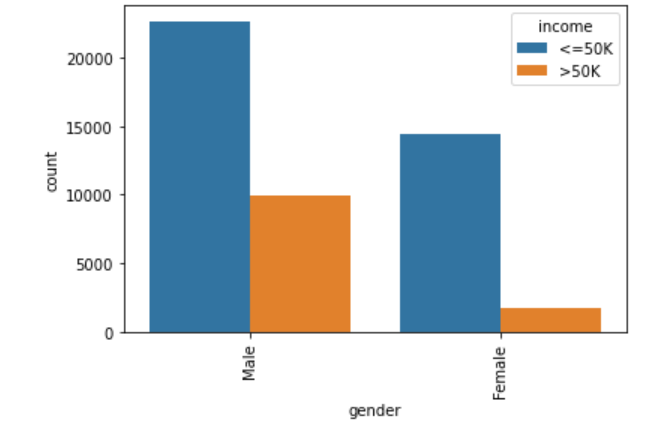


Fig-13 Gender and income

Male white people are earning more

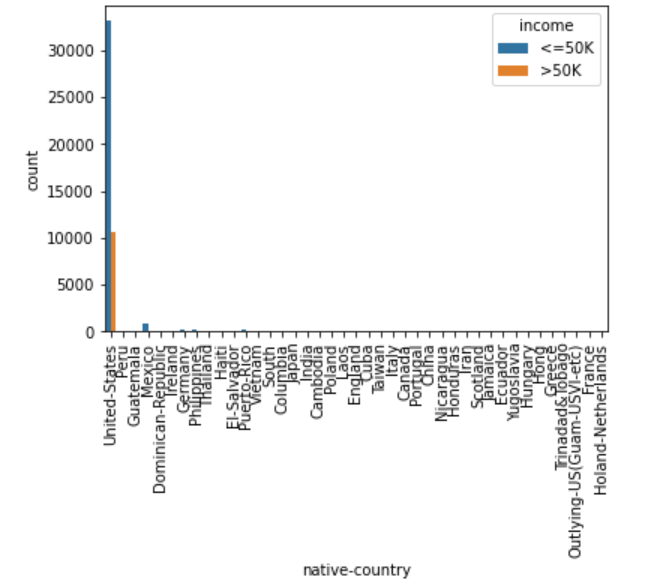


Fig-14 Native country and income

The data mainly consist of people from US.

**4.PREPROCESSING**

It is technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training machine learning models. The quality of the data should be checked before applying machine learning. Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

* **Accuracy**: To check whether the data entered is correct or not.
* **Completeness**: To check whether the data is available or not recorded.
* **Consistency:** To check whether the same data is kept in all the places that do or do not match.
* **Timeliness**: The data should be updated correctly.
* **Believability**: The data should be trustable.
* **Interpretability**: The understandability of the data.

It is observed that some of the categories are not contributing anything to the target and that should be removed. The missing values should be properly replaced without affecting the performance. Different techniques in data preprocessing include,

**MISSING VALUE IMPUTATION**

The columns like workclass, occupation and native country is having missing values. As the missing value fall into the categorical features, we will use mode() method to fill the missing value. Next step is to drop columns which has less significance in predicting the target. The columns like educational-num, capital gain, capital loss and fnlwgt falls in that category and should be removed.

**OUTLIER DETECTION AND REMOVAL**

When checked for outliers using boxplot, column age and hours-per-week has shown outliers.Cappingtechnique is used to remove outliers and in this wecap our outliers data and make the limit i.e, above a particular value or less than that value, all the values will be considered as outliers, and the number of outliers in the dataset gives that capping number.

**ENCODING**

The machine learning models use mathematical equations. So categorical data is not accepted so we convert it into numerical form.

1. Label Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. This technique is applied to different columns like workclass, education, marital status, occupation, relationship and native country.

1. One-Hot Encoding

A one hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.For race and gender column one hot encoding is used.

**5.MODELLING**

The classification model has to be evaluated to analyze real-world programs. Machine Learning classifying models are examined by the performance measures which assess how well the algorithms are performing within a given situation. These performance measures comprise precision, recall score, accuracy, and F1 score. The model’s performance criteria are now becoming a necessity in understanding the strengths and weaknesses of the model while performing new predictions in many cases.

1. **Logistic Regression**

There is a need for building classification models to solve various classification problems thereby finding the main target class within the data sample. Logistic regression is an important algorithm to facilitate organized data analysis. It is part of a supervised machine learning algorithm. This algorithm is based on the sigmoid function and has a conditional value of 0.5. If the input is above 0.5 it belongs to one class and the input value which lies below 0.5 belongs to another class. It is a type of data modeling procedure where the result is the final probability in the discrete outcome based on input values. Binary logistic regression is the most common algorithm consisting of true or false, yes or no values, etc.

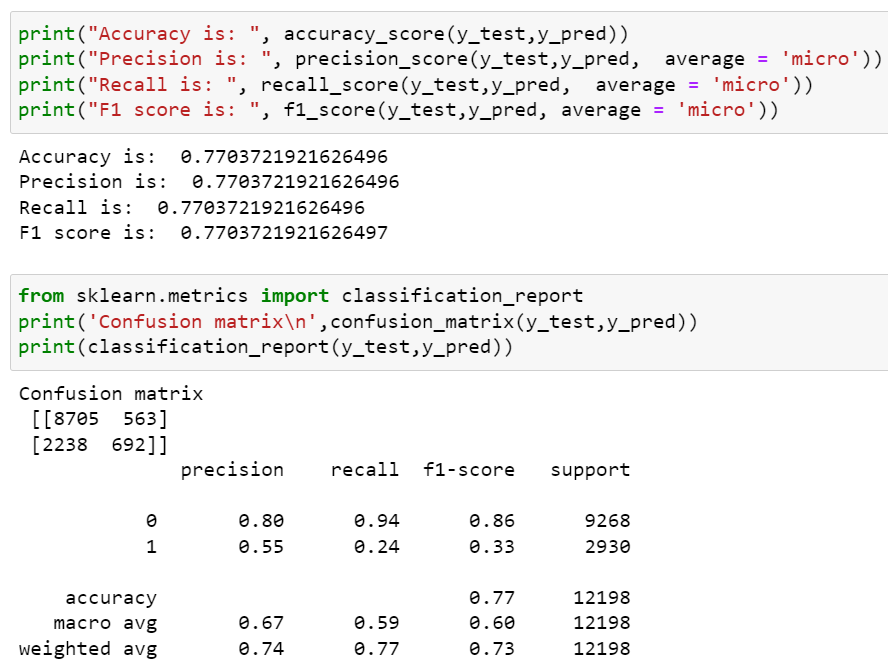


Fig.15 Classification report for logistic regression

1. **Random Forest**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

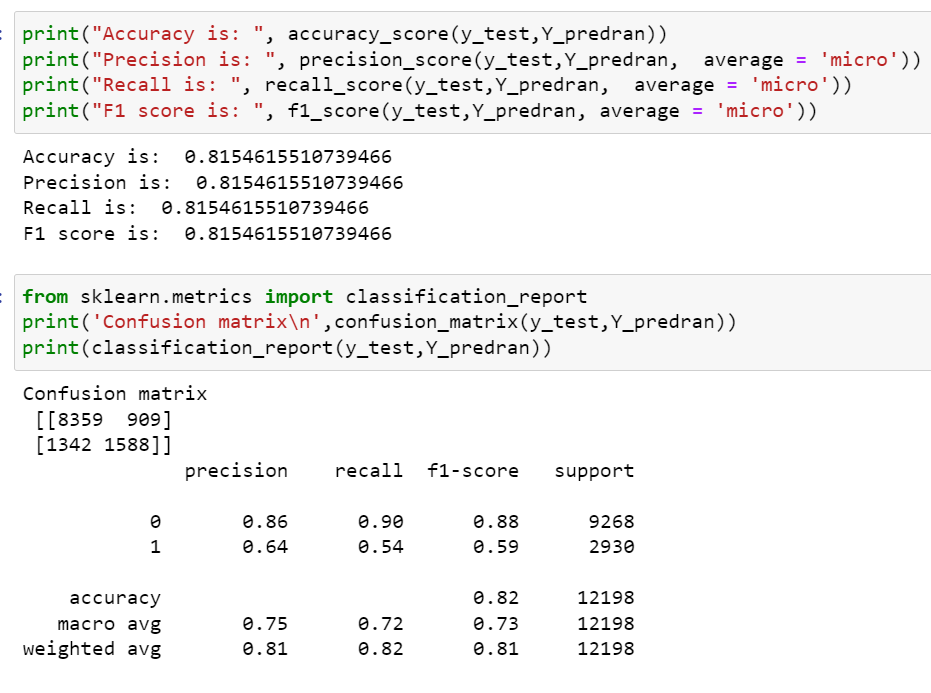
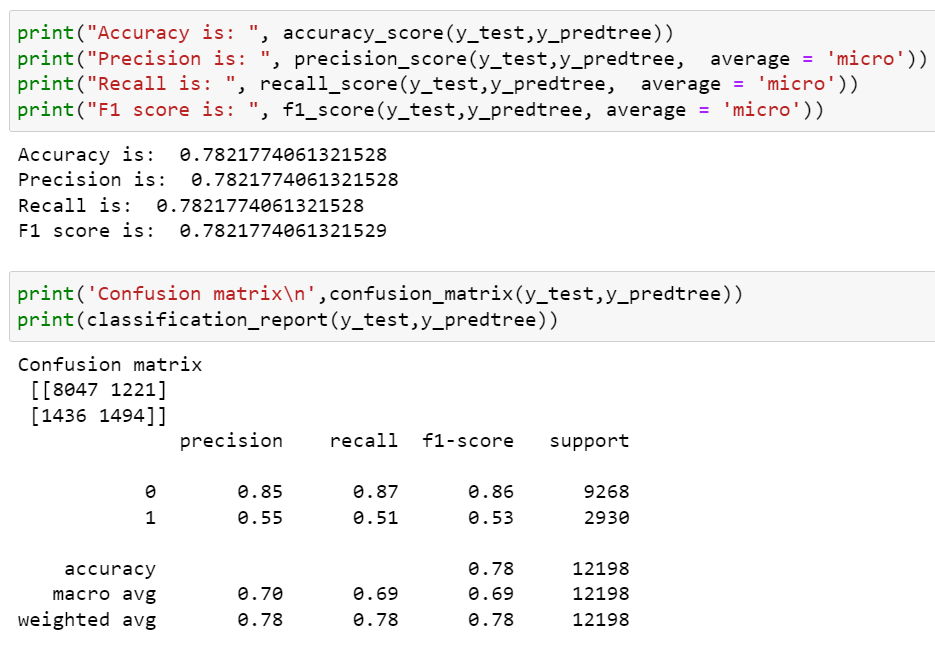


Fig.16 Classification report for Random forest

1. **Decision Tree**

A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node)

holds a class label.

 Fig.17 Classification report for Decision tree

**D. Gradient Boosting**

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment.

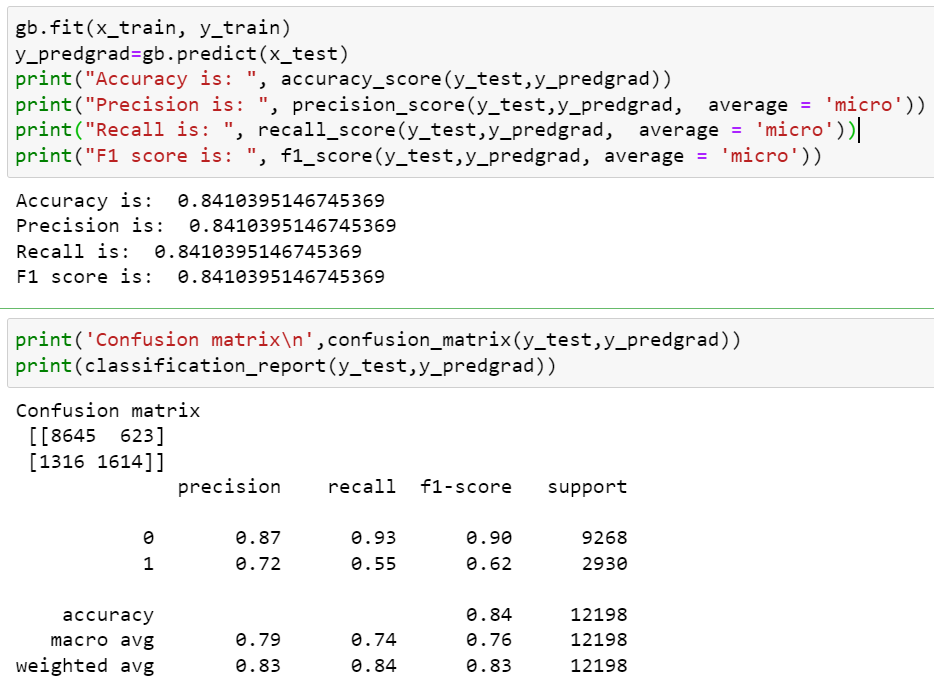


Fig.18 Classification report for Gradient Boosting

1. **XgBoost**

XgBoost stands for Extreme Gradient Boosting, It is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XgBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

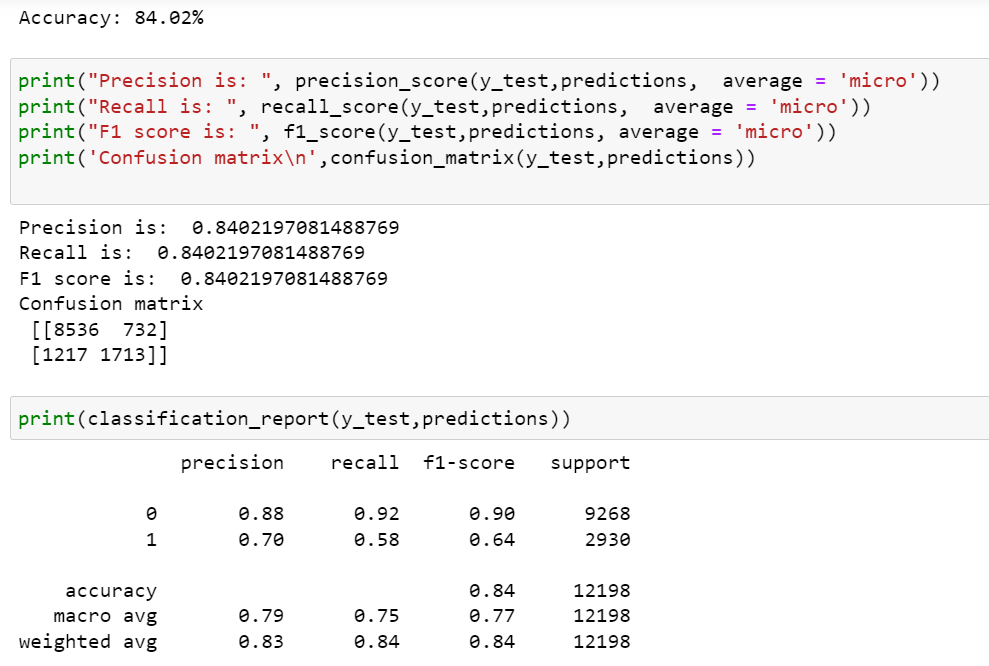


Fig.19 Classification report for XgBoost

**6.HYPERPARAMETER TUNNING**

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

Random Forest

Random forest is tunned by GridSearchCV method that, instead of sampling randomly from a distribution, evaluates all combinations we define.

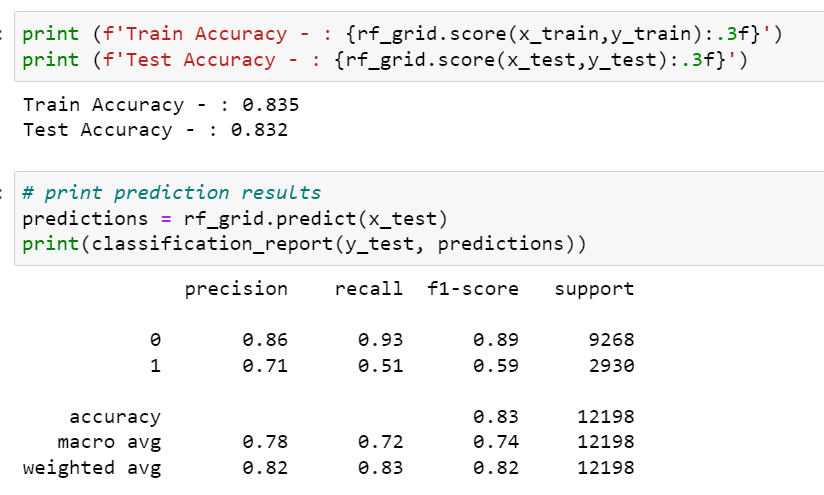


Fig 20: Classification report for Random forest after tunning

**XgBoost**

Xgboost tunning done by RandomizedSearchCV for our model. XgBoost combined with RandomSearchCV is very strong in HR salary prediction. We have compared it with the simple Random forest with GridSearchCV at it outperforms it anyway.

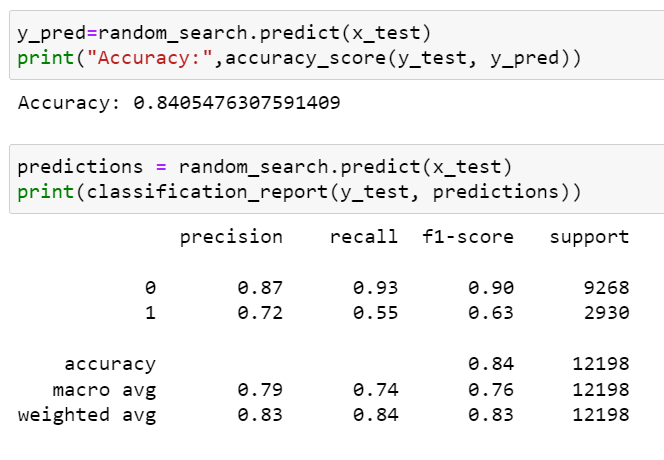


Fig 21 :Classification report for XgBoost after tunning

**CONCLUSION**

The project focuses on finding the right future salary of an applicant based on some parameter of a particular domain. The algorithms used are Logistic Regression, Random forest, Decision tree, Gradient boosting, XgBoost to train and perform predictions using the ML model. The result of the system is calculated by suitable algorithm by comparing it with another algorithms in terms of standard scores and classification accuracy, theF1score, the Precision, Recall etc. Out of the all these, it is XgBoost which is more accurate with an accuracy (F1- Score) of 84.05 % accuracy . All this is only possible if proper data cleaning is done by removing all the missing, incorrect and noisy data from the dataset to get an efficient result. Hence, this model has the capability to act as an aid for HR to predict salary precisely quite conveniently.

LIMITATIONS

1.No. of records is never enough to perfectly classify and predict the salary with complete accuracy as the larger the data, the more accurate is the ML Model.

2. The dataset predicted personal income levels being above or below 50,000 per year based on personal details such as marital status, education level, etc. There could have been many more cases of incomes less than 50K or above $50K, although the skew is not severe.

STREAMLIT

Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps. In this project a simple dashboard is created using streamlit.

LINK:https://share.streamlit.io/sanita88/streamlit1/main/try.py

LINK TO CODE

https://github.com/sanita88/INTERNSHIP-TCS/blob/main/HR%20salary%20prediction.ipynb