MACHINE LEARNING ALGORITHM FOR SOLVING REAL-WORLD CLASSIFICATION AND CLUSTERING PROBLEMS

FACE RECOGNITION USING MACHINE LEARNING MODEL TO PREDICT GENDER

LECTURE: Dr. Omid Chatrabgoun

STUDENT NAME & ID: Jusel Justin (12646627)

Anu Thomas (12386712)

Table of Contents:

Abstract

Face recognition has been done with the help of python programming language and that has predicted gender by various classification models. Naive Bayes, random forest classifiers and Logistic regression are the classifiers techniques which are used to apply the classification of image data and then gender has been predicted based on these classification techniques of image. With this data, machine learning algorithms have been applied to the image of the image data with CSV files and then the gender has been identified. Image classification is the process of assigning a label or category to an image. It is commonly used in machine learning and is used to classify objects, scenes, and activities in images.

Introduction

The use of classification techniques to predict the gender of a person in an image is a relatively new research area. There are a variety of methods have been proposed, including deep learning, traditional machine learning, and even rule-based approaches. In this research, the prediction has been done based on the provided dataset along with the folder of image data and then various classification techniques such as Naive Bayes, random forest classifiers and Logistic regression has been used. By using techniques such as naive Bayes, logistic regression, and random forest, it is possible to predict the gender of a person in a given image. Logistic regression is a supervised machine learning algorithm that is used to predict the probability of an event occurring. Random forest is an ensemble machine learning algorithm that builds multiple decision trees and combines their results for better accuracy. Naive Bayes is a supervised machine learning algorithm that uses probabilistic models to classify data.

Problem and data set

The challenge in this research is to pinpoint the gender based on the given images of a person whether it is an image of a male or female and in the CSV dataset, various types of structure of images have been provided [1]. After feature selection and data cleaning, classification techniques are employed, which are based on data exploration and an evaluation of the data's nature [2].

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	_id	5_o_Clock	Arched_E	Attractive	Bags_Und	Bald	Bangs	Big_Lips	Big_Nose	Black_Hai	Blond_Hai	Blurry	Brown_Ha	Bushy_Eyi	Chubby	Double_C	Eyeglasse	Goatee	Gray_Hair	Heavy_Mc	High
1	1.jpg	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	
2	2.jpg	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	
3	3.jpg	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	
4	4.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
5	5.jpg	-1	1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	
6	6.jpg	-1	1	1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	
7	7.jpg	1	-1	1	1	-1	-1	1	1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	
8	8.jpg	1	1	-1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
9	9.jpg	-1	1	1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	
10	10.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	
11	11.jpg	-1	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
12	12.jpg	-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	
13	13.jpg	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
14	14.jpg	-1	1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	1	
15	15.jpg	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
16	16.jpg	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
17	17.jpg	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
18	18.jpg	-1	1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	
19	19.jpg	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	

Figure 1: Dataset used for image prediction

The above figure has shed light on the CSV dataset in which the structure of the shape of a person has been provided in detail. The target variable is also provided to check if the person is male or female which can be done with the help of other attributes which can be said in this dataset as the independent variables. This dataset is excellent for building and implementing face detection models, especially for identifying facial characteristics such as identifying people who have brown hair, a smile, or glasses [3]. Vast position changes, backdrop clutter, and a variety of persons are all covered in the photographs, which are accompanied by many images and detailed comments [4].

Method

The approaches used with the image folder data use machine learning algorithms which are classification techniques. The Jupyter Notebook must first be imported with several libraries before the dataset can be read using Pandas [5]. Basic checks would be made, such as removing null values from dataset attributes after they had been verified to be true. Then the data exploration has been done based on the provided attributes of the CSV dataset and then the images have been recognised and read with the folder of images [6]. The data splitting should be done based on the target variable and independent variable. Exploratory data analysis is the process of creating graphs and charts based on the observation of the dataset to extract the nature of the dataset.

The dataset is then split into train and test where the size of the test dataset is provided at 25% and the train size is provided at 75%. The classification techniques are then applied to the model where all three techniques are applied on the train by fitting and prediction is done based on the test data [7]. Various metrics are then obtained based on the prediction which is accuracy score, classification report and confusion matrix. The comparison graph is then created based on the accuracy score of all three models with a bar plot to extract the more accurate model, that is, which algorithm predicts a more accurate prediction of the gender of the image data [8].

Experimental setup

A mass sufficient data that has the characteristics and labels the machine learning classifier requires. The numerical data should be normalized or standardized before processing to reduce noise and fill in missing values. Separate the data into a training set and a test set, and then use the test set to assess the classifier's performance after training on the training set. The classifier should be trained using the training set [9]. A different algorithm may be employed depending on the classifier. Use the test set to gauge how well the classifier performed. Performance may be assessed using a variety of elements, including precision and accuracy. Using the information from the test set, alter the classifier's settings to enhance performance. To use the concept in the real world, deploy it to a production server.

Result

The classification algorithms which are naive Bayes, logistic regression and random forest are applied to the image data to classify the image with gender as a target variable in this research.

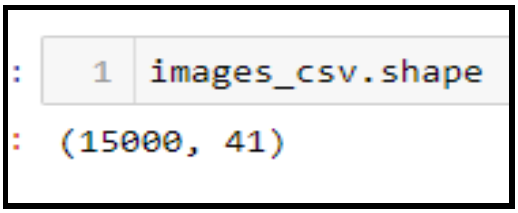

Figure 2: Shape of the provided CSV dataset

The shape function has been used here to show the image_csv shape where the row numbers are 15000 and the column number is 41. The above image has been done by the researcher where the shape function has been used.

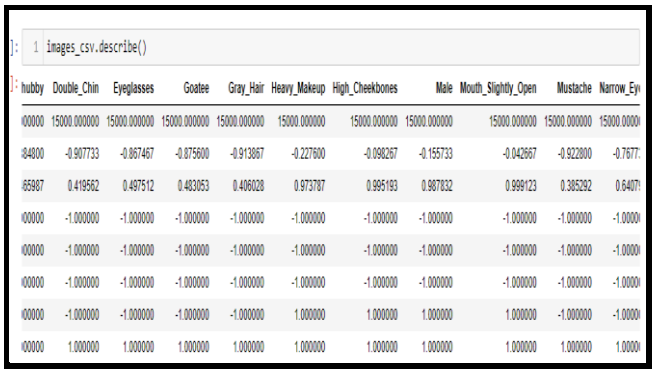

Figure 3: Describing the statistics of the provided CSV dataset

The described function has been illustrated by the researcher in Jupyter notebook software using the python programming language. The above image shows the mean and median values of each column and row.

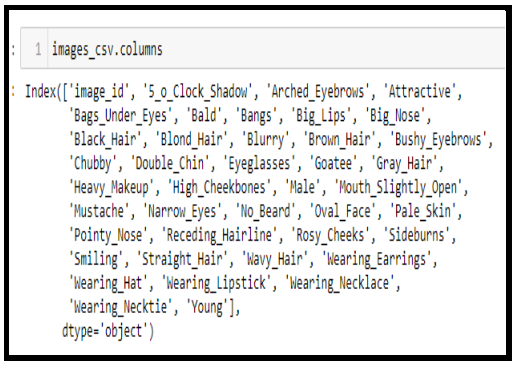

Figure 4: Columns name of the dataset

The column's function has been used in the image where this has been used to show the column names which have been seen in the images_csv. The above image shows the column names and that has been done with the help of the user.

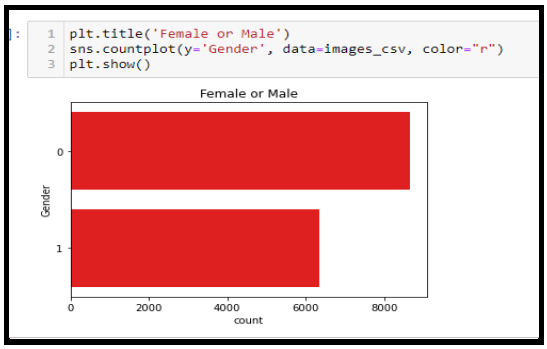

Figure 5: Bar graph for female or male

The bar graph shows the difference between males and females in the dataset where it has been shown that the female value is more than the male count. In the x-axis, the count has been taken and in the Y section, the Gender has been taken.

Figure 6: Visualization of the image from the image folder

The above image shows the identifying human image which has been taken from the image dataset folder named img_align_celeba. The image name is 000001.jpg and the show function has been used here to show the image that has been collected in the imgplot variable.

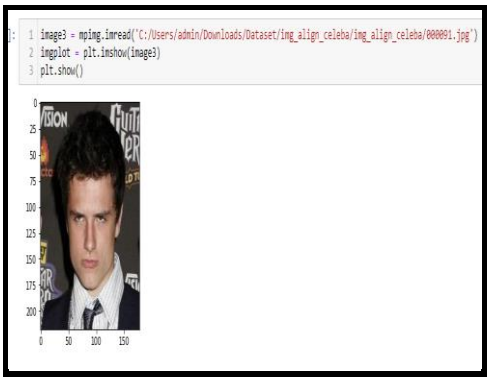

Figure 7: Visualization of the image from the image folder

The above image illustrates a human image which has been taken from the image dataset folder named img_align_celeba. The image name is 000091.jpg and the show function has been used here to show the image that has been collected in the imgplot variable [9].

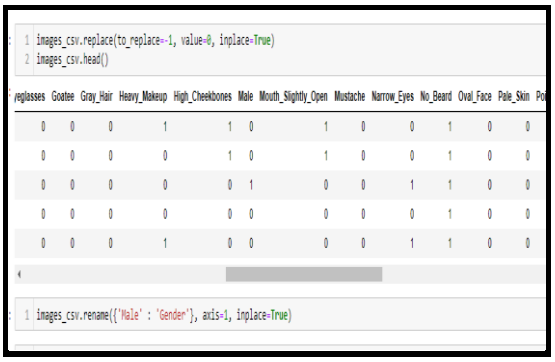

Figure 8: Visualization of the image from the image folder

Here, the researcher has shown the first five rows where replace is -1 and value is 0 and the place is true. The head function has been used in this image with python language [10].

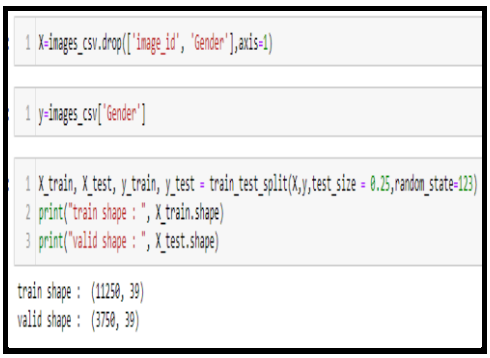

Figure 9: Splitting of the dataset into train and test

The figure above clearly shows how the dataset was split into train and test groups. The test dataset is 25% smaller than the training dataset, which is 75% of its total size [11]. The dataset can be fitted into train and test sets by randomly selecting a portion of the data points to serve as the test set and the remaining data points to serve as the train set.

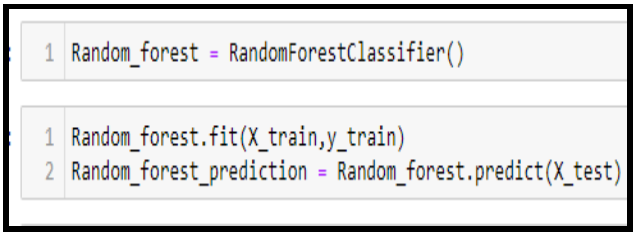

Figure 10: Applying the random forest classifier by fitting and making the prediction

The above figure illustrates the fitting of the random forest classifiers on the dependent variables to predict the target variable, which is the gender attribute of the dataset. The X_train and y_train have

been fitted into the random forest classifiers and based on that prediction have been defined on the X_test dataset.

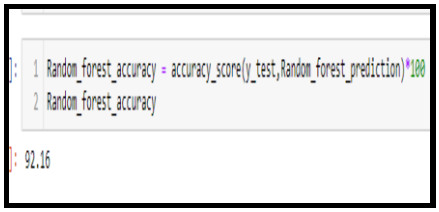

Figure 11: Accuracy score based on the test dataset by Random Forest Classifier

The above figure illustrates the accuracy score of the test dataset predicted on the target variable which is the gender attribute. The accuracy score has been evaluated with the Random Forest Classifier as 92.16% as shown in the figure.

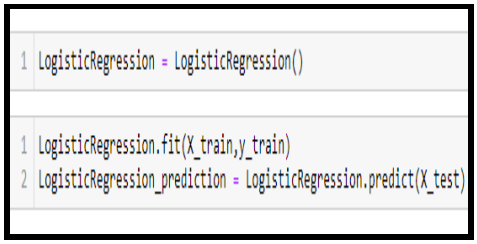

Figure 12: Applying the Logistic regression by fitting and making the prediction

The above figure illustrates the fitting of the Logistic regression on the dependent variables to predict the target variable which is the gender attribute of the dataset. The X_train and y_train have been fitted into the Logistic regression and based on that prediction have been defined on the X_test dataset.

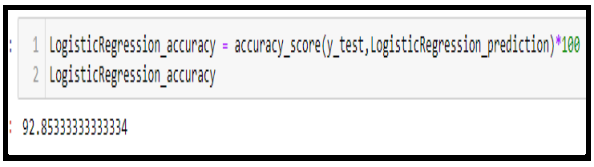

Figure 13: Accuracy score based on the test dataset by Logistic regression

The above figure illustrates the accuracy score of the test dataset predicted on the target variable which is the Gender attribute. The accuracy score has been evaluated with the Logistic regression as 92.85% as shown in the figure.

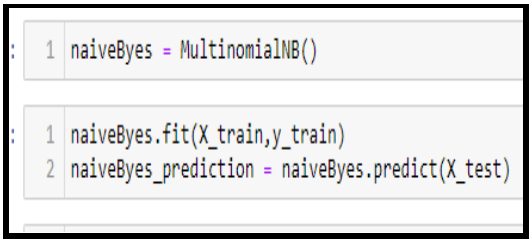

Figure 14: Applying the Naive Bayes by fitting and making the prediction

The above figure illustrates the fitting of the Naive Bayes on the dependent variables to predict the target variable which is the gender attribute of the dataset. The X_train and y_train have been fitted into the Naive Bayes and based on that prediction have been defined on the X_test dataset.

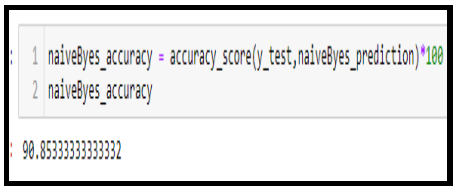

Figure 15: Accuracy score based on the test dataset by Naive Bayes

The above figure illustrates the accuracy score of the test dataset predicted on the target variable, which is the gender attribute [12]. The accuracy score has been evaluated with the Naive Bayes as 90.85% as shown in the figure.

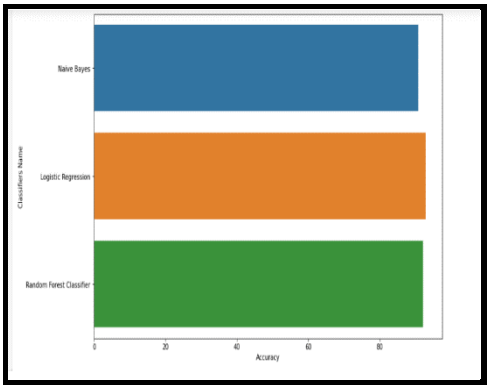

Figure 16: Comparison bar graph with all three-classification techniques

The code that was run in the Jupyter Notebook to compare the three models is shown in the above picture along with the comparison model graph. It is clear from the graph that the Logistic Regression model predicts gender the most accurately, whereas the other model does so somewhat less well.

Discussion

The image dataset has been used to test the performance of the Naive Bayes, Logistic Regression, and Random Forest classifiers. The accuracy of Naive Bayes was rated at 90.85%, the accuracy of Random Forest classifiers at 92.16%, and the accuracy of Logistic Regression at 92.85% [13]. These measurements showed that the Logistic Regression were the best reliable model for predicting the attitudes based on the gender attribute of the image data. Based on the Bayes theorem, Naive Bayes is a probabilistic classification algorithm. This straightforward algorithm can be applied to classification tasks on tweet text data [14]. Random Forest is an algorithm for ensemble learning that builds multiple

decision trees and then combines the output from each tree's analysis. The probability of a particular class is predicted using the linear classification algorithm logistic regression [15].

Conclusion

Based on the provided dataset, machine learning classifiers have been used to ascertain the gender and predict the more accurate gender of the image person. Naive Bayes, Logistic Regression and Random Forest are all popular classification techniques for determining the gender of image data. Each of these techniques has advantages and disadvantages, and the best choice will depend on the specific image data and the desired results. Naive Bayes is a simple and effective technique that is suitable for small datasets and can help identify gender with a high level of accuracy. Logistic regression is a more sophisticated approach that can help identify gender with a higher level of accuracy. Finally, Random Forest is a powerful technique that can help identify gender with even higher accuracy but requires more data and computing resources. This research has shown that logistic regression is emerged as the most accurate model to identify the gender of the image data.

Reference

1. Shaker, S.H. and Al-Khalidi, F.Q., 2022. Human Gender and Age Detection Based on Attributes of Face. International Journal of Interactive Mobile Technologies, 16(10).
2. Sunitha, G., Geetha, K., Neelakandan, S., Pundir, A.K.S., Hemalatha, S. and Kumar, V., 2022. Intelligent deep learning-based ethnicity recognition and classification using facial images. Image and Vision Computing, 121, p.104404.
3. Dukić, D. and SovicKrzic, A., 2022. Real-time facial expression recognition using deep learning with application in the active classroom environment. Electronics, 11(8), p.1240.
4. Siddiqui, H., Rattani, A., Ricanek, K. and Hill, T., 2022. An Examination of Bias of Facial Analysis based BMI Prediction Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2926-2935).
5. Wahlang, I., Maji, A.K., Saha, G., Chakrabarti, P., Jasinski, M., Leonowicz, Z. and Jasinska, E., 2022. Brain Magnetic Resonance Imaging Classification Using Deep Learning Architectures with Gender and Age. Sensors, 22(5), p.1766.
6. Morales Sánchez, D., Moreno, A. and Jiménez López, M.D., 2022. A White-Box Sociolinguistic Model for Gender Detection. Applied Sciences, 12(5), p.2676.
7. Mandis, I.S., Reducing Racial and Gender Bias in Machine Learning and Natural Language Processing Tasks Using a GAN Approach.
8. d'Eon, G., d'Eon, J., Wright, J.R. and Leyton-Brown, K., 2022, June. The spotlight: A general method for discovering systematic errors in deep learning models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1962-1981).
9. Salman, F.M. and Abu-Naser, S.S., 2022. Classification of Real and Fake Human Faces Using Deep Learning. International Journal of Academic Engineering Research (IJAER), 6(3).

10. Albert, K. and Delano, M., 2022. Sex Trouble: Common pitfalls in incorporating sex/gender in medical machine learning and how to avoid them. arXiv preprint arXiv:2203.08227.

11. Firdaus, F. and Munir, R., 2022, March. Masked Face Recognition using Deep Learning based on Unmasked Area. In 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T) (pp. 1-6). IEEE.

12. Bucher, T., 2022. Facing AI: conceptualizing 'fAIcecommunication'as the modus operandi of facial recognition systems. Media, Culture & Society, p.01634437211036975.

13. Cao, Y., Berend, D., Tolmach, P., Amit, G., Levy, M., Liu, Y., Shabtai, A. and Elovici, Y., 2022. Fair and accurate age prediction using distribution aware data curation and augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3551-3561).

14. Ikae, C. and Savoy, J., 2022. Gender identification on Twitter. Journal of the Association for Information Science and Technology, 73(1), pp.58-69.

15. Pagano, T.P., Loureiro, R.B., Araujo, M.M., Lisboa, F.V.N., Peixoto, R.M., Guimaraes, G.A.D.S., Santos, L.L.D., Cruz, G.O.R., de Oliveira, E.L.S., Cruz, M. and Winkler, I., 2022. Bias and unfairness in machine learning models: a systematic literature review. arXiv preprint arXiv:2202.08176.

Appendix

Contribution

Jusel Justin: I processed with Jupiter notebook software and using Python programming language. Various libraries like pandas, NumPy, matplotlib and seaborn were imported. for face recognition dataset. I collected all those relevant data which are associated with the predict the gender [male/ female]. I imported the given dataset from UCI where the image recognition data has been notified.

Mean and median values of each row and column were determined.

Splitting of data, dataset to train and test

Naive base algorithm was run by jusel

Logistic regression algorithm was tested partially

Report was rechecked and completed

Anu Thomas: I proceeded with the writing part and gone through different relevant literatures which are associated with the field of machine learning. As there are different kinds of models are out there which are associated with this kind of face recognition dataset. I went through those portions of the studies and implemented its own research for better understanding and proper explanation.

Data exploration has been done using statistical parameters defined inbuilt function in python.

The bar graph of male/ female was determined. Comparison of bar graph with all three-classification technic was studied.

Random forest algorithm run by Anu

Logistic regression algorithm was tested completely, and the best algorithm was identified.

CODE:

```
“! pip3 install torch torchvisiontorchaudio
importnumpy as np
import pandas as pd
importmatplotlib.pyplot as plt
importmatplotlib.image as mpimg
importseaborn as sns
importos
import sys
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
%pylab inline
import torch

importtorchvision
importtorchvision.transforms as transforms
from torchvision.io import read_image

fromtorch.utils.data import Dataset ,DataLoader

fromtorch.nn import Module
importtorch.nn as nn
importtorch.nn.functional as F
pd.options.mode.chained_assignment = None
images_csv = pd.read_csv('C:/Users/admin/Downloads/Dataset/list_attr_celeba.csv')
images_csv = images_csv.iloc[0:15000,:]
image = 'C:/Users/admin/Downloads/Dataset/img_align_celeba'
pd.pandas.set_option('display.max_columns', None)
```

```
images_csv

images_csv.shape

images_csv.describe()

images_csv.columns

images_csv.info()

images_csv.replace(to_replace=-1, value=0, inplace=True)

images_csv.head()

images_csv.rename({'Male' : 'Gender'}, axis=1, inplace=True)

images_csv.columns

plt.title('Female or Male')

sns.countplot(y='Gender', data=images_csv, color="r")

plt.show()

plt.figure(figsize=(20, 15))

sns.heatmap(images_csv.corr(), cmap="YlGnBu", annot=True)

plt.show()

images_csv.loc[0]

image1 = mpimg.imread('C:/Users/admin/Downloads/Dataset/img_align_celeba/img_align_celeba/000001.jpg')

imgplot = plt.imshow(image1)

plt.show()

image2 = mpimg.imread('C:/Users/admin/Downloads/Dataset/img_align_celeba/img_align_celeba/000010.jpg')

imgplot = plt.imshow(image2)

plt.show()

image3 = mpimg.imread('C:/Users/admin/Downloads/Dataset/img_align_celeba/img_align_celeba/000091.jpg')

imgplot = plt.imshow(image3)

plt.show()

images_csv.columns

fromsklearn.model_selection import train_test_split

X=images_csv.drop(['image_id', 'Gender'],axis=1)

y=images_csv['Gender']
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.25,random_state=123)
print("train shape : ", X_train.shape)
print("valid shape : ", X_test.shape)
fromsklearn.ensemble import RandomForestClassifier
fromsklearn.metrics import accuracy_score
fromsklearn.metrics import classification_report
fromsklearn.metrics import confusion_matrix
Random_forest = RandomForestClassifier()
Random_forest.fit(X_train,y_train)
Random_forest_prediction = Random_forest.predict(X_test)
Random_forest_accuracy = accuracy_score(y_test,Random_forest_prediction)*100
Random_forest_accuracy
print(classification_report(Random_forest_prediction,y_test))
print(confusion_matrix(Random_forest_prediction,y_test))
fromsklearn.linear_model import LogisticRegression
LogisticRegression = LogisticRegression()
LogisticRegression.fit(X_train,y_train)
LogisticRegression_prediction = LogisticRegression.predict(X_test)
LogisticRegression_accuracy = accuracy_score(y_test,LogisticRegression_prediction)*100
LogisticRegression_accuracy
fromsklearn.naive_bayes import MultinomialNB
naiveByes = MultinomialNB()
naiveByes.fit(X_train,y_train)
naiveByes_prediction = naiveByes.predict(X_test)
naiveByes_accuracy = accuracy_score(y_test,naiveByes_prediction)*100
naiveByes_accuracy
data = [['Naive Bayes', naiveByes_accuracy], ['Logistic Regression',
LogisticRegression_accuracy],['Random Forest Classifier',Random_forest_accuracy]]
df = pd.DataFrame(data, columns=['Classifiers Name', 'Accuracy'])
plt.figure(figsize=(12,8))
sns.barplot(x='Accuracy', y='Classifiers Name', data = df)
```

plt.show() ”