AGE AND GENDER DETECTION

A Report submitted

in partial fulfillment for the Degree of

B. Tech in

Computer Science and Engineering

By

Mansi Gaur

Muskan Bansal

pursued in

Department of Computer Science and Engineering

To

MERI COLLEGE OF ENGINEERING AND TECHNOLOGY

46th Milestone Rohtak Road (near Sampla, Bahadurgarh, Haryana 124501

ACKNOWLEDGEMENT

The successful realization of the project is an outcome of consolidated effort of people from different

fronts. Their wisdom and support motivated me to learn quickly and contribute to “Age and Gender Detection”.

One cannot even imagine the power of the force that guides us all and neither can we succeed without acknowledging it. Our deepest gratitude to almighty God for holding our hands and guiding us throughout our lives.

I would like to acknowledge my college faculty for trusting me while doing this project. This exposure has enriched me with technical knowledge and has also introduced me to the attributes of a successful professional.

I am thankful to Dr. (Prof.) Umesh Gupta, Director & Dean, MERI College of Engineering and Technology, Maharshi Dayanand University, Rohtak for his consistent encouragement, valuable suggestions and moral support.

I would like to wholeheartedly thank my faculty supervision Dr. Ritu Aggarwal, Assistant Professor, Department of Computer Science and Engineering for their expert guidance, encouragement and valuable suggestions at every step.

I shall remain indebted to the faculty members of MERI College of Engineering and Technology, Maharshi Dayanand University, Rohtak for their cooperation, kindness and general help extended tome during the completion of this work.

My greatest thanks are to all who wished me success, especially my parents, my brother and sister. Above all I render my gratitude to the almighty who bestowed self-confidence, ability and strength on me to complete this work.

I am extremely happy to acknowledge and express my sincere gratitude to my family and friends for their constant support.

FORMATTING CERTIFICATE

This is to certify that this is a bonafied record of the project work done satisfactorily at MERI COLLEGE OF ENGINEERING AND TECHNOLOGY by Ms. Mansi Gaur and Muskan Bansal Roll no. 4075358 & 40753 in partial fulfillment of B. TECH 8th Semester Examination.

This report or a similar report on the topic has not been submitted for any other examination and does not form part of any other course undergone by the candidate.

Signature of Candidate

Place:

Date:

Signature of Project Guide:

Name:

Designation:

Address:

Name & Seal of the Institute

DECLARATION

We,Mansi Gaur and Muskan Bansal, student of Bachelor of Technology (Computer Science & Engineering), in MERI College of Engineering and Technology, Maharshi Dayanand University, Rohtak, for the session 2019- 2023, hereby declare that dissertation entitled “Age and Gender Detection” towards fulfillment of the requirement for the award of the Degree of Bachelor of Technology in Computer Science & Engineering and submitted in MERI College of Engineering and Technology, Maharshi Dayanand University, Rohtak, Haryana, India is an authentic record of my own work carried out, under the supervision of Dr. Ritu Aggarwal (Guide) Assistant Professor, MERI College of Engineering and Technology, Maharshi Dayanand University, Rohtak, Haryana. The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other Institute/University. I have taken care of all aspects to honor intellectual property rights and have acknowledged the contribution ofothers for using them in this academic purpose. I further declared that in case of any violation of intellectual property rights or copyright, I would be fully responsible for the same.

Mansi Gaur

Muskan Bansal

B.Tech (CSE)

INDEX

What is Machine Learning?

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.

The goal of AI is to create computer models that exhibit "intelligent behaviors" like humans, according to Boris Katz, a principal research scientist and head of the InfoLab Group at CSAIL. This means machines that can recognize a visual scene, understand a text written in natural language, or perform an action in the physical world.

Machine learning is one way to use AI. It was defined in the 1950s by AI pioneer Arthur Samuel as "the field of study that gives computers the ability to learn without explicitly being programmed."

"The function of a machine learning system can be **descriptive**, meaning that the system uses the data to explain what happened; **predictive**, meaning the system uses the data to predict what will happen; or **prescriptive**, meaning the system will use the data to make suggestions about what action to take," the researchers wrote."

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Need for Machine Learning

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

The importance of machine learning can be easily understood by its use's cases, Currently, machine learning is used in **self-driving cars**, **cyber fraud detection**, **face recognition**, and **friend suggestion by Facebook**, etc. Various top companies such as Netflix and Amazon have built machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

Following are some key points which show the importance of Machine Learning:

- Rapid increment in the production of data
- Solving complex problems which are difficult for a human
- Decision making in various sector including finance
- Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Semi Supervised learning**
4. **Reinforcement learning**

Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below:

- **Random Forest Algorithm**
- **Decision Tree Algorithm**
- **Logistic Regression Algorithm**

- **Support Vector Machine Algorithm**

- **Regression**

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

- **Simple Linear Regression Algorithm**
- **Multivariate Regression Algorithm**
- **Decision Tree Algorithm**
- **Lasso Regression**

Advantages and Disadvantages of Supervised Learning

Advantages:

- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.

Disadvantages:

- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

- **Image Segmentation:** Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
- **Medical Diagnosis:** Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with

labels for disease conditions. With such a process, the machine can identify a disease for the new patients.

- **Fraud Detection -** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
- **Spam detection -** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
- **Speech Recognition -** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

- **Clustering**
  The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.
  Some of the popular clustering algorithms are given below:

- **K-Means Clustering algorithm**
- **Mean-shift algorithm**
- **DBSCAN Algorithm**
- **Principal Component Analysis**
- **Independent Component Analysis**

- **Association**

  Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in **Market Basket analysis, Web usage mining, continuous production**, etc.

  Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

Advantages and Disadvantages of Unsupervised Learning Algorithm

Advantages:

- These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
- Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

Disadvantages:

- The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.
- Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

Applications of Unsupervised Learning

- **Network Analysis:** Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
- **Recommendation Systems:** Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
- **Anomaly Detection:** Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.

- **Singular Value Decomposition:** Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.

Semi-Supervised Learning

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced. The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

We can imagine these algorithms with an example. Supervised learning is where a student is under the supervision of an instructor at home and college. Further, if that student is self-analysing the same concept without any help from the instructor, it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of an instructor at college.

Advantages and disadvantages of Semi-supervised Learning

Advantages:

- It is simple and easy to understand the algorithm.
- It is highly efficient.
- It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages:

- Iterations results may not be stable.
- We cannot apply these algorithms to network-level data.

Reinforcement Learning

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.**

A reinforcement learning problem can be formalized using **Markov Decision Process(MDP).** In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

Categories of Reinforcement Learning

Reinforcement learning is categorized mainly into two types of methods/algorithms:

- **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
- **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

Advantages and Disadvantages of Reinforcement Learning

Advantages

- It helps in solving complex real-world problems which are difficult to be solved by general techniques.

- The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
- Helps in achieving long term results.

Disadvantage

- RL algorithms are not preferred for simple problems.
- RL algorithms require huge data and computations.
- Too much reinforcement learning can lead to an overload of states which can weaken the results.

Applications of Reinforcement Learning

- **Video Games:** RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are **AlphaGO** and **AlphaGO Zero**.
- **Resource Management:** The "Resource Management with Deep Reinforcement Learning" paper showed that how to use RL in computer to automatically learn and schedule resources to wait for different jobs in order to minimize average job slowdown.
- **Robotics:** RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning. There are different industries that have their vision of building intelligent robots using AI and Machine learning technology.
- **Text Mining**: Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.

Introduction

UTKFace dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. This dataset could be used on a variety of tasks, e.g., face detection, age estimation, age progression/regression, landmark localization, etc.

- The objective of the project is to detect gender and age using facial images. Convolutional Neural Network is used to classify the images. There are 2 output types namely, gender(M or F) and age.

- Download link: https://www.kaggle.com/datasets/jangedoo/utkface-new

- Environment: Kaggle

- In this Python Project, I had used Deep Learning to accurately identify the gender and age of a person from a single image of a face.

- The predicted gender may be one of ‘Male’ and ‘Female’, and the predicted age may be one of the following ranges- 0-116.

- It is very difficult to accurately guess an exact age from a single image because of factors like makeup, lighting, obstructions, and facial expressions. I have used both classification and regression neural network in my project.

Background

A wide range of machine learning algorithms is available for the learning process. This section describes the classification algorithm used in gender detection and regression algorithm used in age detection.

Classification:

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

The task of the classification algorithm is to find the mapping function to map the input(x) to the discrete output(y).

Example: The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

Types of ML Classification Algorithms:

Classification Algorithms can be further divided into the following types:

- Logistic Regression
- K-Nearest Neighbours
- Support Vector Machines
- Kernel SVM
- Naïve Bayes
- Decision Tree Classification
- Random Forest Classification

Regression:

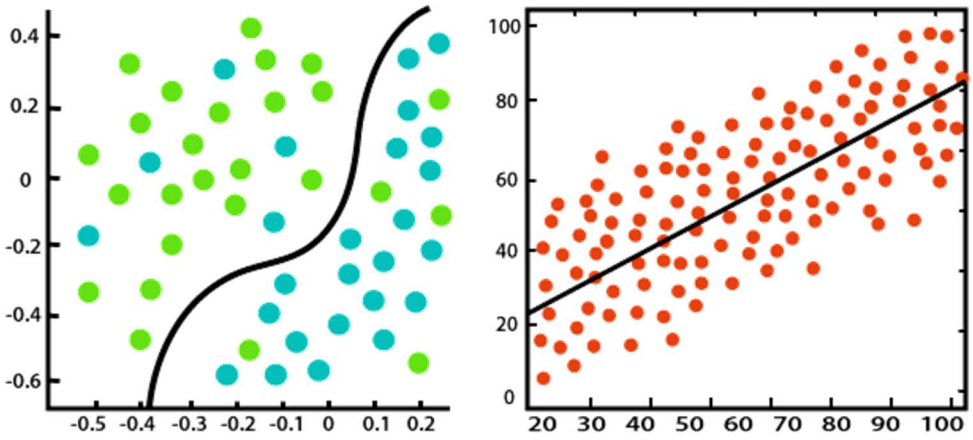
Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of **Market Trends**, prediction of House prices, etc.

The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

Types of Regression Algorithm:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression



Classification Regression

RandomForestClassifier:

Random forests or random decision forests are an ensemble learning method for classification ,regression and other tasks that operate by constructing a multitude of decision trees at training time.

Naive Bayes:

It is used to classify objects .It assumes strong, independent attributes of data points.it also includes spam filters, text analysis and even medical diagnosis.

Decision tree classifier:

A decision tree is a decision support tool that uses a tree-like model of decision and their possible Consequences ,including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Grey Scale

A greyscale image is simply one in which the only colours represented are different shades of grey. Whilst we often refer to such images as “black and white” in everyday conversation, a truly “black and white image” would consist of only these

two distinct colours, which is very rarely the case; making ‘greyscale’ the more accurate term.

As there is no colour information to represent for a greyscale image, less information needs to be stored for each pixel and an additive colour model is not required! For greyscale images, the only information we require is a single value to represent the intensity of each pixel; the higher this value, the lighter the shade of grey. As such, greyscale images usually consist of a single channel, where each pixel intensity is just a single number ranging from 0 to 255.

ReLU

The **rectified linear activation function** or **ReLU** for short is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance.

Max Pooling

Max Pooling is a pooling operation that calculates the maximum value for patches of a feature map, and uses it to create a downsampled (pooled) feature map. It is usually used after a convolutional layer. It adds a small amount of translation invariance - meaning translating the image by a small amount does not significantly affect the values of most pooled outputs.

LIBRARIES:

- **Pandas**

Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc**.**

- **Numpy**

The name “Numpy” stands for “Numerical Python”. It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like TensorFlow use

Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

- **Matplotlib**
  This library is responsible for plotting numerical data. And that's why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc.

- **Keras**
  Keras is a high-level library for deep learning, written in Python, that can run on different back-end engines like Theano, TensorFlow, or CNTK. It is open-source, user-friendly, and scalable for faster neural network experimentation.

- **Tensorflow**
  This library was developed by Google in collaboration with the Brain Team. It is an open-source library used for high-level computations. It is also used in machine learning and deep learning algorithms. It contains a large number of tensor operations. Researchers also use this Python library to solve complex computations in Mathematics and Physics.

- **scikit-learn**
  It is a famous Python library to work with complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc. This library works in association with Numpy and SciPy.

Kaggle:

Kaggle is a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle was first launched in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and Jeremy Howard. Nicholas Gruen was the founding chair succeeded by Max Levchin. Equity was raised in 2011 valuing the company at $25.2 million. On 8 March 2017, Google announced that they were acquiring Kaggle.

Competitions:

Many machine-learning competitions have been run on Kaggle since the company was founded. Notable competitions include one improving gesture recognition for Microsoft Kinect, making a football AI for Manchester City, and improving the search for the Higgs boson at CERN.

Competitions have resulted in successful projects such as furthering HIV research, chess ratings and traffic forecasting. Geoffrey Hinton and George Dahl used deep neural networks to win a competition hosted by Merck. Vlad Mnih (one of Hinton's students) used deep neural networks to win a competition hosted by Adzuna. This resulted in the technique being taken up by others in the Kaggle community. Tianqi Chen from the University of Washington also used Kaggle to show the power of XGBoost, which has since replaced Random Forest as one of the main methods used to win Kaggle competitions.

Several academic papers have been published on the basis of findings made in Kaggle competitions. A contributor to this is the live leaderboard, which encourages participants to continue innovating beyond existing best practices.

Kaggle Progression System

Kaggle has implemented a progression system to recognize and reward users based on their contributions and achievements within the platform. This system consists of five tiers: Novice, Contributor, Expert, Master, and Grandmaster. Each tier is achieved by meeting specific criteria in competitions, kernels (code-sharing), and discussions.

The highest and most prestigious tier, Kaggle Grandmaster, is awarded to users who demonstrate exceptional skills in data science and machine

learning. Achieving this status is extremely challenging. As of April 4, 2023, out of 12 million Kaggle users, only 2,331 (about 1 out of every 5500 users) have reached the Master level.

Among these Masters, only 422 (approximately 1 out of every 5 Masters) have achieved the coveted Kaggle Grandmaster status.

The other tiers in the progression system include:

- 13 thousand Experts
- 200 thousand Contributors
- 12 million Novices.

The progression system serves to motivate users to continuously improve their skills and contribute to the Kaggle community.

Details About Hardware and Software's Used

Hardware Used:

Operating System Used: Windows 10
Processor: Intel i5
RAM: 8.00 GB
System Type: 64-bit

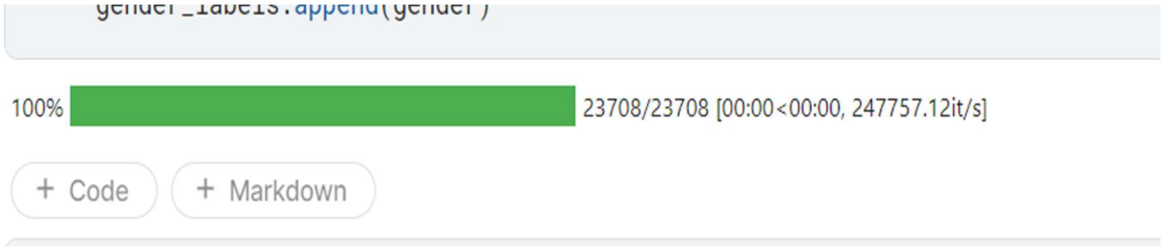
Software Used

Kaggle Environment (Kaggle allows to find datasets to use in building AI models, publish datasets, etc.)
Version: Latest
Language Used: Python

Code and Output

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from tqdm.notebook import tqdm
warnings.filterwarnings('ignore')
%matplotlib inline


from tensorflow.keras.preprocessing.image import load_img
import tensorflow as tf
#from keras.preprocessing.image import load_img
from keras.models import Sequential, Model
from keras.layers import Dense, Conv2D, Dropout, Flatten, MaxPooling2D,
Input



BASE_DIR = '/kaggle/input/dataset'




image_paths = []
age_labels = []
gender_labels = []

for filename in tqdm(os.listdir(BASE_DIR)):
    image_path = os.path.join(BASE_DIR, filename)
    temp = filename.split('_')
    age = int(temp[0])
    gender = int(temp[1])
    image_paths.append(image_path)
    age_labels.append(age)
```

```
gender_labels.append(gender)
```



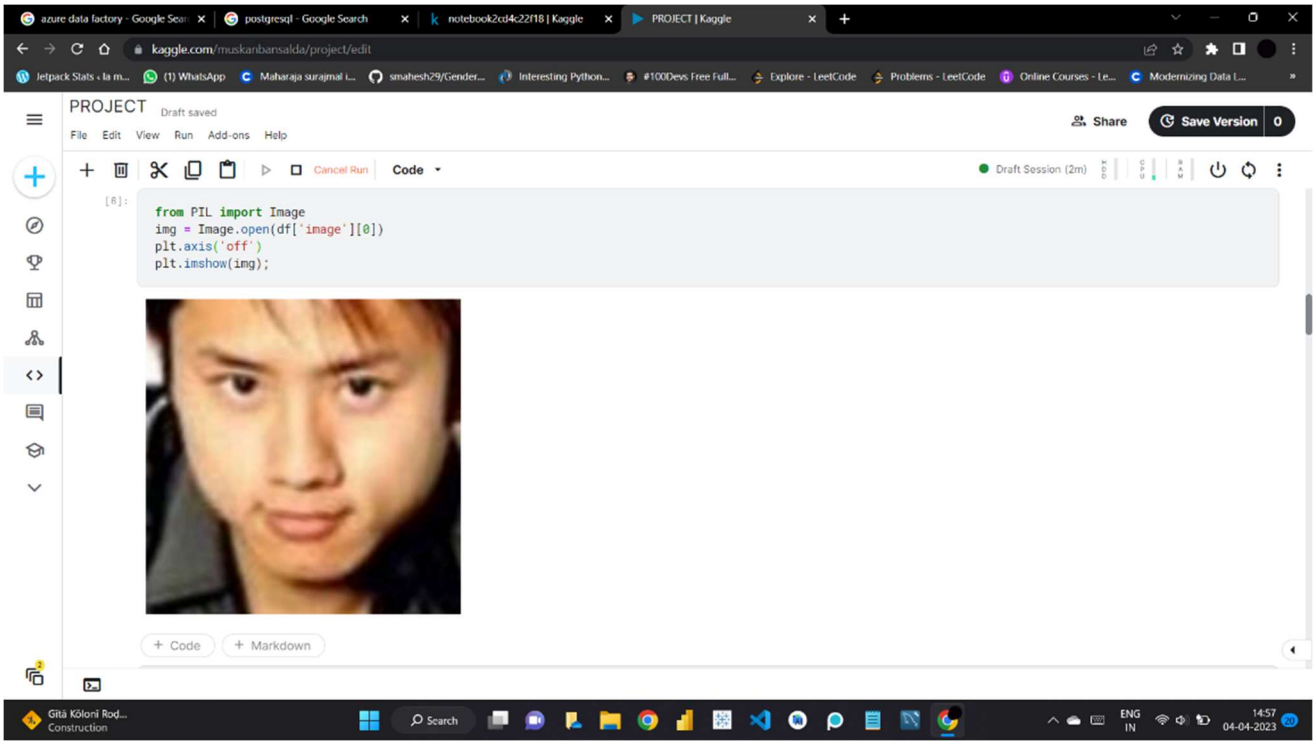

```
df = pd.DataFrame()
df['image'], df['age'], df['gender'] = image_paths, age_labels, gender_labels
df.head()

gender_dict = {0:'Male', 1:'Female'}
df.head()
```

[7]:

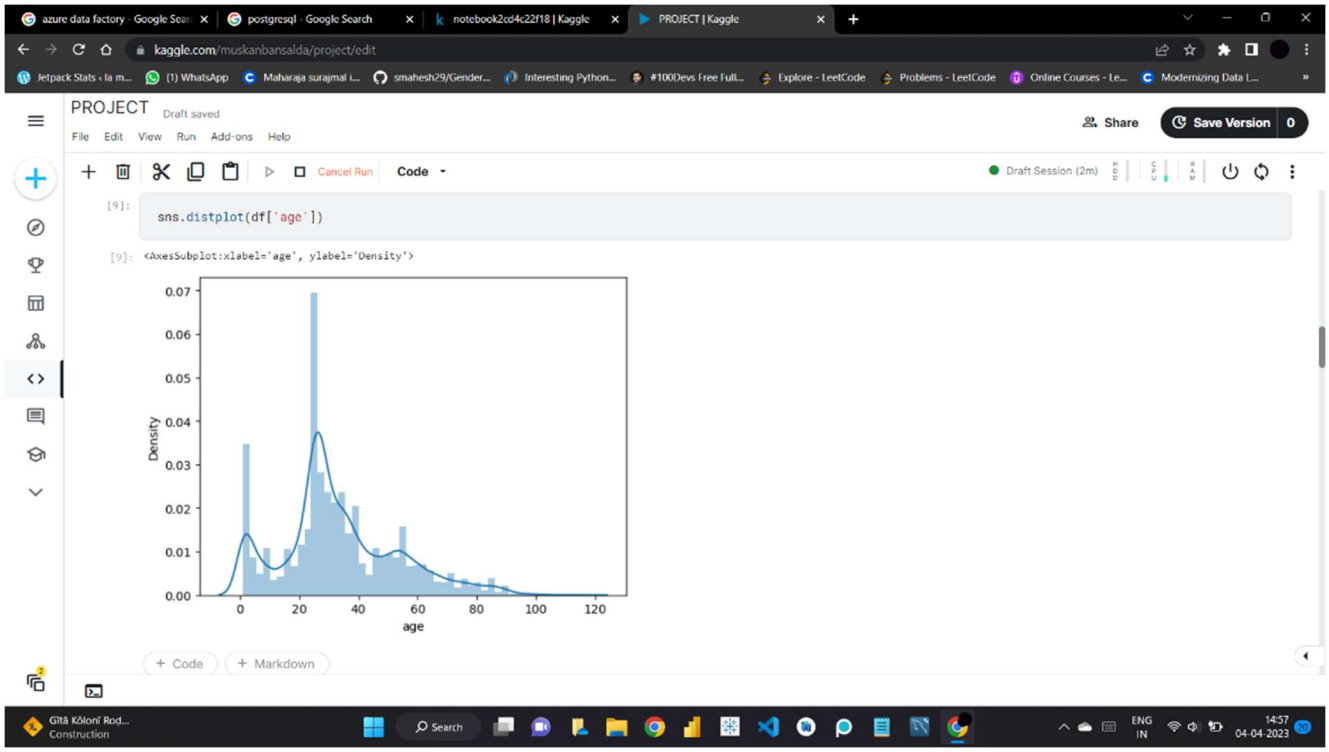
	image	age	gender
0	/kaggle/input/utkface-new/UTKFace/26_0_2_20170...	26	0
1	/kaggle/input/utkface-new/UTKFace/22_1_1_20170...	22	1
2	/kaggle/input/utkface-new/UTKFace/21_1_3_20170...	21	1
3	/kaggle/input/utkface-new/UTKFace/28_0_0_20170...	28	0
4	/kaggle/input/utkface-new/UTKFace/17_1_4_20170...	17	1

```
from PIL import Image
```

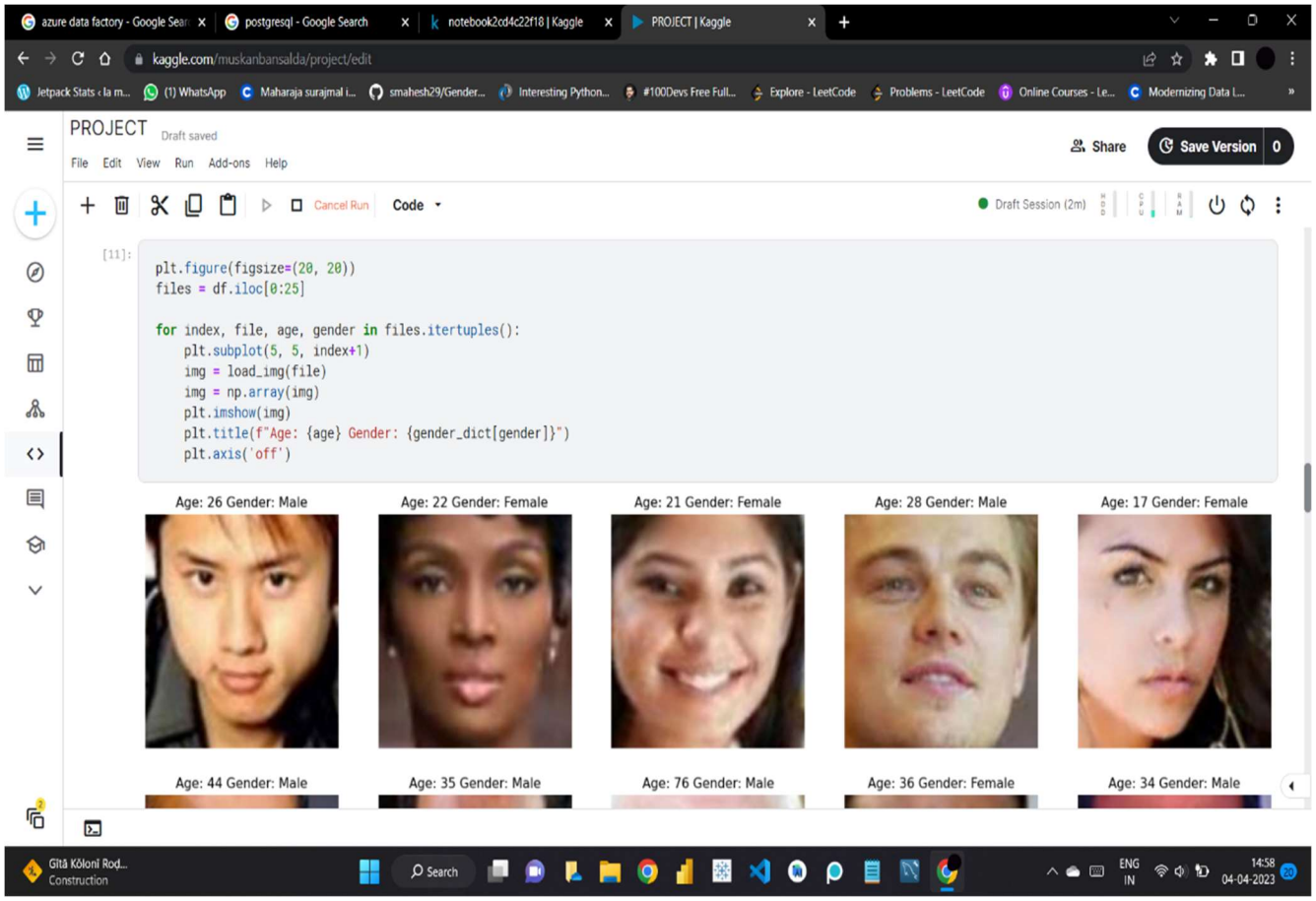
```
img = Image.open(df['image'][7])
plt.axis('off')
plt.imshow(img);
```




```
sns.distplot(df['age'])
```




```
sns.countplot(df['gender'])
plt.figure(figsize=(20, 20))
files = df.iloc[0:25]

for index, file, age, gender in files.itertuples():
    plt.subplot(5, 5, index+1)
    img = load_img(file)
    img = np.array(img)
    plt.imshow(img)
    plt.title(f"Age: {age} Gender: {gender_dict[gender]}")
    plt.axis('off')
```

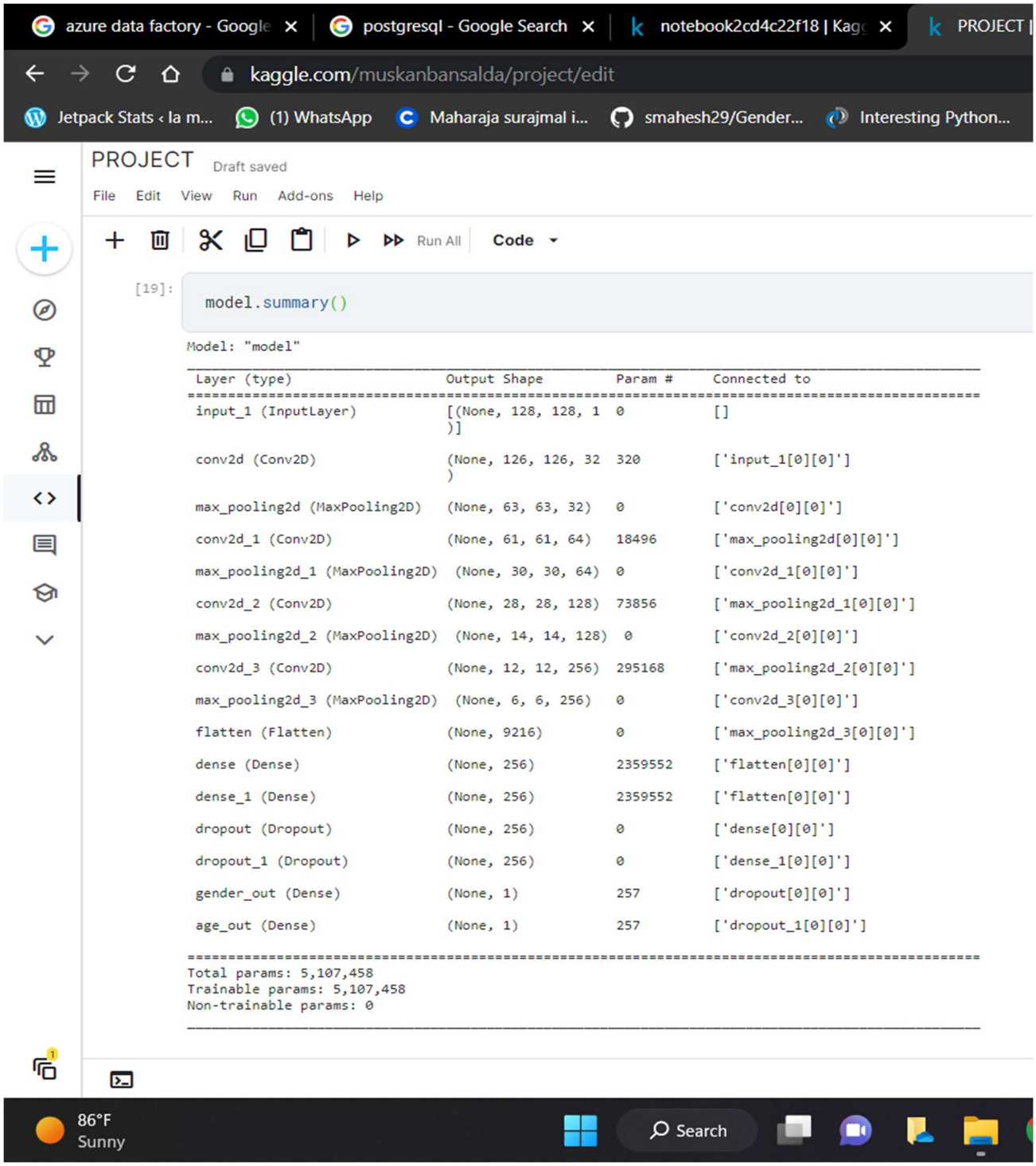



```
def extract_features(images):
    features = []
    for image in tqdm(images):
        img = load_img(image, grayscale=True)
        img = img.resize((128, 128), Image.ANTIALIAS)
        img = np.array(img)
        features.append(img)

    features = np.array(features)
    # ignore this step if using RGB
    features = features.reshape(len(features), 128, 128, 1)
    return features
```

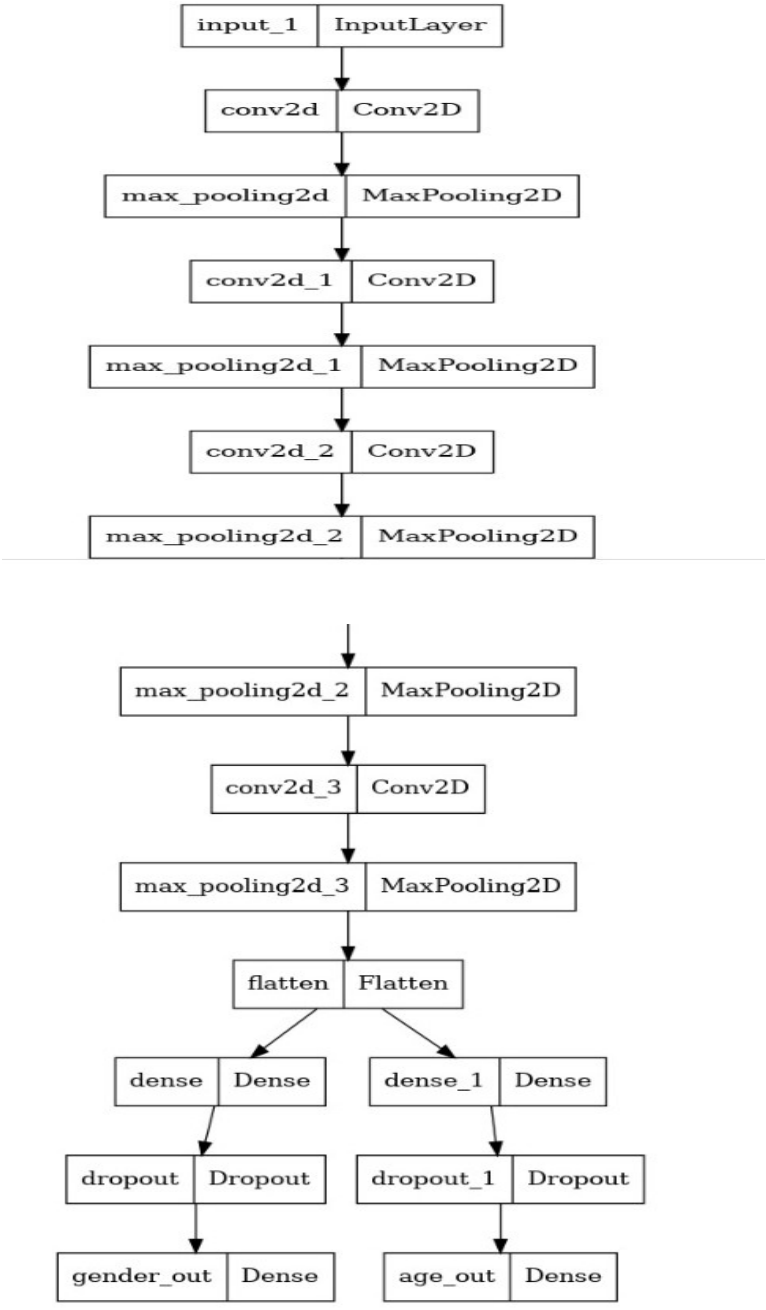
```
X = extract_features(df['image'])
X.shape
```

```
X = X/255.0


y_gender = np.array(df['gender'])
y_age = np.array(df['age'])



input_shape = (128, 128, 1)



inputs = Input((input_shape))
# convulational layers
conv_1 = Conv2D(32, kernel_size=(3, 3), activation='relu') (inputs)
maxp_1 = MaxPooling2D(pool_size=(2, 2)) (conv_1)
conv_2 = Conv2D(64, kernel_size=(3, 3), activation='relu') (maxp_1)
maxp_2 = MaxPooling2D(pool_size=(2, 2)) (conv_2)
conv_3 = Conv2D(128, kernel_size=(3, 3), activation='relu') (maxp_2)
maxp_3 = MaxPooling2D(pool_size=(2, 2)) (conv_3)
conv_4 = Conv2D(256, kernel_size=(3, 3), activation='relu') (maxp_3)
maxp_4 = MaxPooling2D(pool_size=(2, 2)) (conv_4)

flatten = Flatten() (maxp_4)

# fully connected layer
dense_1 = Dense(256, activation='relu') (flatten)
dense_2 = Dense(256, activation='relu') (flatten)

dropout_1 = Dropout(0.3) (dense_1)
dropout_2 = Dropout(0.3) (dense_2)

output_1 = Dense(1, activation='sigmoid', name='gender_out') (dropout_1)
output_2 = Dense(1, activation='relu', name='age_out') (dropout_2)

model = Model(inputs=[inputs], outputs=[output_1, output_2])

model.compile(loss=['binary_crossentropy', 'mae'], optimizer='adam',
metrics=['accuracy'])
```

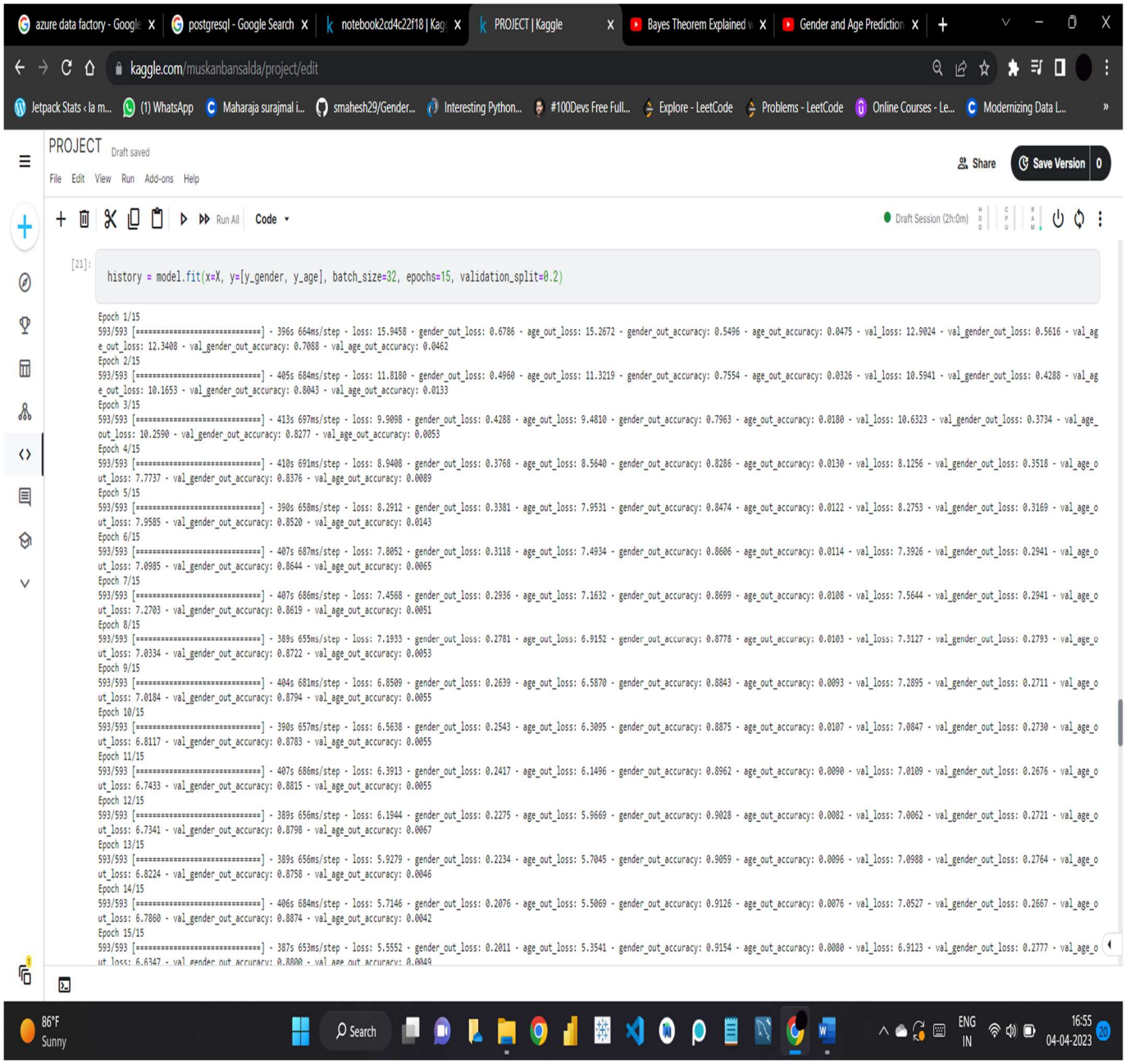
model.summary()



```
from tensorflow.keras.utils import plot_model
plot_model(model)
```

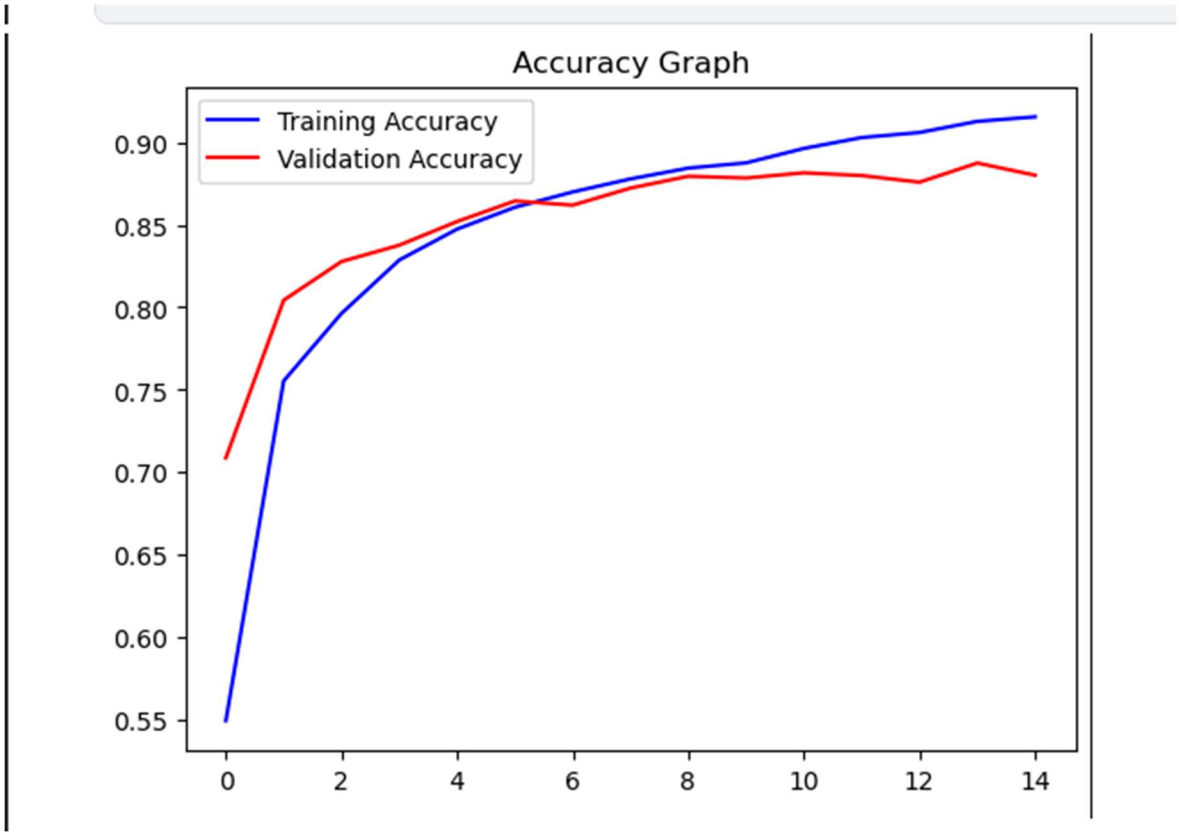


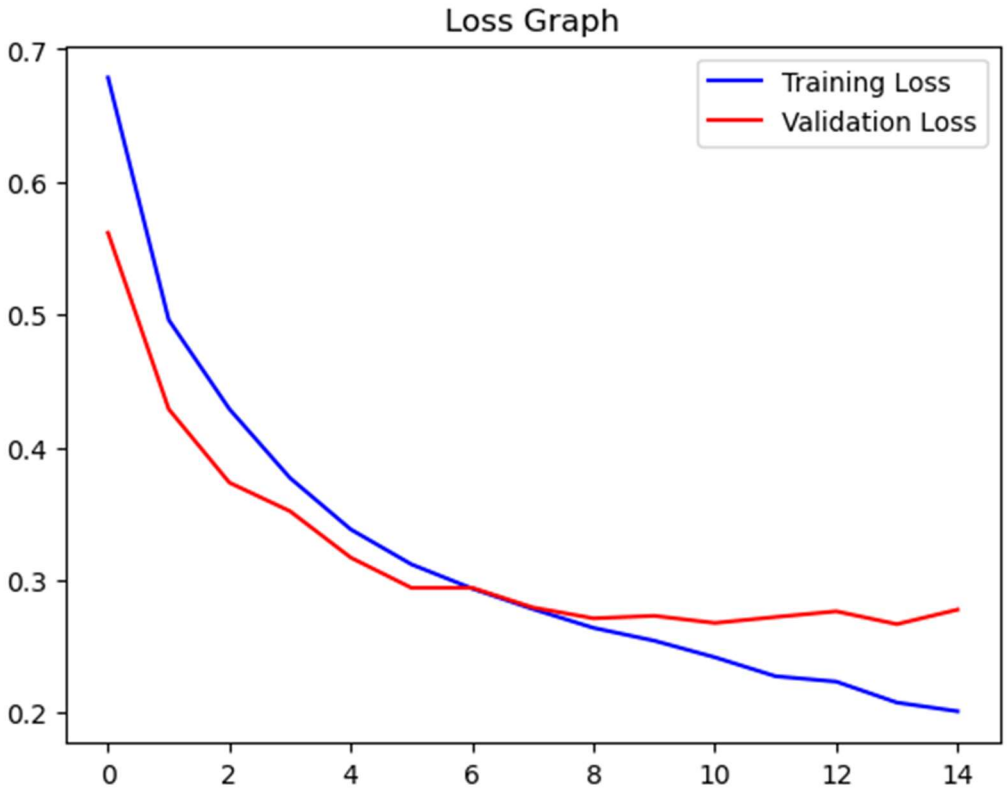

```
history = model.fit(x=X, y=[y_gender, y_age], batch_size=32, epochs=15, validation_split=0.2)
```




```
acc = history.history['gender_out_accuracy']
val_acc = history.history['val_gender_out_accuracy']
epochs = range(len(acc))

plt.plot(epochs, acc, 'b', label='Training Accuracy')
plt.plot(epochs, val_acc, 'r', label='Validation Accuracy')
plt.title('Accuracy Graph')
plt.legend()
plt.figure()
```

```
loss = history.history['gender_out_loss']
val_loss = history.history['val_gender_out_loss']

plt.plot(epochs, loss, 'b', label='Training Loss')
plt.plot(epochs, val_loss, 'r', label='Validation Loss')
plt.title('Loss Graph')
plt.legend()
plt.show()
```






```
loss = history.history['age_out_loss']
val_loss = history.history['val_age_out_loss']
epochs = range(len(loss))

plt.plot(epochs, loss, 'b', label='Training Loss')
plt.plot(epochs, val_loss, 'r', label='Validation Loss')
plt.title('Loss Graph')
plt.legend()
plt.show()
```

```
image_index = 1000
print("Original Gender: ", gender_dict[y_gender[image_index]], "Original
Age: ", y_age[image_index])
#predict from model
pred = model.predict(X[image_index].reshape(1, 128, 128, 1))
pred_gender = gender_dict[round(pred[0][0][0])]
pred_age = round(pred[1][0][0])
print("Predicted Gender: ", pred_gender, "Predicted Age: ", pred_age)
plt.axis('off')
plt.imshow(X[image_index].reshape(128, 128), cmap='gray');
```

```
Original Gender:  Male Original Age:  43
1/1 [==============================] - 0s 36ms/step
Predicted Gender:  Male Predicted Age:  43
```

```
image_index = 5
print("Original Gender: ", gender_dict[y_gender[image_index]], "Original
Age: ", y_age[image_index])
#predict from model
pred = model.predict(X[image_index].reshape(1, 128, 128, 1))
pred_gender = gender_dict[round(pred[0][0][0])]
pred_age = round(pred[1][0][0])
print("Predicted Gender: ", pred_gender, "Predicted Age: ", pred_age)
plt.axis('off')
plt.imshow(X[image_index].reshape(128, 128), cmap='gray');
```

```
Original Gender:  Male Original Age:  22
1/1 [==============================] - 0s 119ms/step
Predicted Gender:  Male Predicted Age:  20
```

```
image_index = 500
print("Original Gender: ", gender_dict[y_gender[image_index]], "Original
Age: ", y_age[image_index])
#predict from model
pred = model.predict(X[image_index].reshape(1, 128, 128, 1))
pred_gender = gender_dict[round(pred[0][0][0])]
pred_age = round(pred[1][0][0])
print("Predicted Gender: ", pred_gender, "Predicted Age: ", pred_age)
plt.axis('off')
plt.imshow(X[image_index].reshape(128, 128), cmap='gray');
```

```
Original Gender:  Female Original Age:  26
1/1 [==============================] - 0s 35ms/step
Predicted Gender:  Female Predicted Age:  29
```

DRAWBACK

Since, we only did gender detection using classification for male and female that is, 0 and 1.

We did not inlude transgender which was a major drawback of this project, apart from this we could not include a feature that can detect age and gender straight from the camera, we first need to take a picture and run the program and only then the detection will take place.

CONCLUSION

Name of our project is Age and Gender Detection, we used machine learning and neural network subjects and used Python language since we wanted to learn Python and make a project using Python, as for ML and NN, we find these subjects very interesting and wanted to study a bit more on these topics and hence we finally chose this project.

The aim for this project was to find the approximate age and gender from a picture when given as an input.

We had to use classification and regression model for the detection of gender and age respectfully. We used a lot of libraries like pandas, numpy, os, matplotlib, seaborn, warnings, tensorflow, keras. We also used RELU function. Dataset was imported from kaggle datasets wince it is a big datahouse.

One can take their picture and run the code, you will get an approximate age and gender, the accuracy is 0.8 and we can increase it by inserting more datasets and training the algo more.

BIBLIOGRAPHY

1. www.google.com
2. www.geeksforgeeks.org/machine-learning/
3. https://www.javatpoint.com/machine-learning
4. https://www.hackersrealm.net/post/gender-and-age-prediction-using-python
5. https://www.javatpoint.com/regression-vs-classification-in-machine-learning