Insightful Identity Analysis: Detecting Age, Gender, and Ethnicity

Team Members:
Sarvagya Kaushik
Kunal Sharma
Vansh

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

1. For those working in the fields of computer vision and facial recognition, the UTKFace dataset is a helpful resource.
2. The unique characteristics of this dataset and its potential applications across numerous domains serve as the inspiration for training on it.

Reasons UTKFace is revolutionary in the field of face data sets:

1. Age Diversity
2. Real-World Challenges
3. Annotations for Gender and Ethnicity
4. Multi-Task Learning

Benchmark:

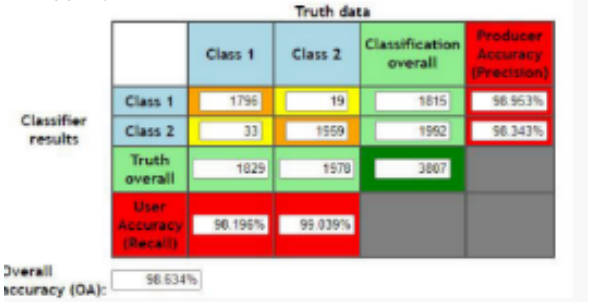
1. Contribute to the advancement of state-of-the-art techniques in facial analysis.
2. Compare our algorithm against existing state-of-the-art methods using the UTKFace dataset.
3. Understand how your solutions measure up in terms of accuracy, efficiency, and robustness.

Literature Survey

1. GRA_Net:

- Consists of multiple layers, each containing an attention block.
- Each attention block combines features from the previous layer with attention weights to produce refined feature representation.
- Gating mechanism dynamically controls the influence of attention on the feature at each layer.
- GRA_Net is trained using standard deep learning techniques, such as backpropagation and gradient descent.

The classification accuracies achieved by the proposed GRA_Net model for UTKFace datasets was found to be 99.2%.



Thorough comparison with various alternative models, GRA_Net has unequivocally demonstrated its supremacy by consistently yielding superior results.

Model	Gender(%)	Age(%)
Facenet	91.2	56.9
Finetuned Facanet (FFNet)	96.1	64
MTCNN	98.23	70.1
RAN (Wang et al. (2017))	97.5	85.4
Proposed model	**99.2**	**93.7**

Literature Survey

2. Feature Extraction based Face Recognition, Gender and Age Classification algorithm

- The algorithm yields good results with small training data.
- Steps involved:
  - Preprocessing:
    - Color Conversion
    - Noise Reduction
    - Edge detection
  - Feature Extraction:
    - Computation: Ratios* are calculated.
    - Gender Classification: Naive Bayes
  - Training on the dataset
    - Artificial Neural Network(ANN): carried out in two parts:
      - Feed-forward path
      - Feedback path
    - Back Propagation

Performance Analysis

Gender	Sample size	Correctly Labeled(CL)	Correct Rate(CR)	Total CR
Male	40	38	95%	94.82%
Female	18	17	94.44%	

Algorithm	AG	Sample size	CL	CR	Total CR
FEBFRGAC	Y	28	25	89.3%	89.65%
	M	20	18	90%	
	O	10	09	90%	
CAGBFF	Y	44	37	84.4%	78.49%
	M	32	25	78.1%	
	O	17	11	64.7%	

*Ratios that were taken into account were left-to-right eye distance upon eye-to-nose distance, left-to-right eye distance upon eye-to-lip distance, eye-to-nose distance upon eye to chin distance, and eye-to-nose distance upon eye-to-lip distance.

Dataset Description

The UTKFace dataset is a fairly large dataset with over 20,000 face images with annotations of age, gender and ethinicity.

The subjects covered in the dataset consisted of people ranging from the age of 0 to 116 years old, over 4 ethnicities.

Use Cases: Face detection, Age estimation, Age progression/regression,Landmark localization, etc.

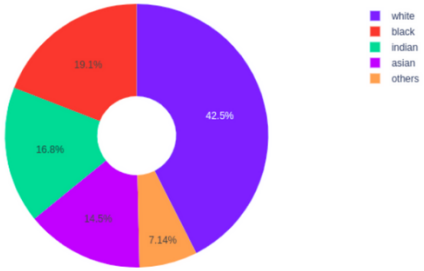


Figure 12. Race Distribution

- Data was collected from a wide number of sources across the internet
- Model may have slightly higher bias with context to race

Dataset Description

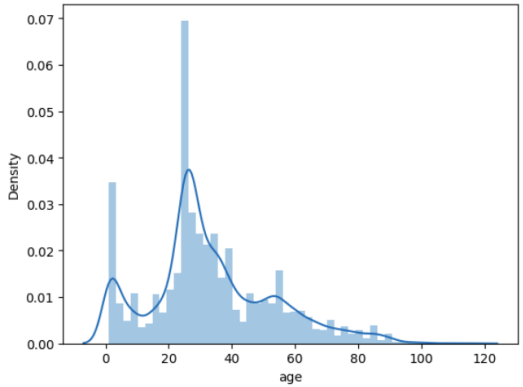
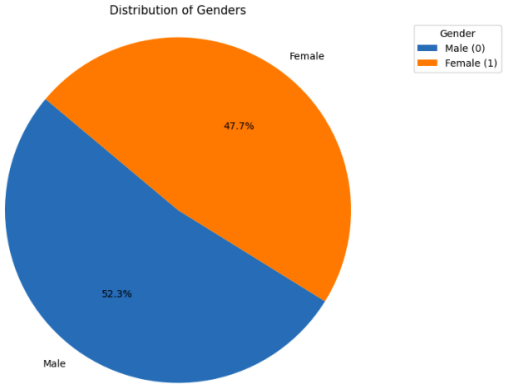


Figure 14. Age distribution plot

- The plot shows us that the data is normally distributed.
- Surface Observation shows us that the data is skewed to left ie most people in hf the dataset are less than 40 years old



- Data spread was fairly symmetrical with regard to gender
- Not from IIIT D !

Pre-processing

- Effective Preprocessing is critical for ML tasks.
- Extracted features from image path name.
- Resizing is essential for overall quality and adaptability: Resized from 128 x 128 pixels to 28 x 28 pixels
- Converted from RGB Scale to grayscale to reduce data complexity and processing resource requirements
- Normalized the pixel values.

Methodology

Models Used:-

- Logistic Regression for gender prediction
- K-Nearest Neighbours for gender prediction
- Naive Bayes for gender, ethnicity and age prediction
- SVM for gender, ethnicity and age prediction
- Random Forest for gender, ethnicity and age prediction
- CNN for gender, ethnicity and age prediction

Logistic Regression

- **Splitting the dataset** into training and test set. For example, we can have 70% data for training, 20% for validation and the rest for testing.
- Creation of model
- $P(Y = 1) = 1 / 1 + e^{-(\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp)}$, where $P(Y = 1)$ is the probability of the event occurring and $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ are coefficients that represent the relationship between the independent variables X1, X2, . . . , Xp and the probability of the event.
- Model Performance
- Tuning hyperparameters and regularization
- Gender prediction

K-Nearest Neighbours

- **Splitting the dataset** into training and test set. For example, we can have 70% data for training, 20% for validation and the rest for testing.
- **Creation of a classifier model** based on the KNN algorithm which is a non-parametric, instance-based algorithm that classifies data points based on their similarity to the k-nearest neighbors in the training data.
- Model training
- Model performance
- Tuning hyperparameters and regularization
- Gender prediction

Naive Bayes

- Since we are using Gaussian Naive Bayes 0architecture we had to change the RGB scale to Grey scale and apply the image transformation and we also flattened the image
- We then applied PCA on the flattened image to reduce the dimensionality of the image
- Naive Bayes model :

$$P(Y|X_1, X_2, \ldots, X_n) = \frac{P(Y) \cdot P(X_1|Y) \cdot P(X_2|Y) \cdot \ldots \cdot P(X_n|Y)}{P(X_1) \cdot P(X_2) \cdot \ldots \cdot P(X_n)}$$

- Here Xs are the features while Y is label
- Performance Metrics:
  - **Accuracy on Gender**: 0.79
  - **Accuracy on Ethnicity**: 0.56
  - **Accuracy on Age**: 0.37
- We used a number of diiferent components in PCA and best accuracy was acieved in 100 components

Support Vector Machine

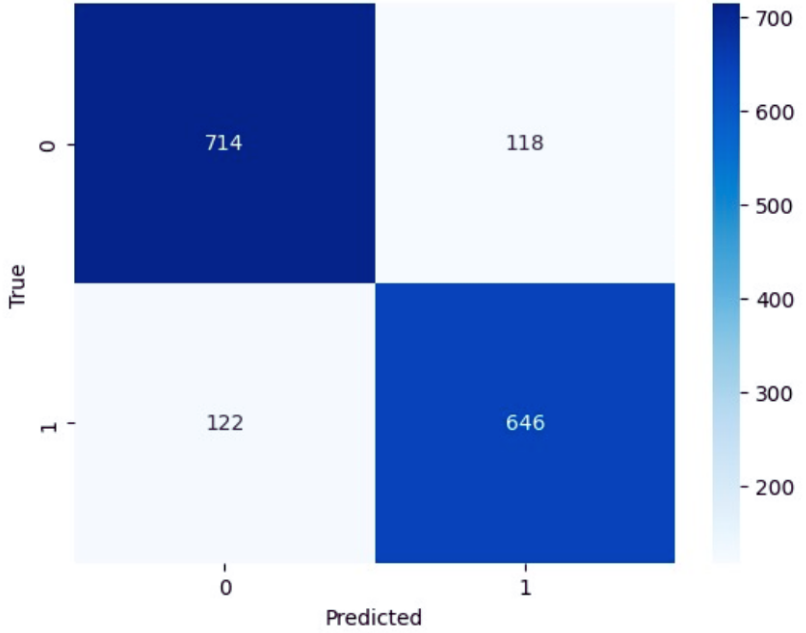
- The SVM model was trained by flattening the image dimensions and we used the sklearn library to train our model.
- We used RBF kernel for our model. It performed better than the linear kernel.
- Performance Metrics:

  - **Accuracy on Gender**: 0.83
  - **Accuracy on Ethnicity**: 0.69
  - **Mean Absolute Error(MAE) on Age**: 11.0

- SVM performed quite well relative to other previous models for gender and ethnicity prediction.
- Classic ML models failed to provide good prediction for age.

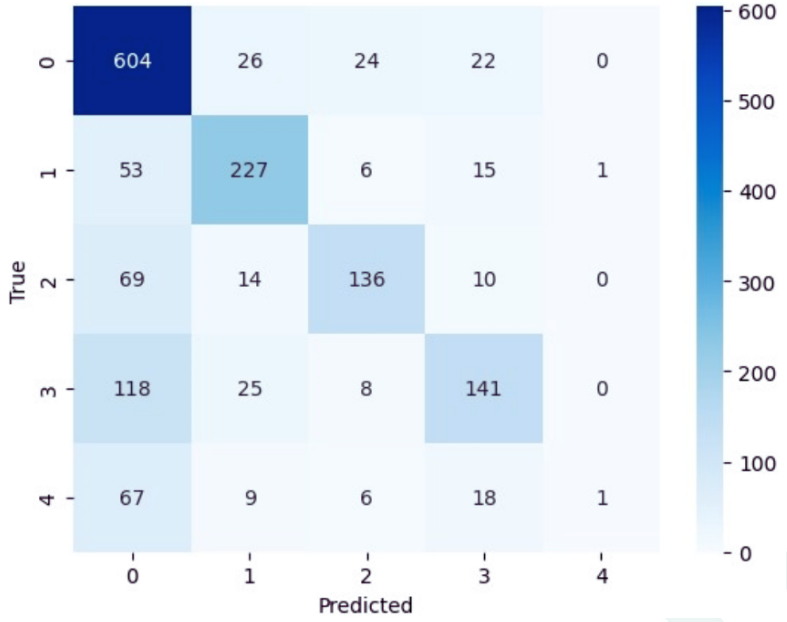
Support Vector Machine

Gender Prediction



Accuracy: 0.83
Precision: 0.8398
Recall: 0.8396
F1-Score: 0.8397

Ethnicity Prediction



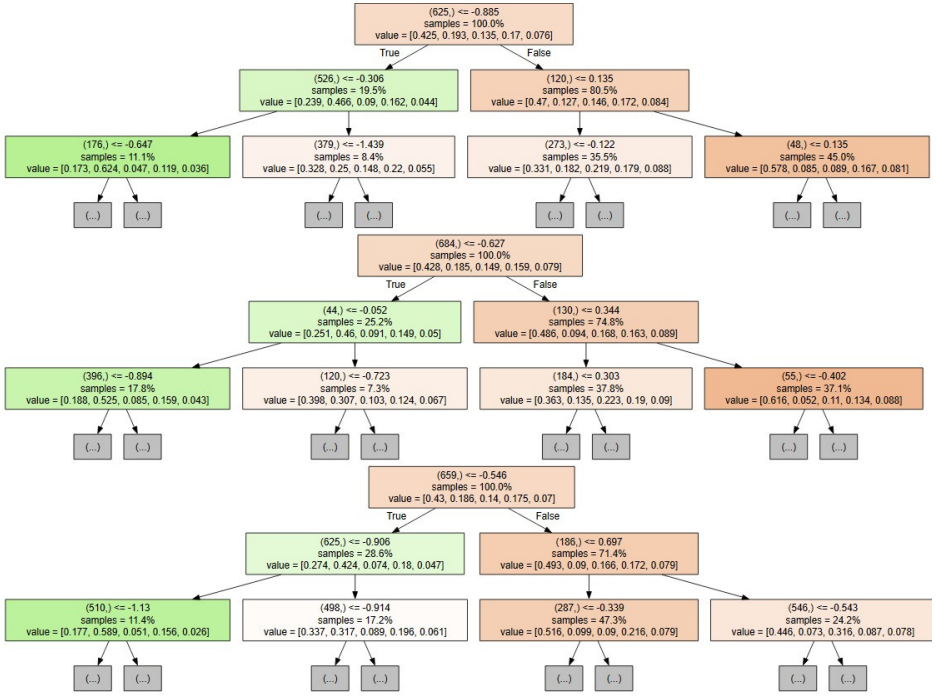
Accuracy: 0.69
Precision: 0.6714
Recall: 0.5436
F1-Score: 0.5529

Random Forests

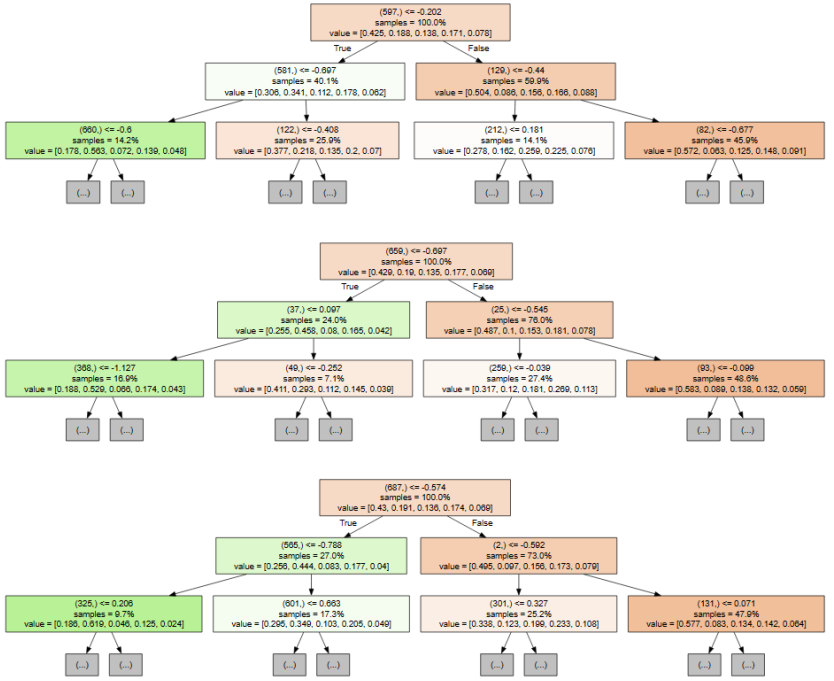
- The model was trained by flattening the image dimensions and we used sklearn library to train our model.
- We normalized the features before feeding them to our model for training. Following this prepossessing helped us gain the maximum accuracy possible with the random forest model.
- We used 100 estimators for our ensemble learning process and used gini impurity as criterion.
- Performance Metrics:
    - **Accuracy on Gender**: 0.80
    - **Accuracy on Ethnicity**: 0.61
- Age prediction was not worth noting for this model
- It performed quite well for ethnicity prediction in comparison with logistic regression and naive bayes model.

Random Forests

First 3 Decision Trees



Gender Prediction



Ethnicity Prediction

Convolution Neural Networs

- **Splitting the dataset** into training and test set. For example, we can have 70% data for training, 20% for validation and the rest for testing.

- Since we were using CNN, no flattening was required to convert and no conversion to grayscale was required either and ADAM optimiser was used.

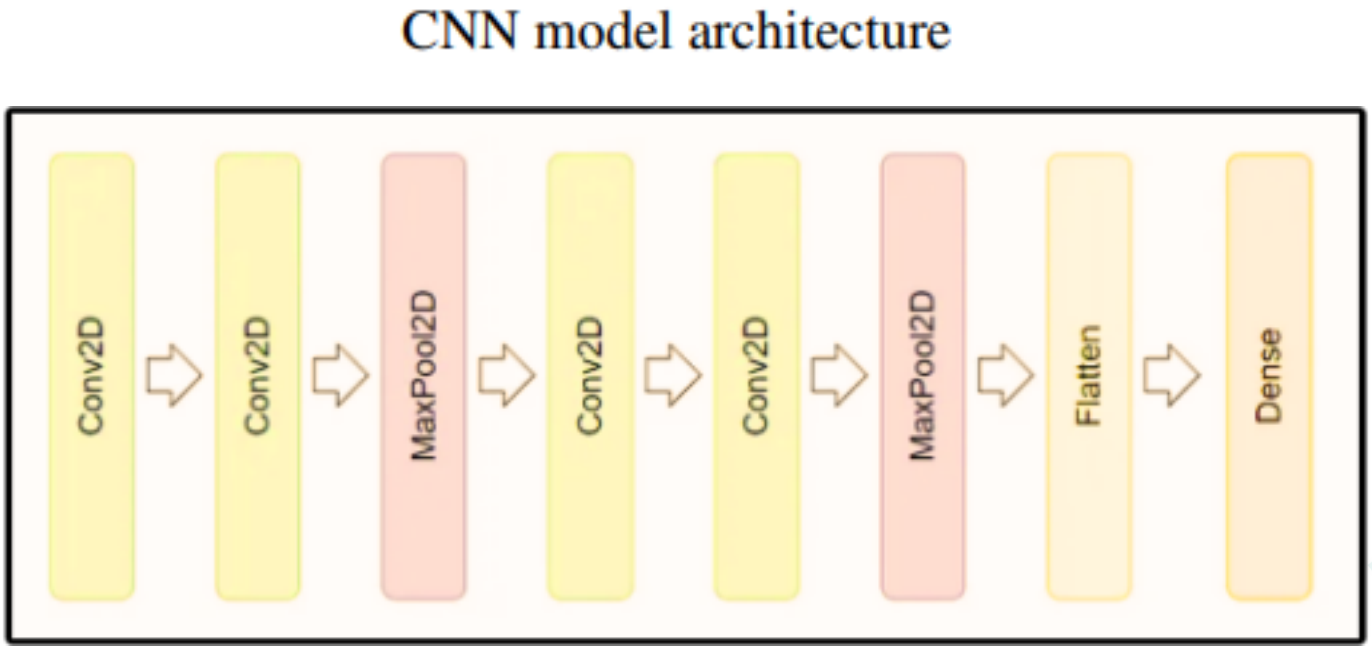
- Performance Metrics:

    - **Accuracy on Gender**: 0.87
    - **Accuracy on Ethnicity**: 0.72
    - **Accuracy on Age**: 0.47

- We divided our training data into batch sizes of 32. Given the complexity of our architecture, it converged in 6 iterations only.

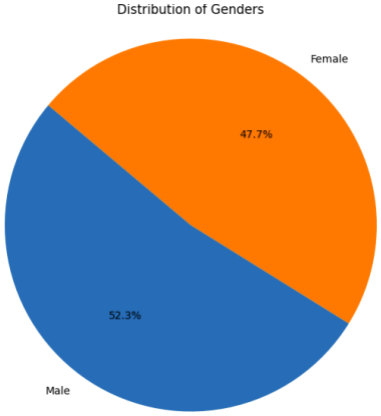
- This wasn't possible with other models since they weren't as complex as CNN architecture.

Convolution Neural Networs



Results and Analysis

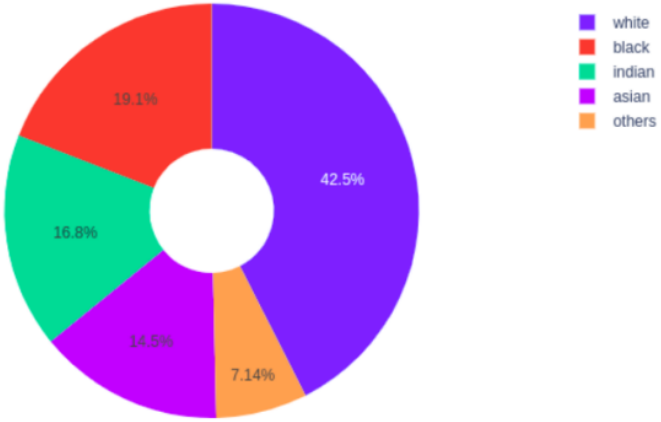
Gender Distribution



The above figure gives a visualization of gender distribution. We can see that the percentage of the male population is slightly greater than females but the difference is minor. It's not capable of creating high bias.

Results and Analysis

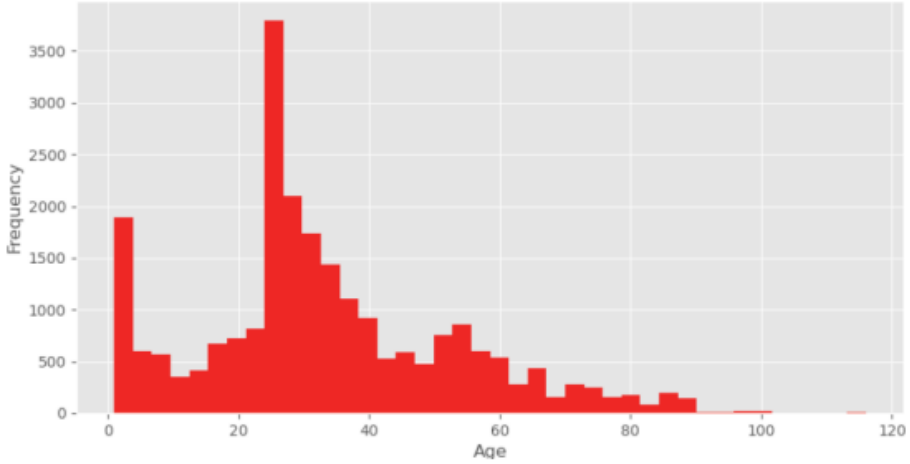
Ethnicity Breakdown



The above figure gives visualization for ethnicity distribution in our data set. Our data set majorly consists of images of white ethnicity with 42.5 percent. It is followed by black at 19.1 percent, Indian with 16.8 percent and Asian with 14.5 percent. Rest of the population are categorized by others.

Results and Analysis

Age Distribution



The above figure gives visualization for age distribution in our data set. From surface observation, we can see that the data is skewed to the left. Thus our data set mainly consists of a population less than 40 years. From the figure we can also see that the data is normally distributed.

Results and Analysis

Logistic Regression

Performance Metrics

The model was trained using a batch size of 32, binary cross-entropy as the loss function, and stochastic gradient descent (SGD) as the optimization algorithm. After 10 epochs of training, we achieved an accuracy of 80

The following statistics summarize the model's performance:

- **Training Loss: 0.3654**
- **Test Loss: 0.3598**
- **Test Accuracy: 84.41**

These results indicate that our logistic regression model performs well in classifying images into male and female categories. The relatively low training and test losses suggest that the model effectively minimized the classification error, and the test accuracy of 84.41 percent demonstrates its ability to correctly classify the gender of previously un□seen images

Results and Analysis

K-Nearest Neighbours

Performance Metrics

The model was trained by flattening the image dimensions into one dimension and setting the k parameter to 20.

The following statistics and the classification report summarize the model's performance:

- **Accuracy: 0.7344**

Class	Precision	Recall	F1-Score	Support
0 (Male)	0.70	0.85	0.77	2468
1 (Female)	0.79	0.61	0.69	2273
Accuracy			0.73	4741
Macro Avg	0.75	0.73	0.73	4741
Weighted Avg	0.74	0.73	0.73	4741

1. True Positive: 2095
2. False Negative: 373
3. False Positive: 886
4. True Negative: 1387

Results and Analysis-2

Naive Bayes

Performance Metrics

The model was trained by changing the RGB scale to grayscale and applying PCA on the flattened image to reduce the dimensionality of the image. The best accuracy was achieved with 100 components in PCA.

The following statistics summarize the model's performance:

- **Accuracy on Gender**: 0.79
- **Accuracy on Ethnicity**: 0.56
- **Accuracy on Age**: 0.37

The model works fairly well for gender and ethnicity prediction. In fact it has better gender prediction than accuracy than KNN.

Results and Analysis-2

Support Vector Machine

Performance Metrics

The model was trained by using the sklearn library after flattening the image.

RBF kernel was used as it gave better accuracy

The following statistics summarize the model's performance:

- **Accuracy on Gender**: 0.83
- **Accuracy on Ethnicity**: 0.69
- **Mean Absolute Error(MAE) on Age**: 11.0

Confusion matrix for gender prediction ---> $\begin{bmatrix} 714 & 118 \\ 122 & 646 \end{bmatrix}$

SVM performed quite well relative to other previous models for gender and ethnicity prediction.

Results and Analysis-2

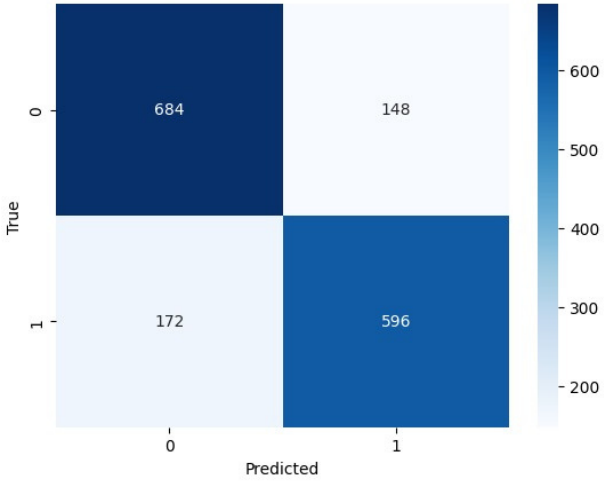
Random Forests

Performance Metrics

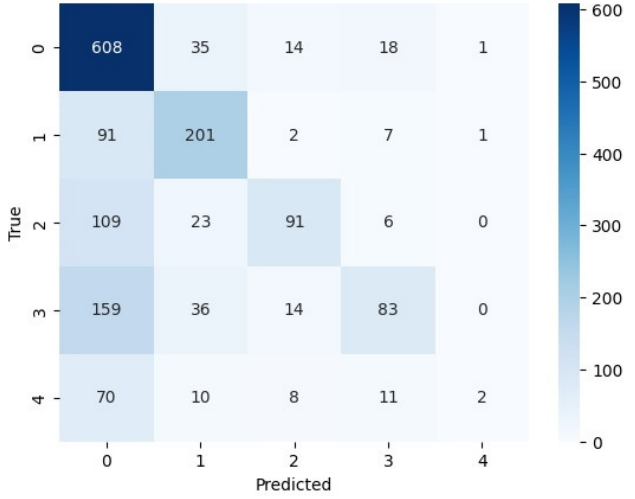
Features were normalized before feeding them to our model for training to gain the maximum accuracy possible with the random forest model.

100 estimators were used for our ensemble learning process and gini impurity was used as the criterion.

The following statistics summarize the model's performance:



Gender Prediction
Accuracy: 0.8
Precision: 0.8
Recall: 0.799
F1 score:0.799



Ethnicity Prediction
Accuracy: 0.61
Precision: 0.623
Recall: 0.453
F1 score:0.463

Results and Analysis-2

Convolutional Neural Networks

Performance Metrics

We didn't gray scale our image and we trained our model in 3 channels i.e RGB. We also normalized the RGB values

This helped us get the best result in comparision with other models

The following statistics summarize the model's performance:

- **Accuracy on Gender**: 0.87
- **Accuracy on Ethnicity**: 0.72
- **Accuracy on Age**: 0.47

Given the complexity of CNN's architecture, it converged in 6 iterations only which wasn't possible for other relative less complex models

CNN gave the best results for prediction of age, gender and ethnicity.

Conclusion

In our study of image data, six models were employed to predict age, ethnicity, and gender. While binary classification, particularly for gender, saw strong performance across multiple models, challenges arose in multi-class classification (e.g., ethnicity) and regression tasks like age prediction. CNN architecture demonstrated rapid learning, excelling in all prediction labels and outperforming others in multi-class scenarios. SVM performed well but was computationally expensive, while Random Forest showed comparable results but lagged behind SVM. KNN, logistic regression, and custom architectures struggled, with Tiny VGG net standing out as the most effective. CNN's speed and simplicity led to convergence in just 5 iterations, far surpassing other models requiring 20 iterations with the same batch size and optimizer.

Timeline Progress (Till Mid Sem)

August 30 - September 30	September 30 - October 15	October 15 - November 15	November 15 - End sems
Data Analysis • Visualizations • Insights	**Data Cleaning** • RGB to Grayscale • Rescaling • Normalisation	**Model Preparation** • Logistic Regression • K-Nearest Neighbours • More Models	**Optimization and Finalization** • Code cleanup • Optimization • Final adjustments

Timeline Progress (Updated)

August 30 - September 30	September 30 - October 15	October 15 - November 15	November 15 - End sems
Data Analysis • Visualizations • Insights	**Data Cleaning** • RGB to Grayscale • Rescaling • Normalisation	**Model Preparation** • Logistic Regression • K-Nearest Neighbours • More Models (NB, SVM, RF, CNN)	**Optimization and Finalization** • Code cleanup • Optimization • Final adjustments

Individual Contribution

- Sarvagya Kaushik : Training and Testing Models, Data Visualisation, Code Optimisation (CNN, RF, SVM)
- Kunal Sharma : Training and Testing Models, Results and Analysis, Code Optimisation (Naive Bayes, Logistic Regression, KNN)
- Vansh : Training and Testing Models, Hyper parameter Testing, Code Optimisation (CNN, RF, SVM)