Jacob Redden ([jjr0017@auburn.edu](mailto:jjr0017@auburn.edu), Coordinator)

Yifeng Wu ([yzw0134@auburn.edu](mailto:yzw0134@auburn.edu))

Robert Smith (res0059@auburn.edu)

COMP 6600 Artificial Intelligence

Group 17

04/18/2024

Project Progress Report

Group 17 has been tasked with implementing 5 different Artificial Intelligence models. The database chosen to train these models is the CelebA-attrs dataset on huggingface. This is a large database of celebrity headshots. The pictures in this database are 178x218 resolution pictures with RGB encoding. The dataset contains 167k training samples, 19.9k testing samples, and 20k validation samples. Each sample has 40 boolean attributes identified.

The group has decided to train these Artificial Intelligence models to predict hair color. Each model will use the RGB pixels of each picture as inputs to the model and will try to classify the pictures into different classes of hair. The four attributes of hair color provided in the database are Black\_Hair, Brown\_Hair, Blond\_Hair, and Gray\_Hair. Because hair color is being used for classification, if the database entry does not have one of these attributes set or if it has more than one set, that entry is thrown out and not used for training or testing. 36% of the picture entries in the database do not have one of these attributes set so the actual number of training samples is 107k.

Our group has chosen to use 3 main evaluation metrics to evaluate the effectiveness of our models in classifying these picture inputs into 4 different categories of hair color. The first metric is precision. This metric will describe for each hair color prediction the fraction of those predictions that are correct. However, this metric alone will not be able to describe the correctness of our model. If only one prediction was made for a classification and it was correct, the precision would be 100%. For this reason, the group also decided to use recall as a metric to evaluate our models. Recall is more focused on the truth data and describes how many of the positives in the truth data were correctly identified as positives by the model. It is important that precision and recall are used together because precision alone will not accurately portray the correctness of the model. If all the entries are predicted by the model to be one classification, that classification will have 100% recall. The last metric we will use is a confusion matrix. This matrix similarly describes True and False Positives and Negatives but does it in a way to describe where the model is potentially getting confused. For example, if the model is having a hard time differentiating between Black Hair and Brown Hair, a confusion matrix will be able to describe a lot of False Negatives in the Brown Hair column that are true Black Hair entries or vice versa. This will help the group identify strengths and weaknesses within each model.

The two models the team has started working on first are the Naive Bayes Classifier and the Linear Regression Classifier. A subsample of the data has been used to do efficient tuning and debugging of the dataset cleaning, pruning, and model implementation. In this subsample, we are using 10K training samples and 2K testing samples. Of these samples, they are further downselected to clean the data of any entries that do not have classification data for the attributes of the pictures we are trying to classify.

The metrics obtained from the preliminary run of the Naive Bayes Classifier are below in Table 1 and Table 2.

Table 1. Confusion Matrix of Naive Bayes Classifier

|  | Black Hair | Blond Hair | Brown Hair | Gray Hair |
| --- | --- | --- | --- | --- |
| Black Hair | 288 | 37 | 103 | 2 |
| Blond Hair | 20 | 222 | 28 | 5 |
| Brown Hair | 119 | 71 | 288 | 1 |
| Gray Hair | 24 | 35 | 14 | 22 |

Table 2. Precision and Recall Metrics for Naive Bayes Classifier

|  | Precision | Recall | Support |
| --- | --- | --- | --- |
| Black Hair | 64% | 67% | 430 |
| Blond Hair | 61% | 81% | 275 |
| Brown Hair | 67% | 60% | 479 |
| Gray Hair | 73% | 23% | 95 |
| Weighted Average | 65% | 64% | Total: 1279 |

The metrics obtained from the preliminary run of the Linear Regression Classifier are below in Table 3 and Table 4.

Table 3. Confusion Matrix of Linear Regression Classifier

|  | Black Hair | Blond Hair | Brown Hair | Gray Hair |
| --- | --- | --- | --- | --- |
| Black Hair | 284 | 12 | 117 | 17 |
| Blond Hair | 3 | 203 | 42 | 27 |
| Brown Hair | 60 | 39 | 342 | 38 |
| Gray Hair | 2 | 9 | 4 | 80 |

Table 4. Precision and Recall Metrics for Linear Regression Classifier

|  | Precision | Recall | Support |
| --- | --- | --- | --- |
| Black Hair | 81% | 66% | 430 |
| Blond Hair | 77% | 74% | 275 |
| Brown Hair | 68% | 71% | 479 |
| Gray Hair | 49% | 84% | 95 |
| Weighted Average | 73% | 71% | Total: 1279 |

To fairly compare these metrics, a random baseline model was created to evaluate the performance of each model and its learning potential. The metrics for the random baseline are provided in Table 5 and Table 6.

Table 5. Confusion Matrix of Random Baseline Classifier

|  | Black Hair | Blond Hair | Brown Hair | Gray Hair |
| --- | --- | --- | --- | --- |
| Black Hair | 111 | 111 | 97 | 111 |
| Blond Hair | 75 | 61 | 67 | 72 |
| Brown Hair | 133 | 111 | 114 | 121 |
| Gray Hair | 27 | 24 | 21 | 23 |

Table 6. Precision and Recall Metrics for Random Baseline Classifier

|  | Precision | Recall | Support |
| --- | --- | --- | --- |
| Black Hair | 32% | 26% | 430 |
| Blond Hair | 20% | 22% | 275 |
| Brown Hair | 38% | 24% | 479 |
| Gray Hair | 7% | 24% | 95 |
| Weighted Average | 30% | 24% | Total: 1279 |

The preliminary results for both the Naive Bayes Classifier and the Linear Regression Classifier look promising. There is lots of room for improvements with more training and better hyperparameters. However, the models are learning how to classify the pictures and are performing much better than the Random Baseline. The average precision and recall of the Naive Bayes Classifier are 65% and 64%, respectively, for the Linear Regression Classifier they are 73% and 71%, respectively, and for the Random Baseline they are 30% and 24%, respectively. The Linear Regression is out performing the Naive Bayes slightly and it seems based on the Confusion Matrix for the Naive Bayes, the model is having a hard time distinguishing Black and Brown Hair, and classifying Gray Hair in general.

So far, the main challenge the group has faced is resources. 100K+ samples of 178x218 resolution pictures is a lot of data that takes a lot of time to train. We plan to try doing some overnight runs of data on some better hardware. We also plan to break up the training into batches and train the models in stages if possible.

Loss training curves were created with the preliminary models using the log-loss function. This loss should decrease as the models become more accurate in their testing after more training occurs. The log-loss plots for the Naive Bayes Classifier and the Linear Regression Classifier are show in Figure 1 and Figure 2.

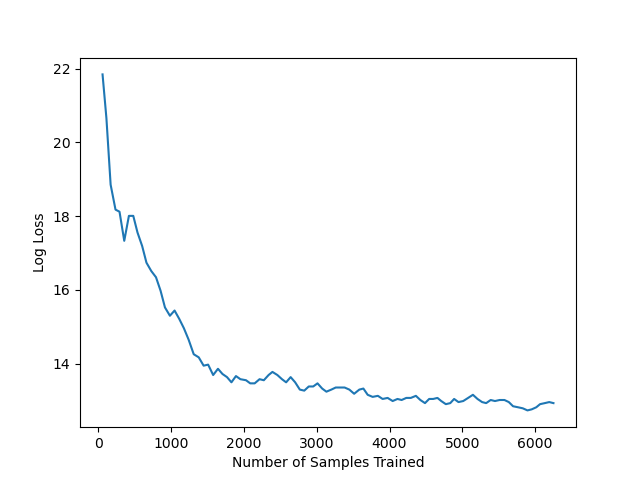


Figure 1. Log-Loss for Naive Bayes Classifier

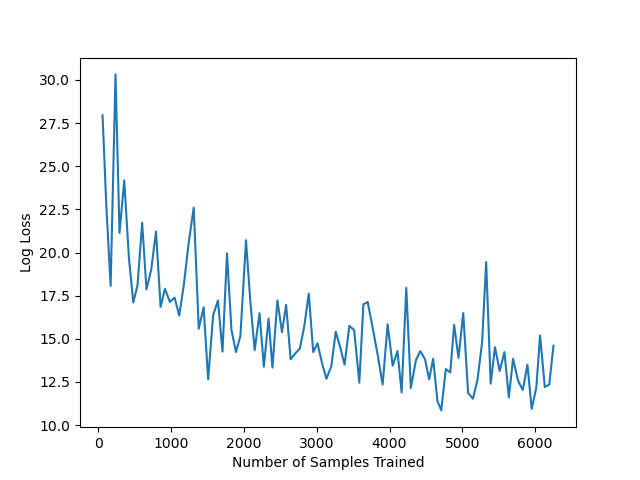


Figure 2. Log-Loss for Linear Regression Classifier