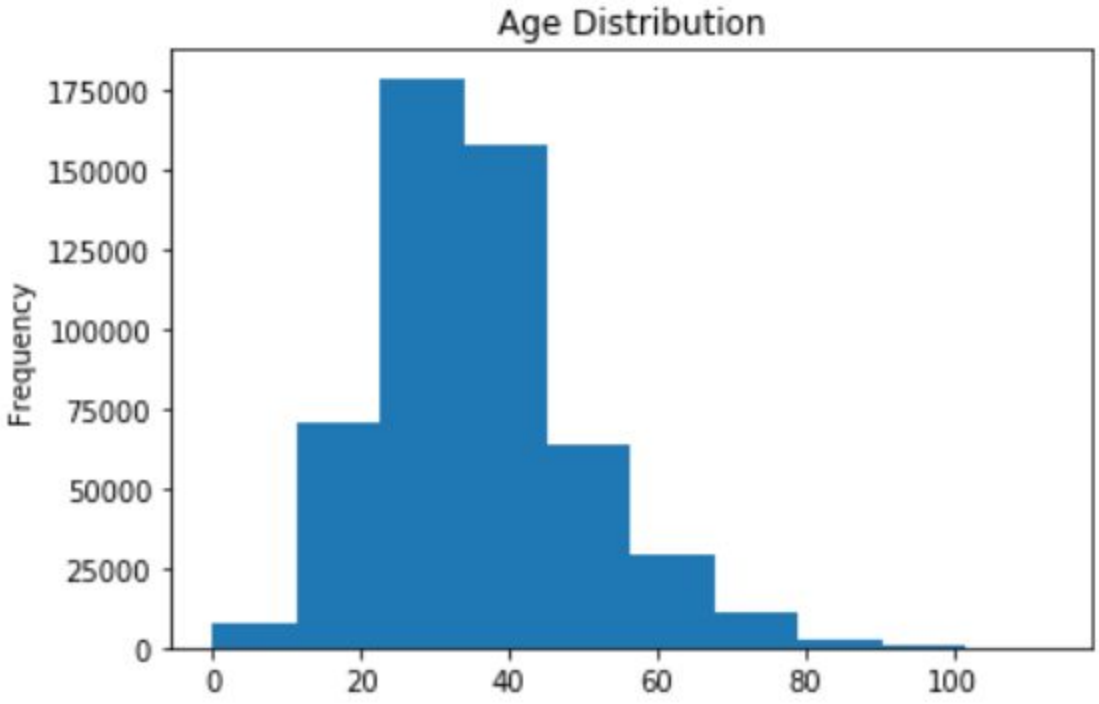
Data Task Report

Arun Ramesh

Task 1

Data was found here: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

Task 2



Task 3

Roughly 1.9 percent of the data consists of 30 year old males

Task 4

I would first partition the dataset into a training and testing set, and use the training dataset for training the algorithm, and the testing set for evaluating the algorithm - I will use a 70-30 split for this.

My next step would be to implement an artificial neural network, and define an activation function as well as initial weights and a learning rate for the training. As this is a form of supervised learning, I can compare the predicted age and actual age, and by utilizing the loss function, backpropagate and tune the weights, until the error is minimized (or the ages are close within a certain defined threshold). By finding the right set of parameters, it will enable me to optimize the algorithm in terms of time complexity and accuracy.

Task 5

The most obvious tradeoff we would have to make is the one between run-time and accuracy of prediction. Depending on the context the algorithm will be used for, we would have to prioritize between the two. For the loss function, I would use a common loss function used in deep learning - the softmax loss function. This would return a probability distribution of possible ages.

Task 6

We can start by building a confusion matrix to see the level of True Positives, True Negatives, False Positives and False Negatives. We must mainly look at the False Positive and False Negative rates, and see what is the probability of a Type I and Type II error. However, it is imperative to first look at what are the benefits of the algorithm getting the age right, as well as the costs - for example, if the person is a minor but is classified as an adult, that would have very negative consequences. Hence such an algorithm would require some manual human oversight.

We must also consider the quality of the training data. For example, if the images are mainly skewed to a certain demographic, it may be hard to classify a person who belongs to a demographic that is not represented well in the training data. As airports can be diverse places, it is not yet clear if the dataset also has a similar distribution, otherwise there could be a problem of overfitting. Hence either the algorithm should be used with significant human oversight, or should not be used in such a risky scenario.

Task 7