MODS207 - Applied Projects

Discrimination on Airbnb

Authors:
Daniel Victor FERREIRA DA SILVA
Gabriel MARTINS DE ANDRADE

Palaiseau, July 2023

Summary

1. Introduction

In today's interconnected world, discrimination on web platforms has emerged as a pressing concern, with Airbnb being a particularly noteworthy example. The harmful impact of discrimination on marginalized communities has drawn significant attention, highlighting the power of the web to shape social interactions and economic opportunities. This demands a closer examination of the role played by web platforms in perpetuating discrimination. To address this issue, our project aims to tackle discrimination within the Airbnb platform.

We will analyze a CSV file containing guest profiles, including names and pictures, employing various machine learning algorithms to estimate the gender and race of each guest based solely on their names, pictures, or a combination of both. Our goal is to understand how accurately we can predict these attributes by incorporating or excluding picture information while retaining the use of names.

The project workflow will encompass several tasks, starting with data preparation and cleaning. We will utilize face_recognition to extract features from the pictures, exploring the application of Python libraries such as Genderize.io for gender estimation based on first names and OpenCV for gender estimation from pictures. Additionally, we will leverage Ethnicolor, NamePrism, DeepFace Race estimation, and other relevant prediction models to estimate race. By thoroughly examining different approaches, we aim to improve the accuracy of our predictions and gain insights into the potential impact of different data sources on discrimination within the Airbnb platform.

2. Background

The multiculturalism present in the centers of capitalism brings with it racism of the majority groups, whether this racism is historical, as in the case of the USA, due to recent immigration waves, as in some European countries, or even a mixture of the two. Minority groups that are targets of racism are usually victims of daily harassment, exaggerated police coercion, and social exclusion, and have a harder time sustaining themselves when compared to majority groups.

Thus, in everyday life, these groups are seen as inferior, even as repulsive, as inherently belonging to the working class, so business owners are not well accepted. Thus, in the financial aspect racism causes minority businesses to perform significantly worse, as described by Jean-François Ouellet in Consumer Racism and Its Effects on Domestic Cross-Ethnic Product Purchase: An Empirical Test in the United States, Canada, and France. This is because many customers do not want to interact with minorities or do not like the image that buying from this store would convey.

Moreover, this performance worsens the more racist the area in which the business is located. This is an important factor since wealthier areas are usually inhabited by dominant groups, since the capital revolves around them, and minorities do not have access to the most lucrative markets and clientele, since their businesses would not be well accepted in these areas. This generates a cycle of spatial segregation of these groups, which hinders their financial ascension and reinforces a racial hierarchy imposed by the dominant group.

Besides being affected on a daily basis, minority groups are also more affected in periods of crisis. This is because the blame for crises is often sought, which ends up falling on minorities, as described by Michelangelo Rossi in Scapegoating and Discrimination in Times of Crisis: Evidence from Airbnb. This article shows how American-Asian businesses were harmed during the COVID-19 pandemic since this ethnic group was associated with the Chinese, who in the minds of racists were to blame for the disease. Thus, because they were seen as the culprits, they suffered a consequent aversion from society, which stopped renting out their homes on AirBnb. The same happened after 9/11 in the US, where groups of Muslim residents were associated with terrorists and were persecuted, receiving less investment and fewer customers.

In terms of gender, something similar occurs, as in many cultures women are seen as less capable than men and belong in a submissive position, making the image of a businesswoman controversial and negative. In the article Gender Differences in Business Performance: Evidence from the Characteristics of Business Owners Survey by Robert W. Fairlie and Alicia M. Robb we find, in fact, that businesses managed by women have lower performance and profits than those managed by men.

Given this context, it is important to study the impacts of these biases on digital capitalism, since on the Internet anonymity makes bias-driven decisions less explicit and generates less user guilt. In addition, there is no conversation and negotiation that exist in personal exchanges, and decisions are often made on the appearance of the available service provider. The responsibility of the companies is also put aside since they argue that they only mediate these negotiations of interests, so they cannot do anything in cases of service providers being prejudiced by color and gender. It is therefore important to study whether these disadvantages are in fact decisive and present on digital platforms. This is because it can often be a better choice for the service provider to hide his or her face with a different profile picture, or without a picture, or group pictures that hide the person's ethnicity and gender.

For such an analysis, however, you need models that analyze large amounts of data and return accurate and consistent results. Since we are dealing with big data, future studies can use classification models to predict the gender and ethnicity of a platform used as a basis for a case study. For this, it is necessary to have models that have been tested and validated, and this will be the goal of this work, to validate the efficiency of the models and libraries for predicting ethnicity and gender that exist today.

3. Data Description

Our data

Fulfilling the objectives of our study, we have employed guest reviews from the popular online platform Airbnb as our primary data source (Figure 1). The CSV file provided contains comprehensive information about each guest's profile, including our variables of interest first name and the web address linking to their profile picture, and other variables such as languages spoken, occupation, and address. Additionally, the dataset includes information about the review received by each guest, with the corresponding date and review content.

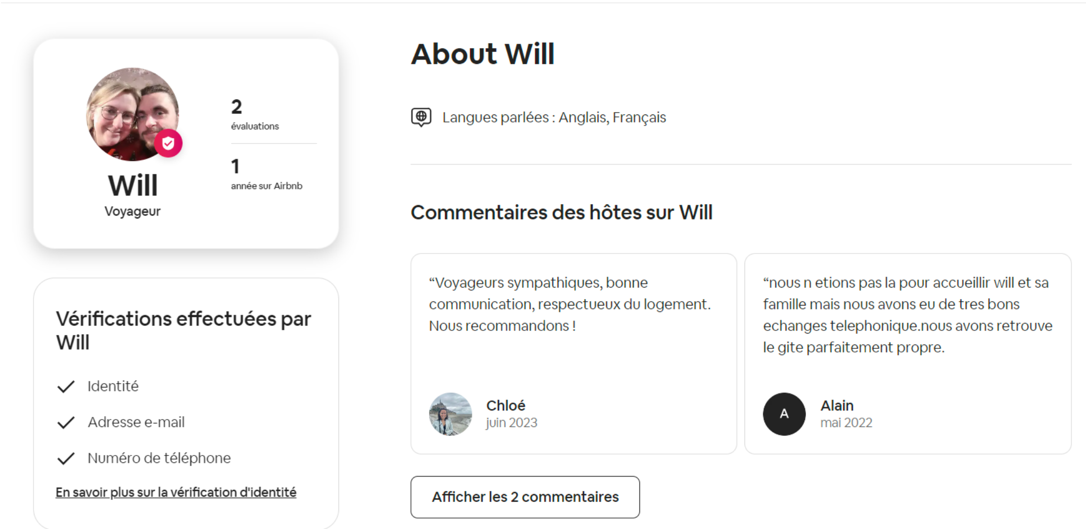


Figure 1: Guest review on Airbnb.

Initially, the dataset comprised a total of 10,000 guest profiles. However, due to limitations regarding data size and cleanliness, extensive preprocessing was necessary to filter out unusable data. Numerous challenges were encountered during this stage, particularly regarding the names and images of the guests. Invalid first names, such as those consisting of a single letter or symbol, as well as accounts shared by multiple individuals or couples, posed difficulties in accurately identifying and classifying the corresponding profiles. Similarly, image-related issues included inaccessible web addresses, profile pictures lacking recognizable faces, pictures of objects, animals, or individuals not directly facing the camera, and images with multiple faces or multiple faces recognized.

Instances containing any of these problems were discarded, as it was not feasible to infer the correct name or distinguish the right face of each sample. However, different preprocessing approaches were employed for each analysis, as necessary. For instance, in gender estimation, samples with one only one detected face among multiple faces of the same gender in the image could still be used, as the predicted gender remained consistent. Nevertheless, accurate analysis of such samples was not possible, potentially impacting the accuracy of the models under study. Consequently, the first step resulted in a significant reduction in the dataset size.

Another limitation encountered was the absence of labels in the dataset. For each analysis proposed, gender and race estimation, distinct sets of labels were required. However, the original data collection did not provide this information, necessitating the manual labeling of each sample. This process, being sensitive and time-consuming, further reduced the size of our workable dataset, as the project's time constraints did not allow for labeling all the samples. Despite additional preprocessing stages, which will be described in subsequent sections, the final sample count for the gender and race prediction task is presented in Table 1 and Figure 2.

class	N° of observations	(%)
male	961	53.75
female	1117	46.25
total	2078	1

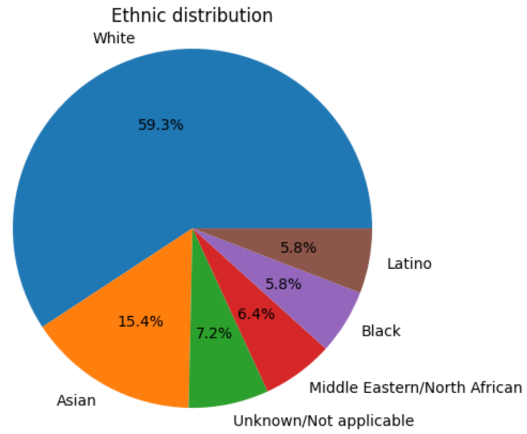


Table 1 and Figure 2: Distribution of gender and ethnic classes and number of samples for each task.

Algorithms

Gender Estimation

The task of gender estimation in our study involves analyzing the accuracy of current gender prediction tools on the Airbnb guest dataset and exploring the potential improvement by combining names and images. For name-based gender prediction, we utilized the Genderize.io Python library, which predicts gender based on a large aggregated dataset of gender information. The image-based prediction model employed was a pre-trained model from OpenCV, a neural network specifically designed for gender prediction based on images.

To facilitate the gender estimation task, additional preprocessing steps were required, specifically face recognition. We utilized the Face_recognition package in Python, which is based on a deep-learning model for face recognition developed by dlib. This package offers high accuracy, with a recognition rate of 99.38% on the Labeled Faces in the Wild benchmark. However, the face recognition model encountered challenges within our dataset, such as low-resolution images, multiple faces being recognized within a single image, and images lacking proper faces, resulting in the loss of numerous observations.

Once the data was prepared, the next step involved predicting gender based on the first names. The Genderize.io API was used for this purpose, offering a simple and convenient way to predict gender based on name frequency, and data collected from various sources across the web, and from multiple countries. The API response is provided in JSON format, including the name, count, assigned gender, and probability, representing the certainty of the assigned gender based on the male-to-female ratio.

It is important to note that the Genderize.io API posed some velocity limitations in our project workflow. The free version allows up to 1000 name predictions per day, and exceeding this limit requires a paid API key with a per-request payment system. However, the API also offers some advantages, such as no sign-up or API key requirement for the free version, and batch usage capability, allowing for the inference of up to 10 names at a time. The API is user-friendly, with library implementations available in various programming languages, including Python.

Similarly, we proceeded with the gender prediction using profile pictures, we employed models from OpenCV, which is an open-source computer vision library that offers optimized and cross-platform support, including a Python interface. The specific version we utilized was the pre-built CPU-only OpenCV package (cv2) for Python, with the gender prediction model architecture and pre-trained weights from the Caffe deep learning framework.

As mentioned earlier, the input images were preprocessed using the Face_recognition library, recognizing the bounding boxes of the faces used as inputs for prediction. The overall pipeline involved retrieving the faces from the recognized bounding boxes, performing some more image preprocessing such as resizing and mean subtraction, gender prediction, and finally cleaning any unrecognized observations.

Furthermore, in the final step of the gender estimation task, we aimed to assess the potential performance improvement by combining both input information (names and images) to create a more robust predictive model. To achieve this, we proposed various models that utilized the classification probabilities from both the name-based and image-based models. Our objective was to enhance the overall accuracy of gender prediction computed with the ground truth labels we obtained through manual classification. Further details about the proposed models and their analysis are provided in the next section, which focuses on algorithmic analysis.

Ethnicity Estimation

For the task of estimating ethnicity, the approach will be similar to that for gender. Image and name-based prediction tools have been used and their accuracy tested. An analysis using both to test the accuracy of the combination has also been done. The data that were used in this part were similar to the data prepared and treated as described in the genre part.

For the prediction using names, the NamePrism API was used, which is a non-commercial nationality/ethnicity classification tool that was trained on 74 million names from 118 different countries. It returns a CSV with the probability that the first name given is from 39 different ethnic groups.

Although it has a high quality of accuracy and was the best tool we found for classifying first names, NamePrism imposes a limit of 60 requests per minute in order not to overload their system. This limitation meant that this was a time-consuming part of the collection, since we had more than 9500 names, this process took more than 3 hours.

For a photo-based ethnicity analysis, we used the DeepFace library, which is a face detection framework that also has the option to return the racial probabilities of a given photo. It has support for predicting 6 different races. Also, it has support for photos with more than one face, making the prediction for all, which will be an important thing to deal with throughout the analysis, since sometimes "ghost" faces are recognized, and sometimes it only recognizes one face in a photo with more than that.

The time for classification was also high, considering that the method first identifies a face, maps it, and then gives its classification, which makes it take an average of 8 hours to fully collect the faces.

For sorting by names the Ethnicolor library was also tested, however, we only had first names, while the library worked with either last name or full name, so its use was not very useful.

Finally, an analysis testing the usefulness of combining the two methods (photo and name) was also performed, testing different confidence intervals for each method and observing the result obtained. Thus, a series of data was labeled randomly, as well for each threshold that we would like to test. This data was used as a ground truth to analyze the accuracy of the models. It is important to point out that this classification was made by the group, so it is not an absolute result, since it depends on the worldview of the classifier. Ideally, a data set with ethnic self-declarations would be ideal, since a third party judging another's ethnicity can result in errors by carrying biases.

So, 2 macro analyses were done, an initial one that took into account a multitude of ethnicities. This was done so that when doing the manual labels, 5 ethnicities were chosen to fit each person, these being: Asian, Black, Latino/Hispanic, Middle Eastern/North African person, and White. In this way, we adapted the outputs of the photo and image classifiers to fit these groups. This analysis, however, carries more errors in ethnic classification of groups, since people doing the labeling process often cannot easily differentiate between different ethnicities that have similar traits. One way out was to give more than one possible ethnicity for each photo, i.e. if there was a doubt about the possible origin of the person, all options were given as labels, and if the classifiers predicted any of the predicted labels, we considered it a correct prediction.

But still, this analysis could be biased, besides the fact that we had few samples from the non-white groups, resulting in few possible tests. So another analysis was done considering only white groups and clustering non-white groups as a new classification. Further details about the proposed models and their analysis are provided in the next section, which focuses on algorithmic analysis.

4. Algorithm Analysis

Gender Estimation

As previously described, our study focuses on analyzing the accuracy of current gender prediction tools using the Airbnb guest dataset, as well as exploring potential improvements by combining different algorithms. In the gender estimation task, we first utilized the Genderize.io API for predicting gender based on first names. The classification results are summarized in Figure 3 and Table 2. We discarded observations with invalid or unknown names, single-letter or symbol names, and names not found in the Genderize.io database, retaining only the successfully predicted samples in the classification results.

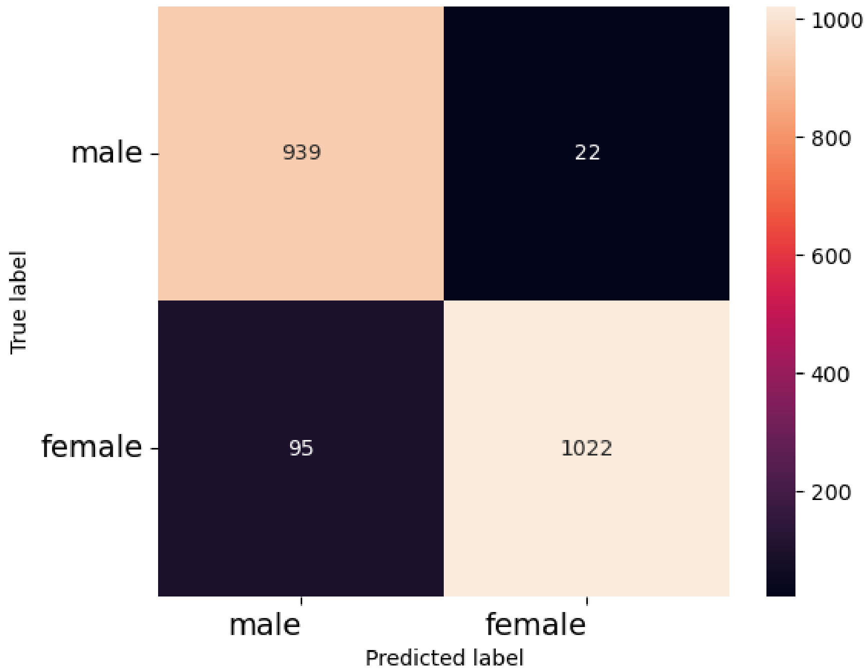


Figure 3: Confusion matrix of Genderize predictions

	precision	recall	f1-score	support
male	0.908124	0.977107	0.941353	961.000000
female	0.978927	0.914951	0.945858	1117.000000
accuracy				0.943696
macro avg	0.943525	0.946029	0.943606	2078.000000
weighted avg	0.946183	0.943696	0.943775	2078.000000

Table 2: Genderize classification report

Considering that the Genderize.io API is based on frequency data collected from various sources on the internet, it was expected that estimating someone's gender solely based on their first name would not produce perfect results. This is due to the existence of many gender-neutral names across multiple countries, and the analysis did not account for any potential local biases. As anticipated, the results confirmed these expectations, revealing some errors in the predictions. However, despite these challenges, the model achieved an overall accuracy of 94.36% across 2078 observations. This outcome is surprisingly good considering the limited input information. Specifically, the model demonstrated a precision of 98% for the female class and a recall of 98% for the male class.

Moving on to the image-based gender estimation using Opencv gender prediction models and Face_Recognition face recognition tools, the results are depicted in Figure 4 and Table 3. The preprocessing stage for the image data presented its own challenges, with the need for additional down- or upscaling of some faces to ensure proper recognition. The poor resolution of many images, combined with other dataset qualities, contributed to somewhat unsuccessful results. It is worth noting that even the difficulty of the labeling task could have led to misidentified observations. The precision for males was 0.75 with a recall of 0.78, while for females, the precision was 0.80 with a recall of 0.77.

Given more time and a better-treated and pre-labeled dataset, further analysis could involve exploring different models and pre-trained weights from the Opencv library.

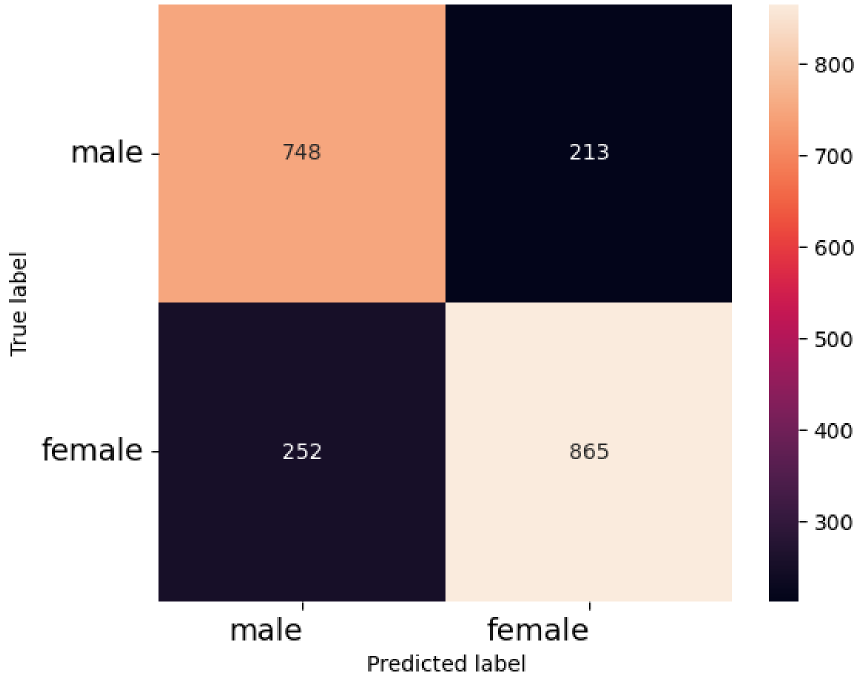


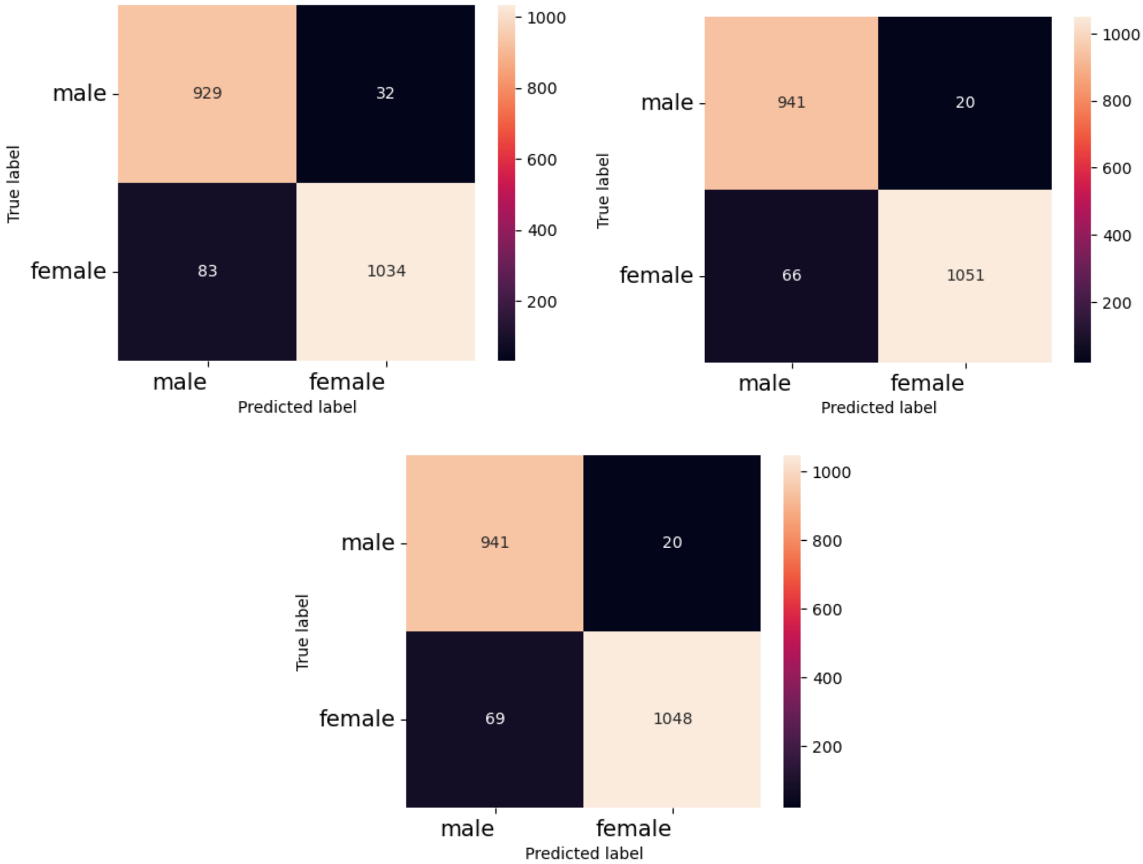
Figure 4: Confusion matrix of Opencv predictions

	precision	recall	f1-score	support
male	0.748000	0.778356	0.762876	961.000000
female	0.802412	0.774396	0.788155	1117.000000
accuracy				0.776227
macro avg	0.775206	0.776376	0.775515	2078.000000
weighted avg	0.777248	0.776227	0.776464	2078.000000

Table 3: Opencv classification report

Moreover, the next step focused on proposing a model with improved performance by combining both image-based and name-based algorithms. Three models were proposed based on the predicted probabilities from both algorithms, aiming to achieve better predictions than the individual algorithms.

The first model selected the output with the highest assigned probability between the two models. The second model considered the higher accuracy of the name-based algorithm as a foundation, using the Genderize output if the probability exceeded a threshold, or selecting the prediction with the highest probability if it did not. The third model used a weighted average of the probabilities from both models to determine the predicted gender. The results of these models are illustrated in Figures 5 to 7 and Table 4.



Figures 5 to 7: Confusion matrices of Models 1 (top left) to 3 (bottom).

	class	Model 1	Model 2	Model 3
precision	male	0.9180	0.9345	0.9317
	female	0.9700	0.9813	0.9813
recall	male	0.9667	0.9792	0.9792
	female	0.9257	0.9409	0.9382
f1-score	male	0.9417	0.9563	0.9549
	female	0.9473	0.9606	0.9593
accuracy		0.9446	0.9586	0.9571

Table 4: Classification report of the proposed models.

The optimal parameters were determined as a probability threshold of 92% for Model 2, weights of 0.43 for Opencv probabilities, and 0.57 for Genderize probabilities for Model 3. Analyzing the model's results, although all performed better than the algorithms isolated, Models 2 and 3 performed better, reaching similar accuracy levels of around 95%. These results confirmed our initial hypothesis that combining name and image information can increase the capacity for gender estimation. Furthermore, they highlighted the greater impact of the name-based model in guaranteeing great accuracy, despite the limitations found in the image-based model.

Race Estimation

For ethnicities, we will start with the DeepFace results. First, before looking at the predicted results for the races, we will look at the quality of the tool's face recognition, since in order to have race prediction, it is first necessary that the faces are properly recognized. When the classification was being done, we assigned different flags for each type of classification, so there was one flag for 0 identified faces, one for 1 face, another for 2 faces, and another for 3 or more faces. So, let's see what the final distribution of these flags looks like:

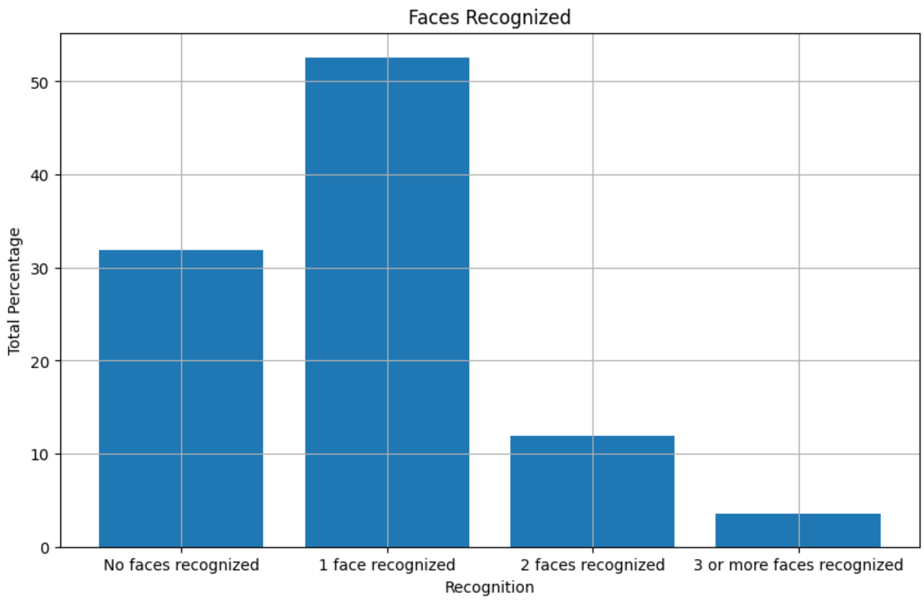


Figure 8: Bar graph of the distribution of recognized faces

We have a lot of photos that will not be classified because they don't have recognizable faces. Furthermore, we chose not to work with the cases of 3 or more recognized faces, since it would be difficult to label different faces and then compare with the results of the tool. The cases of 2 faces, however, could be used, so that classification will be considered correct if both predictions of the faces are met in the labels. The use of 2 faces, however, depends on the quality we find in the test below.

Therefore, we did a test to analyze how accurate this face recognition is, we focused on cases of 0,2 and 3 or more recognized faces (but we also took 100 photos of 1 face recognition) and gave labels saying in which case the photo fits, resulting in 400 labels. A first look at the classified data made us realize that there were many cases of photos with 1 face that were misclassified, cases of upturned or profile faces, and mainly the use of adornments like glasses or hats caused several misclassifications. This was verified because after assigning the label, the predicted number was shown to the person who gave the label, who could compare the results. Thus, we got:

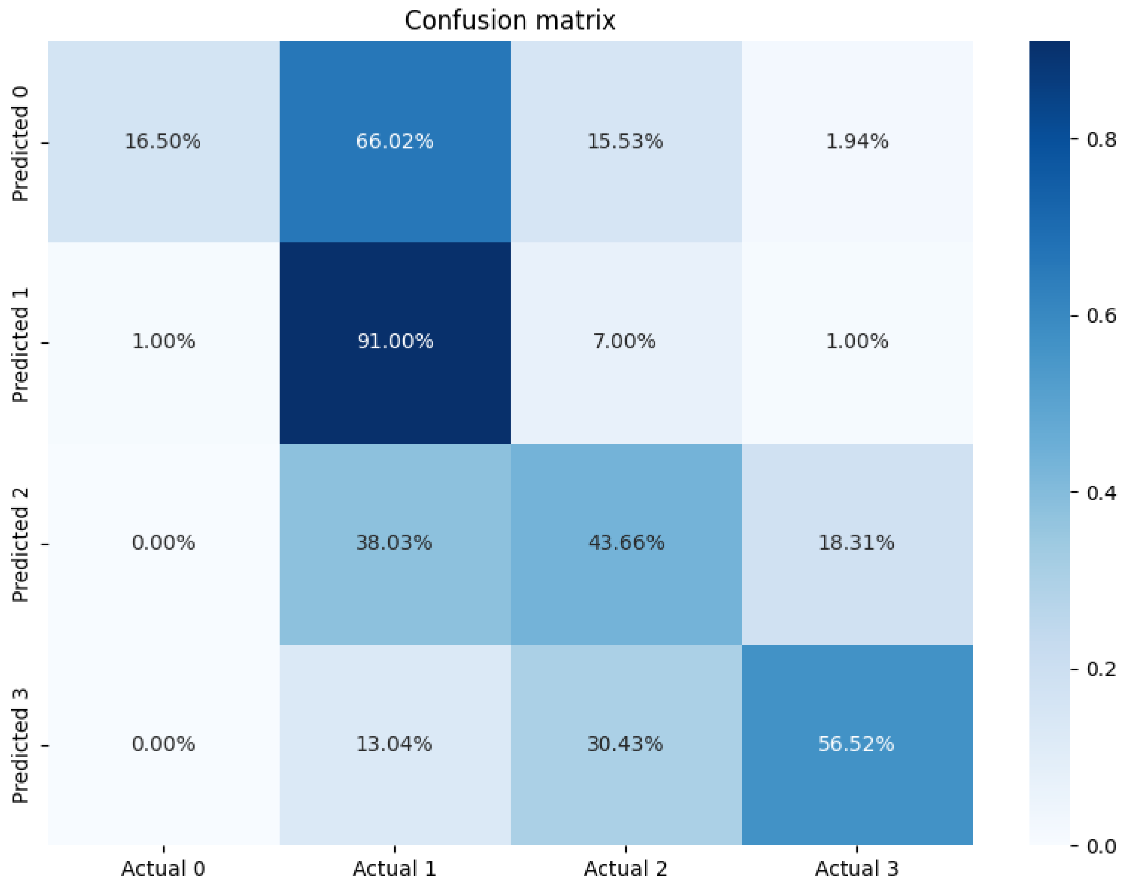


Figure 9: Confusion matrix for the number of faces

We observed that several photos where there was 1 face were classified as 0, which is a significant loss and indicates the difficulty of the classifier to properly identify a face since almost 2/3 of the 0 face predictions had a face there. The prediction of 1 face proved to be very consistent, with the largest case of error being in photos of 2 people the classifier identified only 1. Since by giving labels to the photos, they contemplate all the people in the image, this will not be a problem. For 2 faces, however, we see very low reliability, which will make us not use these cases to test the accuracy of the ethnicity prediction.

Now, for race prediction, DeepFace returns the probability that the person in the photo fits into 6 different groups: Asian, Indian, Black, White, Latino/Hispanic, and Middle Eastern person. Since in our classification, we don't consider "Indian", we will add the probability of this output to the probability of "Asians", so we will have a classification that is similar to the one made in the labels. So we compared the values and found an accuracy of 67.18%, which is a low value considering that for a large amount of data, we initially lost about 20% of the data due to facial recognition error, and of the valid data we classified 1/3 wrong

To better understand our accuracy, we split DeepFace's prediction probability intervals and calculate the accuracy of each one Furthermore, we put a graph of the presence of each prediction probability on the data:

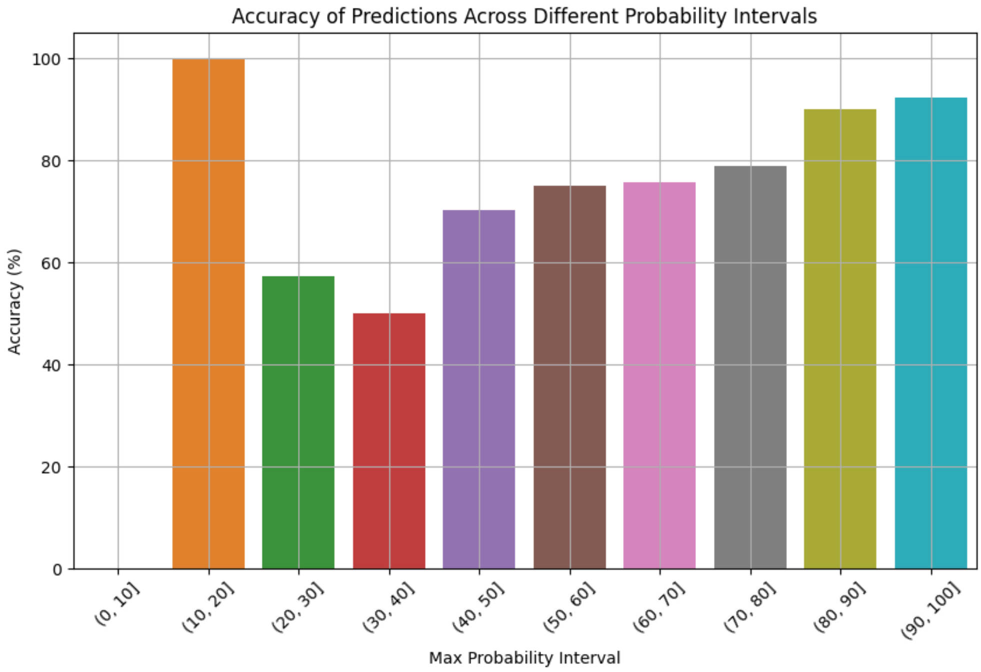


Figure 10: Bar graph with the accuracy per probability interval for faces

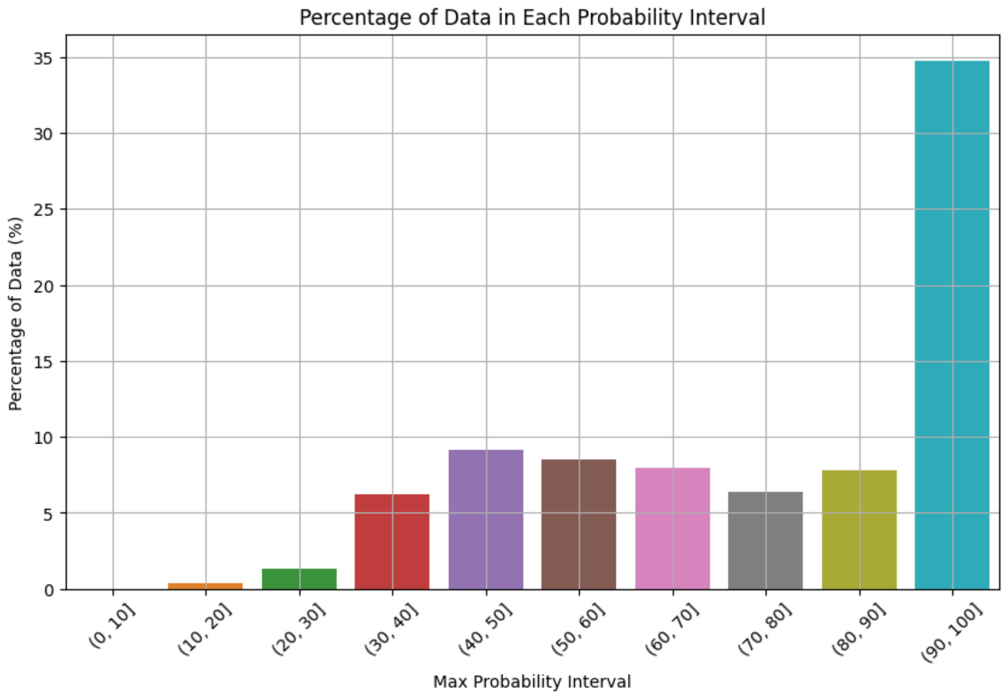


Figure 11: Bar graph with distribution of probabilities for faces

The interval cases up to 30 are anomalies because they show very few predictions (under 1%), and by chance, they were mostly correct. From 30 to 100 we can see a tendency of increasing accuracy, as expected.

Now, since several times minorities are represented by every non-white group in a society, we will do an analysis of the case of white/non-white classification. So we will use the returned probability of being white and for non-white, we will use its complementary value. The labels will also change, merging all non-white labels into one category. In this new model, we had an accuracy of 80%, and distributed as seen below:

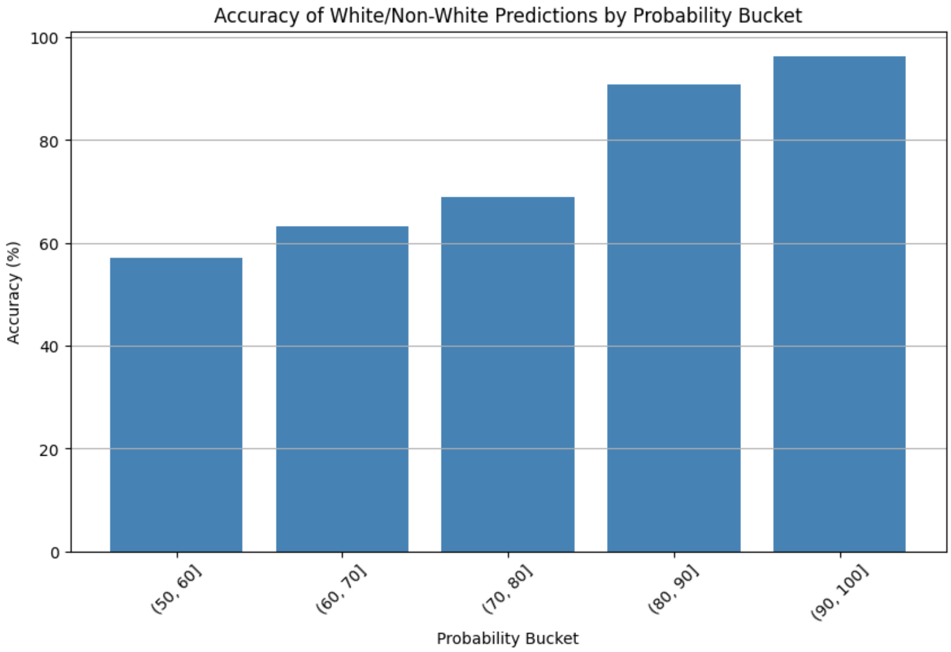


Figure 12: Bar graph with accuracy per probability interval in White/Non-White model for faces

There are no probabilities below 50% since we are in a binary model of classification. Moreover, again we see the tendency of a higher accuracy the higher the certainty of prediction, which is coherent and shows that the model presents a validity in these returned values. Something interesting observed is that if we calculate the accuracy by ethnicity, we will have that white people are usually classified more accurately than people of other ethnicities, while the worst accuracy is for black people. As this model is a DeepLearning tool, we have that it was trained with images and photos of people, however, those who have more access to the Internet and end up putting more photos on it are those who have more money, so, people from the dominant group: white. Below we have the accuracies:

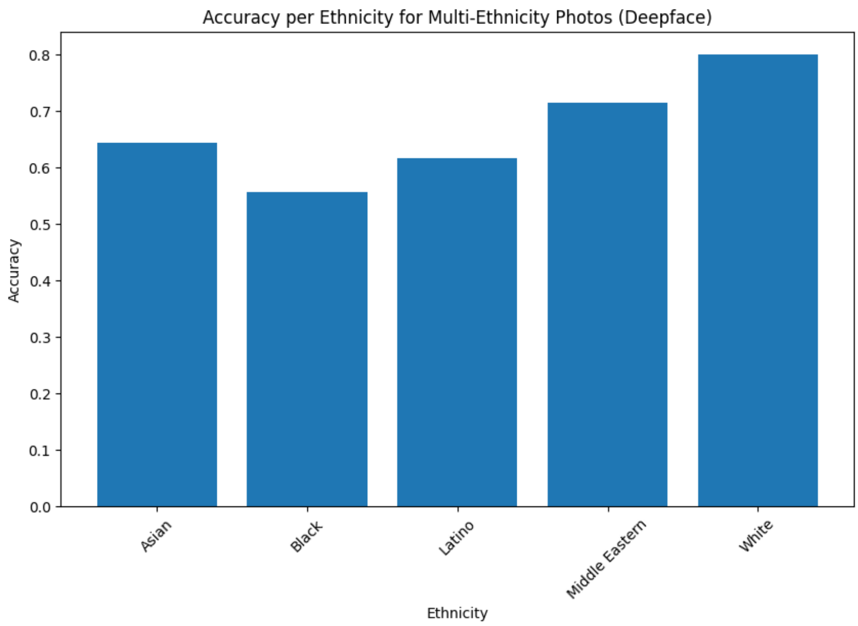


Figure 13: Bar graph with accuracy per ethnicity

Now, using names, we also sort the entire dataset using the NamePrism API. It does, however, return several probabilities for various ethnicities, these being: European-SouthSlavs, Muslim-Pakistanis-Bangladesh, European-Italian-Italy, European-Baltics, African-SouthAfrican, European-Italian-Romania, Muslim-Nubian, European-French, EastAsian-Indochina-Thailand, EastAsian-Indochina-Vietnam, Jewish, Muslim-Turkic-CentralAsian, EastAsian-Indochina-Cambodia, Nordic-Scandinavian-Denmark, EastAsian-Indochina-Myanmar, Nordic-Finland Muslim-Persian, Nordic-Scandinavian-Sweden, Muslim-Maghreb, Greek, Muslim-Pakistanis-Pakistan, Hispanic-Portuguese, European-Russian, Muslim-ArabianPeninsula, African-WestAfrican, EastAsian-Japan, European-German, EastAsian-Chinese, SouthAsian, Hispanic-Spanish, Nordic-Scandinavian-Norway, Muslim-Turkic-Turkey, Hispanic-Philippines, CelticEnglish, EastAsian-Malay-Malaysia, EastAsian-South Korea, African-EastAfrican, European-EastEuropean, EastAsian-Malay-Indonesia.

Thus, we agglutinated these outputs into classifications that were consistent with our classification so that we will make the following groups: Asian: All "EastAsian" and "SouthAsian" categories. Black: Both "African" categories. Latino: All "Hispanic" categories. Middle Eastern/North African: All "Muslim" categories, "Jewish". White: All "European", "Nordic", "Celtic-English", and "Greek" categories. The comparison was made with the same labels used on the model in the photos. With this, we obtained an accuracy of 66.60%. This is also low, considering that we can misclassify 1/3, but in this case, at least we won't waste data, since names can always be classified

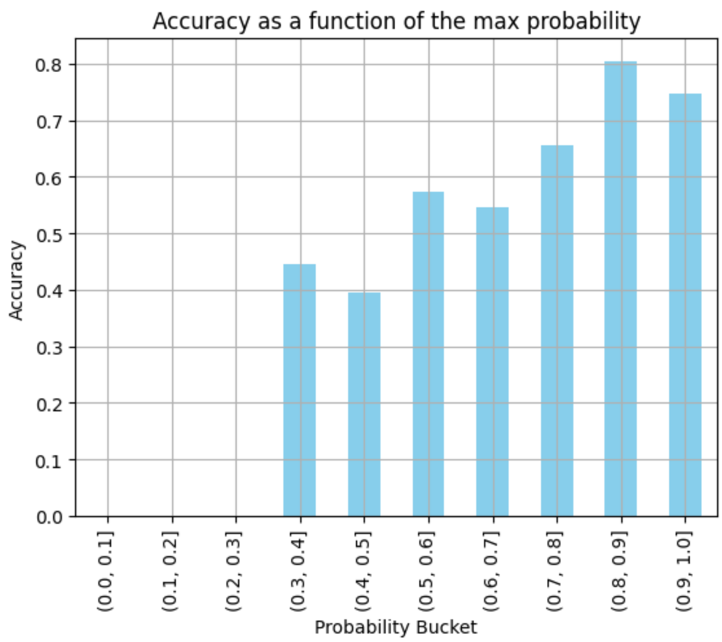


Figure 14: Bar graph with accuracy per probability interval for names

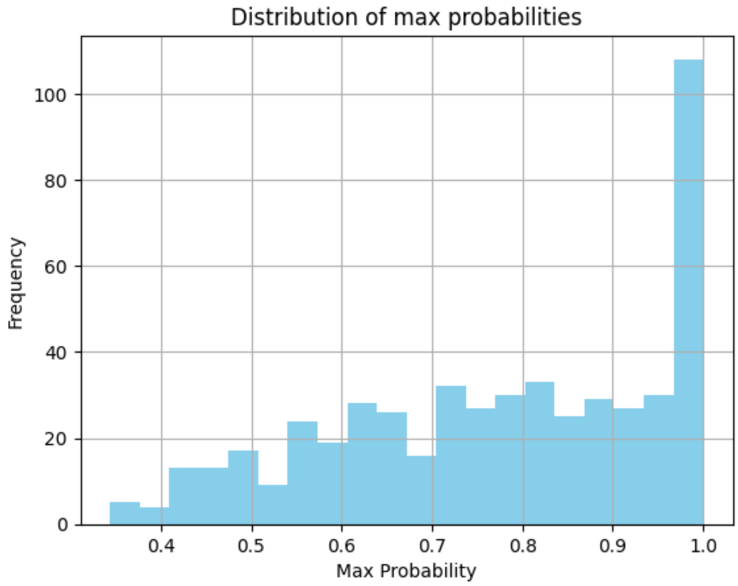


Figure 15: Bar graph with distribution of probabilities for names

Also noticeable in this case is a tendency for the accuracy to increase as the returned probability increases. Now for the binary white/non-white model, we got an accuracy of 74,56%, which is a considerable improvement:

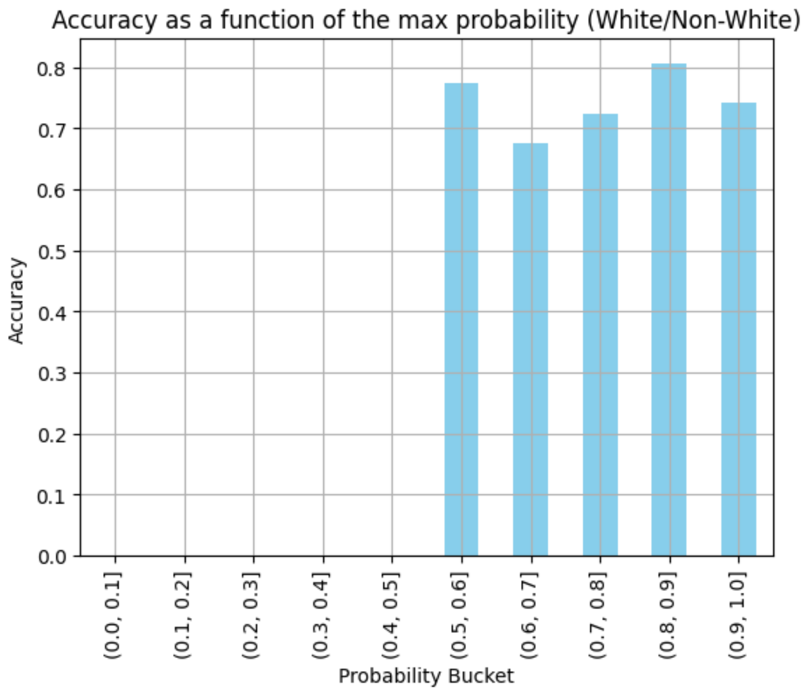


Figure 16: Bar graph with distribution of probabilities for faces in the white/non-white model

It is interesting to note that we no longer have an increasing trend of accuracy. So now we will propose a model that combines both names and pictures. So, since predictions by photo are the limiting factor since we often lose photos by not identifying faces, we will use it as a basis. So we will test different probability values on the names using the probability value of the photos as a base. Thus, we chose data where both models predicted the same ethnicity, and that DeepFace had a probability of the prediction above 50%. Thus, we obtained the following probability distribution, with also the graph showing the count of data for each case:

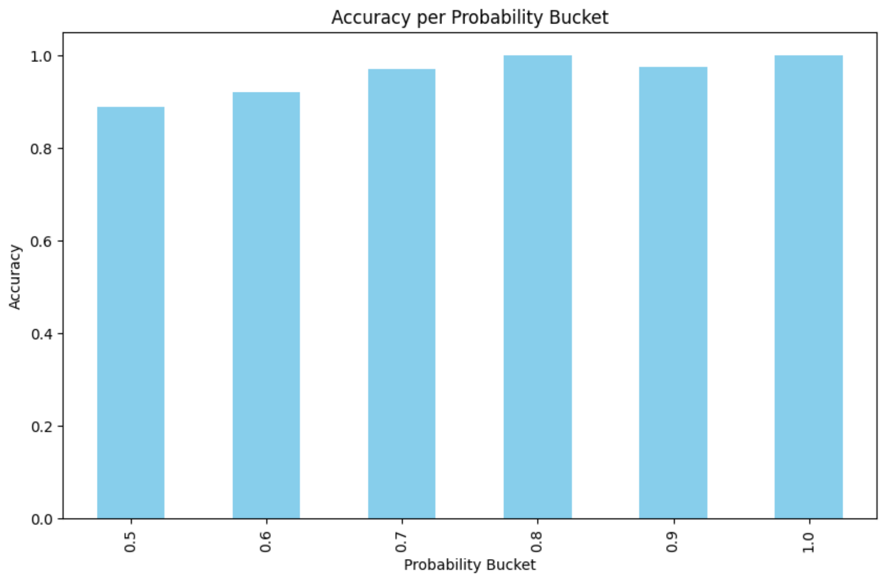


Figure 15: Bar graph with distribution of probabilities for the hybrid model

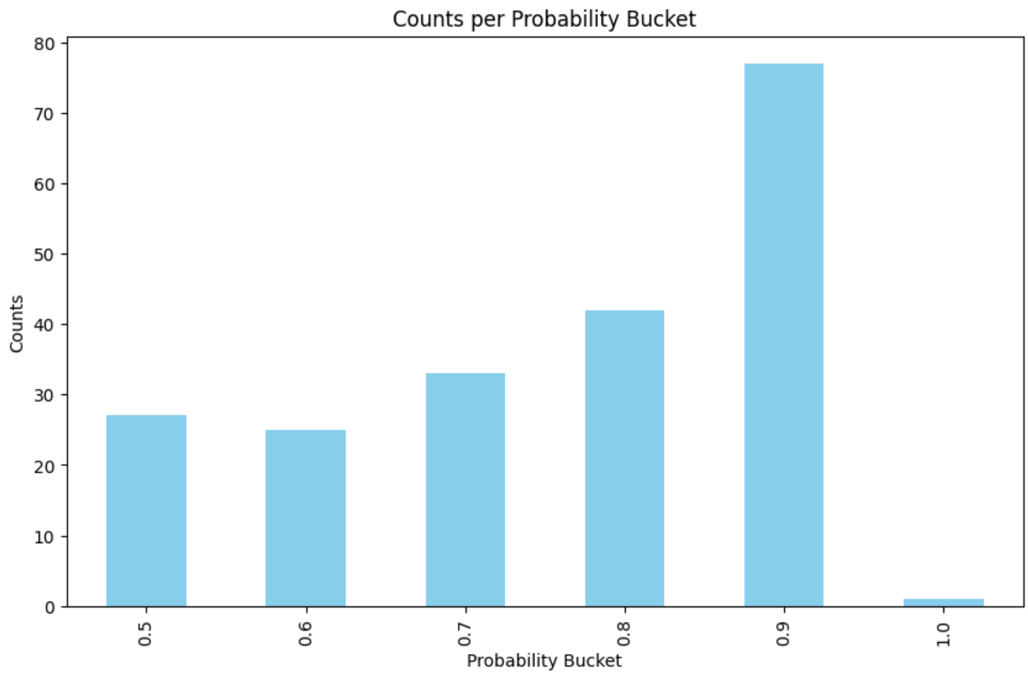


Figure 16: Bar graph with data count for each probability

With this model, we got an average accuracy of 96.10%, which is very good when compared to the previous results, and that's just by comparing the results and taking the data with a probability above 50% from DeepFace (which excludes a small amount of the data, about 17%, which were often facial recognition errors. From what we got in the graph, we can ignore the 100% bar, since we only had 1 data match. For a better visualization, let's look at the restricted probability graph. Then, the bar at 50 will indicate the accuracy if we reset the minimum probability to this value, and so on:

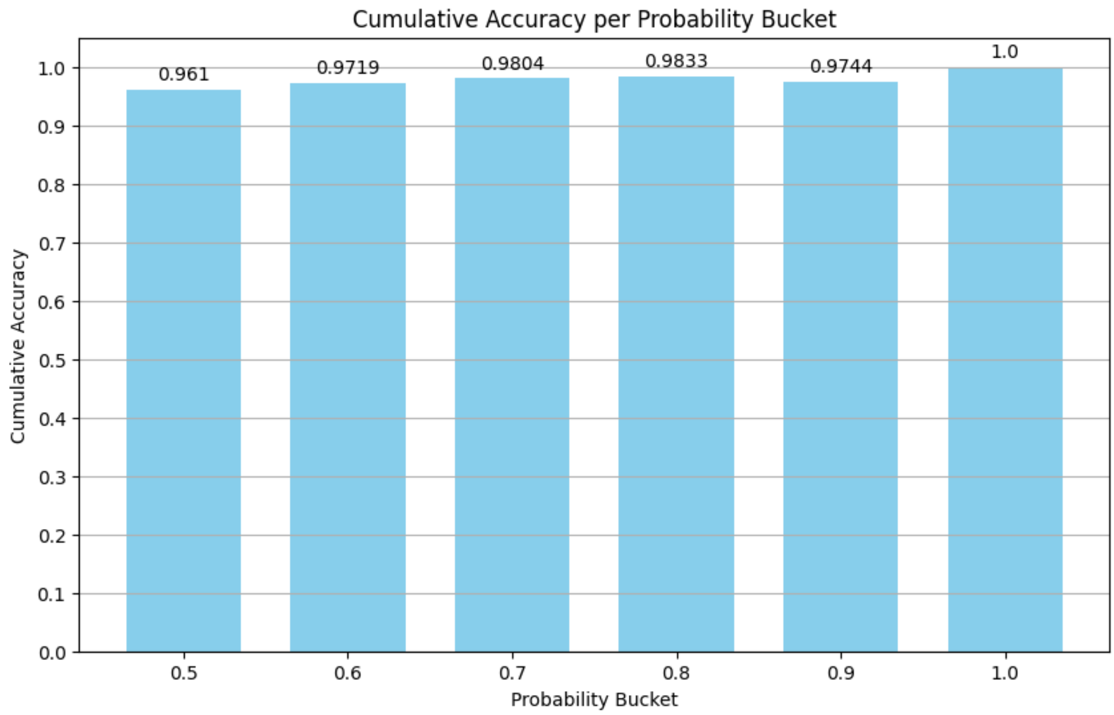


Figure 17: Bar graph with cumulative probability for the hybrid model

So we can get up to 98% accuracy if we take the probability values above 80% from NamePrism. So we can say that this model would be accurate if we want to predict the ethnicity of a large amount of data based on photos and first names.

However, a model that was not done here but would be ideal in this case would be a Stacking model, which does training based on the output of the two models and generates a third prediction model that gives weights to each of the others. This model wouldn't exclude any data besides the ones already collected by DeepFace and would probably present a good accuracy.

For the white/non-white model, this analysis will not be done, since we could not choose a good threshold for the predictions, since this binary model implies the lowest probabilities of 50%, and taking values in the 70/80% range of the probabilities would be a significant loss in the data.

Therefore, we had that the models individually do not perform well, but we confirmed that by combining the two, the results obtained were satisfactory, confirming our hypothesis. It is important to note, however, that the cost of this was a large amount of unusable data.

5. Conclusion

Racism and misogyny are present themes that affect people daily, and the economic aspect would be no different. In a platformized economy, the fight should be constant against racist biases in the choices of service providers, since it is something that can go unnoticed and platforms will not do much without pressure. Therefore, validation and testing of the tools that can be used to find minorities is essential, so that one knows the risks and biases is subject while using it.

In this context, analyzing the tools for gender prediction, we observed that their accuracy was 94% for names-based and 77% for images-based models. For the profile pictures, relying on image prediction models, which depend on the faces recognized and extracted facial features, the model suffered with the poor quality of the input data, among other factors that contributed with the OpenCV model performing worse than the name-based one. For names, despite that there are usually cultural differences about names' genders, they are easier tools to determine someone's gender algorithmically, except in the cases of unusual names that were not present in the model's database or unisex names. And with the 3 proposed models that combined images and names, achieving an accuracy of up to 95.8%, we confirmed the potential to increase the performance of the gender estimation task by utilizing both input information.

As for the prediction of ethnicities, for the attempt to predict several ethnicities the photo model gave us an accuracy of 67% and the name model a 66%. These values are not satisfactory and indicate models that often predict incorrectly. This is because the prediction of ethnicities from photos depends on the machine's attempt to associate faces with ethnicities, which ends up resulting in attributions of facial features to certain ethnicities that may be present in others, indicating that the model has not been trained enough yet to be considered sufficient. As for names, although they are a cultural tradition, we have waves of immigration that cause ethnic groups to change country and language and often adopt local names, as well as cases of names that appear in one region imported from another by fashion.

Thus, the naming model has a low and unsatisfactory accuracy which can hardly be improved upon much. We proposed a model that combined both photos and names and with this, we managed to increase the accuracy to 96% without losing much data, which is an extremely satisfactory result. For the attempt to predict only whether the person is white or non-white, we had an accuracy of 80% for the faces model and 74% for the names model. These numbers are better than before, but considering that we are in a binary model, they could be better.

Another interesting finding was that the face prediction model has better accuracy for white people, and worse for black people, which is a reflection of the data that were used to train it, evidencing a higher amount of pictures of white people. This is, however, expected, since people from the dominant class have more capital, and, consequently, more access to photo devices and the internet, but it still highlights the structural inequality present in society.

Thus, we can conclude that gender forecasting models are currently good and can fulfill their role reliably. Conversely, the ethnicity prediction models are not as good on their own, but by trying to combine them, we get results that are satisfactory and reliable for large data analysis.

6. References

Rossi, Michelangelo (2022). Scapegoating and Discrimination in Times of Crisis: Evidence from Airbnb. Available at https://www.nber.org/papers/w30344

Ouellet, Jean-François (2007). Consumer Racism and Its Effects on Domestic Cross-Ethnic Product Purchase: An Empirical Test in the United States, Canada, and France. Available at: https://journals.sagepub.com/doi/abs/10.1509/jmkg.71.1.113

W. Fairlie, Robert (2008). Gender Differences in Business Performance: Evidence from the Characteristics of Business Owners Survey. Available at: https://repec.iza.org/dp3718.pdf