Universitat Autònoma de Barcelona

Facultat de Ciències

Academy Awards: rewarding WHITENESS and MALENESS since 1928

Gerard Vinyes Sanchez
Grau en Matemàtica Computacional i Analítica de Dades

September 2023

To my parents and my sister, for all their support during all these years and for always believing in me

Abstract

The goal of this Thesis is to do an analysis on the winners and nominees of the awards given annually by the Academy of Motion Picture Arts and Sciences, the Oscars.

This analysis has been performed on the movies from the 74th edition (2002) until the 94th (2022) and a prediction of the winners of this year has been done based on the results. The results, however, were announced on March the 12th so by the time the final version of the project is delivered, it will be easy to see whether the predictions were accurate or not, and an explanation regarding the differences will be given. The main reasons of using these years instead of all the historical data are two: firstly, that times have changed a lot, so using all the information would result in biased information that would surely introduce error in the prediction, and second, because in 2002 the Best Animated Feature was added as a category, being latest added to the current list, so this way all the categories' predictions will be done on the same amount of data (apart from the Best Picture category that every year receives more nominations than the rest).

The data is mainly obtained from csv files found on the internet, plus some extra datasets that have been created manually, and after doing the proper handling and sanitization, the data has been saved onto a non-relational database previously created in MongoDB, having every nomination as a single entry, and easily accessible using queries. The choice of using a non-relational database instead of a relational one is purely based on the idea of being able to have different attributes depending on the category of the nomination instead of having to have the same for all of them and having to fill with NaN values the ones that are not applicable.

The idea of making a prediction of the winners, is not only in order to demonstrate that Machine Learning and Artificial Intelligence can be used in a lot of different ways, but also to use it as a support to the critics that the Academy has been receiving on the recent years about how the configuration of the 23rd chromosome and the type of melanin of the directors and the cast of the films directly influence to whom the statuettes are awarded.

Contents

List of Figures

List of Tables

1 Collection of the Data

The main data for this project has been taken from a dataset in Kaggle [**1**] containing every nomination since the first ceremony in 1928 until 2023. Each entry on the dataset has the year of the film, year of the ceremony, number of the ceremony, category of the nomination, name of the nominee (name of the producers in the case of Best Picture, name of the producers and directors in the case of Animated Feature Film and name of the country in the case of International Feature Film), name of the film and whether the nomination ended up in a win or not.

A similar dataset also found in Kaggle [**2**] was used to get the nominations for the Golden Globes, from where I saved the nominations with categories equal to Best Picture, Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role, Actress in a Supporting Role, Directing, Writing (Original Screenplay), Writing (Adapted Screenplay), International Feature Film, Animated Feature Film, Music (Original Score) and Music (Original Song).

Finally, 5 datasets were created from zero as they seemed very important in the process of creating a predicting model but weren't on the internet:

1. BAFTA Award: dataset with the winners and nominees in the categories of Visual Effect and Film Editing from 2002 until 2023 (length of the dataset: 498 rows) [**3**]
2. Screen Actors Guild Award (SAG): awards destined to recognize the performances of actors and actresses in both film and television. The created dataset has the nominees for Actor in a Leading Role, Actress in a Leading Role, Actor in a Supporting Role and Actress in a Supporting Role and whether they won or not, from 2002 until 2023. (length of the dataset: 440 rows) [**4**]
3. Visual Effects Society Award (VES): dataset with the winners and nominees in the category of Best Visual Effects in an Effects Driven Motion Picture from 2003 until 2023. (length of the dataset: 96 rows) [**5**]
4. Directors for International Feautre Film: dataset with the directors of the films nominated for International Feature Film, so that they can be included in the analysis by race and gender. The information was taken from a google search of every nominee. (length of the dataset: 111 rows)
5. Race and Gender: dataset created manually with the race and gender of the nominees that belong to the categories where I believe these two factors have more influence on the decision of the winner. To simplify the analysis, I have reduced the possible values to the following:

1. Gender:
    - Male: in the analysis it may also be referenced as M
    - Female: in the analysis it may also be referenced as F
    - Female - Male: in the analysis it may also be referenced as F - M or Female and Male

2. Race:
    - White
    - Hispanic: it includes people from Central and South America
    - Black
    - Asian
    - Middle Eastern
    - Indian
    - Hispanic - White: in the analysis it may also be referenced as H - W
    - Black - White: in the analysis it may also be referenced as B - W
    - Asian - White: in the analysis it may also be referenced as A - W
    - Middle Eastern - White: in the analysis it may also be referenced as ME - W
    - Indian - White: in the analysis it may also be referenced as I - W
    - Middle Eastern - Black: in the analysis it may also be referenced as ME - B
    - Asian - Indian - White: in the analysis it may also be referenced as A - I - W

    To choose the race, a google search of the nominee was done and both the place of birth of the nominee as well as the ethnicity of their parents were taken into account. In the cases of mixed races between White and another race, the latest has been chosen as this project intends to show the bias on race and the tendency of choosing caucasians as winners.

    Races and genders with hyphens are used if there are multiple people in the same nomination and there is at least one representative of each value in the nomination, no matter how many of each there are. Finally, in the case of transgender people, the gender at the moment of the nomination has been chosen in order to avoid inconsistencies for the acting categories. (length of the dataset: 941 rows)

2 Data Cleaning

2.1 Academy Award

First of all, the names of the categories have changed across the years so they had to be mapped from the different forms that have existed to the current ones.

In second place, there are some categories where multiple people were included in one nomination specifying name and role for everyone. This was changed to just an alphabetically sorted list of all the people involved regardless of the role they played in the category to simplify the process of associating race and gender to the nominees when needed. The categories affected were the ones involving Screenplay, Music, Costume Design, Makeup and Hairstyling and Directing (regular, Animated or International films)

In third place, there were also some special cases that had different names across the dataset and that were normalized in order to avoid inconsistency issues.

Finally, for the category of International Feature Film the continent where the film was made and the director of the film were added so that it could be included in the analysis.

2.2 Golden Globes

The first change needed was the creation and use of a dictionary to map the categories' names to the ones in the Academy Awards. In most cases the issue was just having a different name, but there were three exceptions: for the acting categories, "Best Performance in a Motion Picture - Drama" and "Best Performance in a Motion Picture - Musical or Comedy" for both Male and Female were mapped to "Actor/Actress in a Leading Role" as the Academy Awards don't make distinctions in the movie genre. The same happened for "Best Motion Picture - Drama" and "Best Motion Picture - Musical or Comedy" which is mapped to "Best Picture". Finally, the Golden Globes don't make distinction between Original and Adapted Screenplay, so the nominations for "Best Screenplay - Motion Picture" were mapped to both "Writing (Original Screenplay)" and "Writing (Adapted Screenplay)".

The second change done was a mapping dictionary regarding the name of the films, because in some cases the names changed slightly. It was always taken as valid the name in the Academy Awards dataset.

Finally, the dataset only had entries until 2020, so the nominations of the three years that were missing were added manually.

3 Exploration of the Data

For both the exploration of the data and the prediction model, only data ranging from 2002 until 2023 will be used for two main reasons: it is the period where all the categories featured in the 2023 ceremony were awarded and also because using all the data would introduce error to the prediction.

In the following doughnut charts we can see the proportion by race and gender of the nominees and winners in 13 of the 23 categories, although for the four acting categories corresponding to the figures 1, 2, 3 and 4, the gender isn't represented as the name of the category makes it obvious. The categories featured in the exploration are:

1. Actor in a Leading Role
2. Actress in a Leading Role
3. Actor in a Supporting Role
4. Actress in a Supporting Role
5. Costume Design
6. Makeup and Hairstyling
7. Writing (Original Screenplay)
8. Writing (Adapted Screenplay)
9. Music (Original Score)
10. Music (Original Song)
11. Directing
12. Animated Feature Film
13. International Feature Film

Nominees are represented in blue and winners are represented in green (which can be seen in the inner layer), race is represented in the outer layer and when necessary, gender is found in the central layer.

In all the categories, there is a clear majority of White winners and nominees, in some of them very vast, and it would be even more extreme if we took into account the entirety of the winners and nominees since 1928.

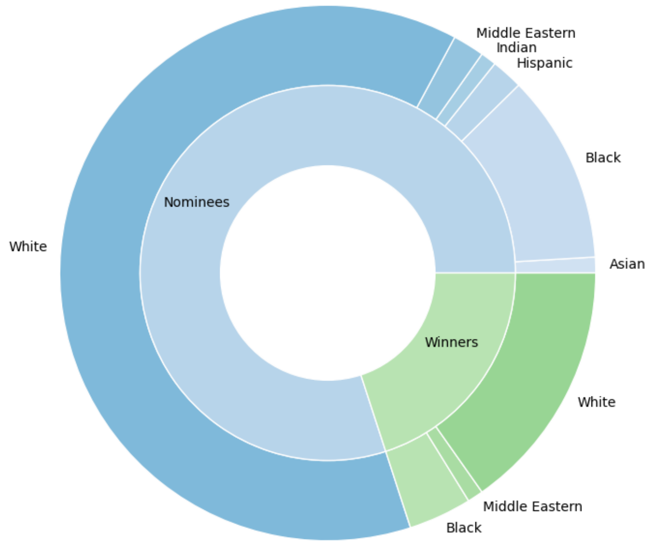

Figure 1: Distribution of Actor in a Leading Role

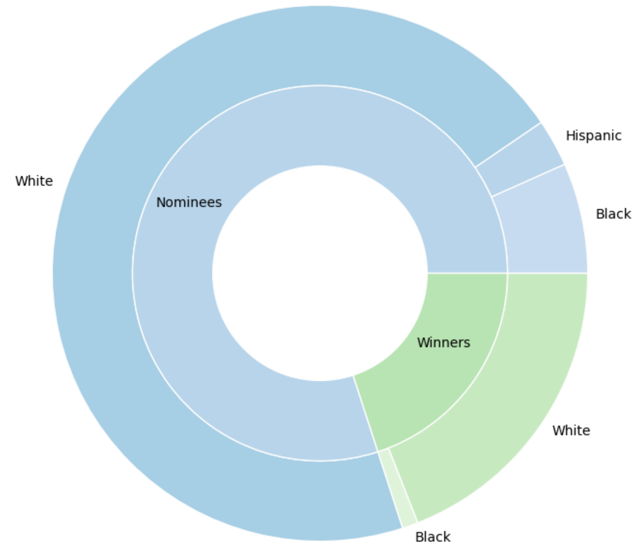

Figure 2: Distribution of Actress in a Leading Role

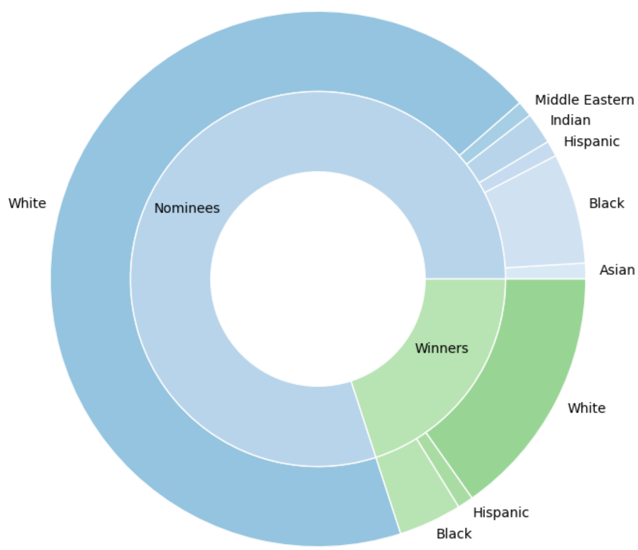

Figure 3: Distribution of Actor in a Supporting Role

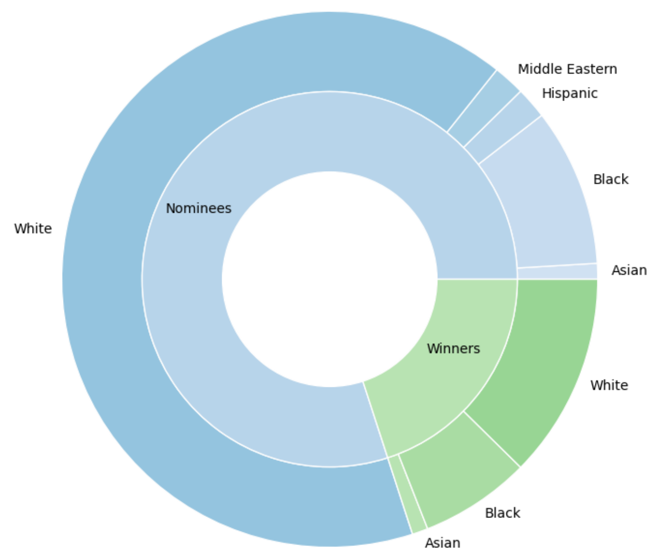

Figure 4: Distribution of Actress in a Supporting Role

In terms of gender, the majority of nominees and winners are males, except for two categories: Costume Design and Makeup and Hairstyling, corresponding to figures 5 and 6 respectively. In Costume Design, females are usually more rewarded and in Makeup and Hairstyling the majority goes to the mix of at least one male and one female.

This turns out to be quite predictable, when we take into account that this two categories are related to tasks traditionally seen as feminine, and which one more time shows how stereotypical this award turns out to be.

Figure 5: Distribution of Costume Design

Figure 6: Distribution of Makeup and Hairstyling

It is also interesting to note that when we focus on the winners (it can also be seen for the nominations but not as clear), apart from the Costume Design and Makeup and Hairstyling categories where it has already been seen that females are more likely to win and the acting categories where gender doesn't play a role, the fact that there are so many male representatives leaves room for different races to be elected (with a clear tendency for caucasians), which cannot happen with females, because the white-preference tendency is even more emphasised due to the low number of winners and this results in just two different races having been elected in the best scenario, being white one of them.

Figure 7: Distribution of Writing (Original Screenplay)

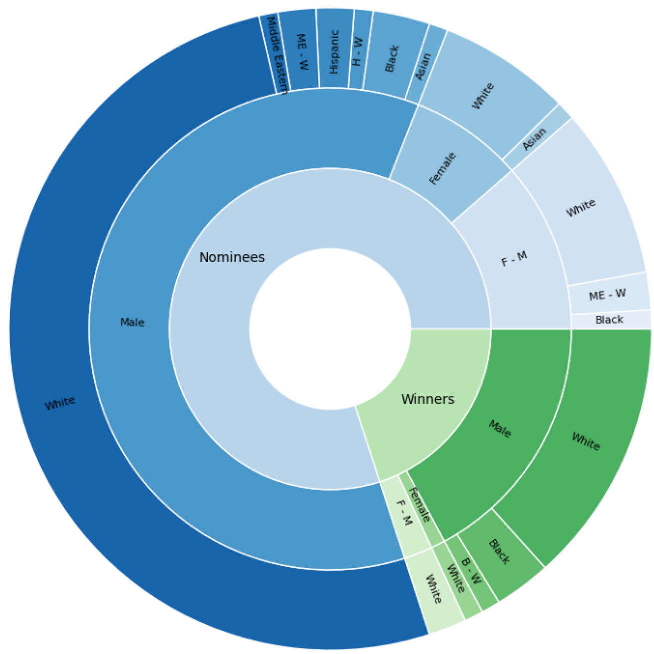

Figure 8: Distribution of Writing (Adapted Screenplay)

Figure 9: Distribution of Music (Original Score)

Figure 10: Distribution of Music (Original Song)

Figure 11: Distribution of Directing

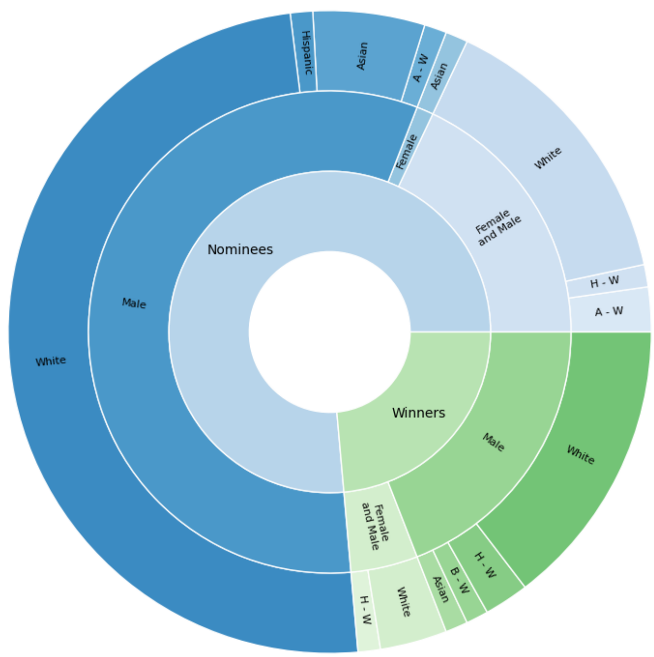

Figure 12: Distribution of Animated Feature Film

For the International Feature Film, the distribution can be seen in the figure 13 and a bar plot by continent has also been created, consisting of a superposition of a blue bar representing the number of nominations and an orange bar representing the number of wins (if the number of nominations and wins were the same, only the color orange would be visible) with the percentage of wins on top of them, which can be seen in the figure 14. It is obvious that Europe has the most winners and nominees but in terms of percentage of wins they are all more or less at the same point.

South America is continent with the highest percentage of wins at 25.0%, followed by Asia and North America. Oceania is the only one with 0.0% of wins, but that is comprehensible as it has only had one film nominated.

Figure 13: Distribution of International Feature Film

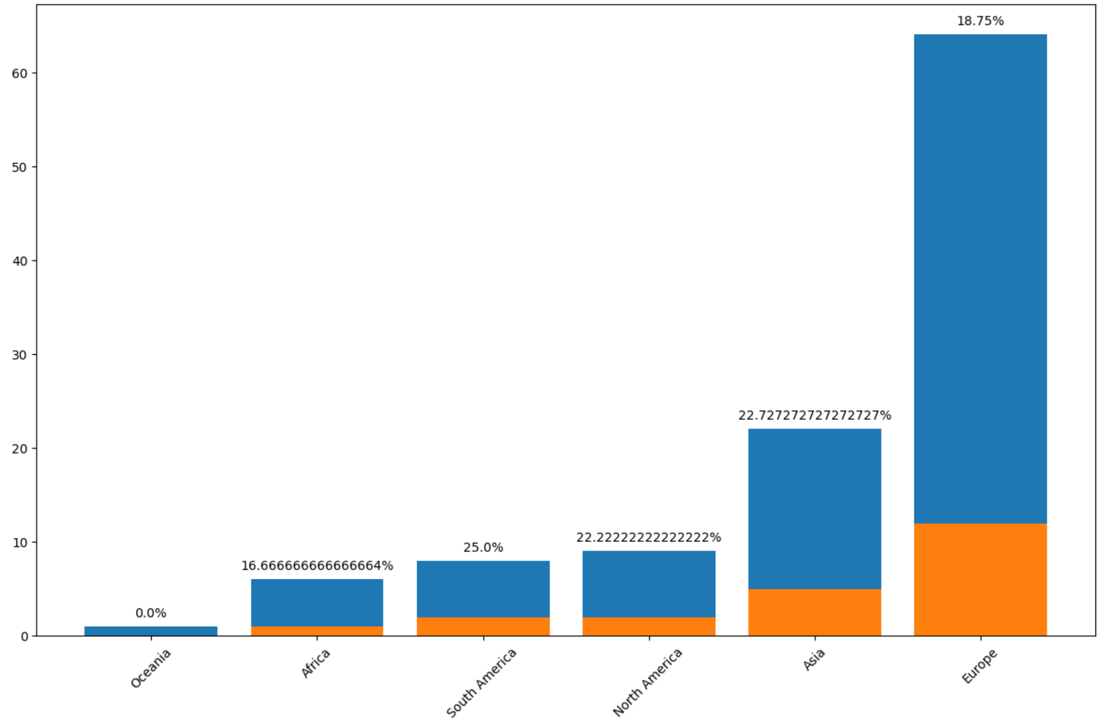

Figure 14: Distribution of wins in International Feature Film across continents

The categories that are missing from this charts but are also included in the prediction model are:

1. Cinematography
2. Documentary Feature Film
3. Documentary Short Film
4. Film Editing
5. Best Picture
6. Production Design
7. Short Film (Animated)
8. Short Film (Live Action)
9. Sound
10. Visual Effects

4 Database

All the information is saved in a non-relational database using MongoDB, and every time that the python script is run, a connection is established and after deleting the database if it already exists, it creates a new one from scratch.

The main reason for using a non-relational database is that there are categories that require certain information that other categories don't need, for example the country in the ”International Feature Film”, or in some categories the race and gender of the nominee. So in order to avoid having the same features for all the entries and filling with NaNs when a feature is irrelevant, it was considered that it would be way better to take advantage of the structure flexibility that a non-relational database provides.

The database begins with the original Academy Awards dataset. From there, all the changes previously mentioned in the Data Cleaning section are done directly to the database and after that, new fields start being added.

The first additions are the attributes winsOscars, nomsOscars, winsGoldenGlobees, nomsGoldenGlobes, winsBaftas and nomsBaftas, which store the number of wins and nominations for every entry in these three awards. This will be saved as a nested object called film_awards and when the data is saved as a dataframe for the analysis and prediction, it will be exploded as six different columns.

In second place, the rating of the films nominated for Best Picutre is added, followed by the race and gender of the people nominated in the categories that we saw in the doughnut charts.

After that, eight fields that will act as flags will be created: the first two will be nom_GG_same_cat and win_GG_same_cat that will have a 1 if there was the name nomination or win respectively in the Golden Globes of that year. The same will happen for the Bafta Awards but only for the Visual Effects and Film Editing categories, and the same for the Visual Effects Society Awards but only for the Visual Effects category. Finally, for the acting categories the same thing will be done but using the SAG award.

It is important to notice that these flags are only created if they need to exist and are set to 1. If the nomination or win doesn't occur in the other award, the flag won't be created in the database, it will only be created when we transform the information in the database into a dataframe, where the flag will be set to 0 in order to avoid issues in the prediction model.

Below, a couple of example can be found of the JSON structure that the final database has. The first one (15) shows an entry for Actor in a Supporting Role, so the attributes ”race”, ”gender” and ”nom_SAG_same_cat” can be found.

Next to it (16), the structure of an entry for Visual Effects is seen, so the attributes ”nom_VES_same_cat” and ”win_VES_same_cat” can be found. This makes evident the idea mentioned before of the use of a non-relational database, where different queries can return outcome with different attributes.


```
[{
  "_id": {
    "$oid": "64ba5d666463d9a61e314584"
  },
  "year_film": 2012,
  "year_ceremony": 2013,
  "ceremony": 85,
  "category": "Actor in a Supporting Role",
  "name": "Robert De Niro",
  "film": "Silver Linings Playbook",
  "winner": false,
  "film_awards": {
    "nomsBaftas": 3,
    "nomsGoldenGlobes": 4,
    "nomsOscars": 8,
    "winsBaftas": 1,
    "winsGoldenGlobes": 1,
    "winsOscars": 1
  },
  "gender": "Male",
  "race": "White",
  "nom_SAG_same_cat": 1
}]
```


Figure 15: JSON for entry in Actor in a Supporting Role


```
[{
  "_id": {
    "$oid": "64ba5d666463d9a61e3148c3"
  },
  "year_film": 2018,
  "year_ceremony": 2019,
  "ceremony": 91,
  "category": "Visual Effects",
  "name": "Dan DeLeeuw, Kelly Port, Russell Earl and Dan Sudick",
  "film": "Avengers: Infinity War",
  "winner": false,
  "film_awards": {
    "nomsBaftas": 1,
    "nomsGoldenGlobes": 0,
    "nomsOscars": 1,
    "winsBaftas": 0,
    "winsGoldenGlobes": 0,
    "winsOscars": 0
  },
  "nom_BAF_same_cat": 1,
  "nom_VES_same_cat": 1,
  "win_VES_same_cat": 1
}]
```


Figure 16: JSON for entry in Visual Effects

5 Prediction Model

One of the main goals of this project is to see if there is a way of creating a prediction model that can be trustworthy enough to make confident predictions. Given that the outcome needs to be True or False, multiple binary classification models such as Logistic Regressions, Gaussian Naive Bayes, Random Forest Classifer and XGB have been tested to see which one gives the best results.

To do so, data from 2002 until 2022 is used to train the model and after that, 2023's nominations are used for the prediction. The main issue was finding a way of being able to choose one and only one winner for every category, and to do so, instead of using the usual function "predict", "predict_proba" was used, which returns the probability attributed to being True or False instead of just giving me the result. From there, the results were sorted from greater to lower probability of being Winner and the first one was chosen as the Winner in the category.

The first approach that was tried involved using only the attributes that were common in all categories so that a larger amount of data could be used for training, but it turned out that it was better to split the prediction by groups of categories and use different sets of attributes every time, as seen below:

1. Visual effects:
    - nom_BAF_same_cat
    - win_BAF_same_cat
    - nom_VES_same_cat
    - win_VES_same_cat
2. Directing:
    - nom_BAF_same_cat
    - win_BAF_same_cat
    - nom_GG_same_cat
    - win_GG_same_cat
    - race
    - gender
3. Best Picture:
    - nom_BAF_same_cat
    - win_BAF_same_cat
    - nom_GG_same_cat

- win_GG_same_cat
- nomsOscars
- nomsGoldenGlobes
- nomsBaftas
- rating

4. Actor/Actress in a Leading Role, Actor/Actress in a Supporting Role:

   - nom_SAG_same_cat
   - win_SAG_same_cat
   - nom_GG_same_cat
   - win_GG_same_cat
   - nomsOscars
   - nomsGoldenGlobes
   - nomsBaftas
   - winsGoldenGlobes
   - winsBaftas
   - race

5. Film Editing:

   - nom_BAF_same_cat
   - win_BAF_same_cat
   - nomsOscars
   - nomsGoldenGlobes
   - nomsBaftas
   - winsGoldenGlobes
   - winsBaftas

6. Production Design, Cinematography, Documentary Feature Film, Documentary Short Film, Short Film (Animated), Short Film (Live Action), Sound:

   - nomsOscars
   - nomsGoldenGlobes
   - nomsBaftas
   - winsGoldenGlobes
   - winsBaftas

nom_BAF_same_cat is a flag for having the same nomination in Baftas (the same goes for wins and for GG meaning Golden Globes, SAG meaning Screen Actors Guild and VES meaning Visual Effects Society).

After trying the multiple prediction models with this data, Gaussian NB turned out to be the best option, as the amount of data wasn't very large, specially in the categories that were predicted separately and because it was easy to encode the race and gender features into discrete values. It correctly predicted 16 out of the 23 categories, and here are some possible reasons to the mispredictions:

1. Costume Design: none of them were nominated in any other award and both designers were Female, but the the one in Everything Everywhere All at Once was white and the one in Black Panther: Wakanda Forever was black, which according to the statistics results in less probability of winning.

2. Makeup and Hairstyling: none of them won any other award but they were both nominated for the Bafta.

3. Music (Original Score): Babylon won the Golden Globe and All Quiet on the Western Front was predicted as third best bus with a difference of just 1%.

4. Best Picture: The Banshees of Inisherin won the Golden Globe and Everything Everywhere All at Once was predicted as second best.

5. Sound: All Quiet on the Western Front won the Bafta and Top Gun: Maveric wasn't even nominated.

6. Writing (Adapted Screenplay): All Quiet on the Western Front won the Bafta and the writers were a combination of Male and Female, and the writer of Women Talking was a woman, which according to the statistics results in less probability of winning. Howerver, it was predicted as second best.

7. Writing (Original Screenplay): The Banshees of Inisherin won the Golden Globe and the Bafta and the writer was White, and the writers in Everything Everywhere All at Once were a Asian and White, which according to the statistics results in less probability of winning. However, it was predicted as second best.

The decision of the model and the features used for each category wasn't just made based on this year's results. Multiple iterations were done across other years and this ended up being the best solution in global. This way, this code can be used in future predictions and not just for this project.

6 Results of the Prediction

The outcome of the predictions generated by the model can be found in the table below 1. The columns of the table contain the information of the category being predicted, the true winner, the predicted winner, the place in which the winner is predicted and the probability of win attributed to the winner.

Category	Winner	Prediction	Place of Prediction	Probability of Win Given by the Model
Leading Actor	Brendan Frasier	Brendan Frasier	1	0.9910967054256101
Leading Actress	Michelle Yeoh	Michelle Yeoh	1	0.999996600245036
Supporting Actor	Ke Huy Quan	Ke Huy Quan	1	0.999996600245036
Supporting Actress	Jamie Lee Curtis	Jamie Lee Curtis	1	0.9995872930673726
Animated Film	Guillermo del Toro's Pinocchio	Guillermo del Toro's Pinocchio	1	0.9999965156677981
Cinematography	All Quiet on the Western Front	All Quiet on the Western Front	1	0.9998460161482006
Costume Design	Black Panther: Wakanda Forever	Everything Everywhere All at Once	4	0.010891848843604621
Directing	Everything Everywhere All at Once	Everything Everywhere All at Once	1	1.0000000000000000
Documentary	Navalny	Navalny	1	0.03998163454505988
Documentary Short	The Elephant Whisperers	The Elephant Whisperers	1	0.032705692617109504
Film Editing	Everything Everywhere All at Once	Everything Everywhere All at Once	1	0.9978321711479365
International Film	All Quiet on the Western Front	All Quiet on the Western Front	1	0.9966432348626613
Makeup and Hairstyling	The Whale	All Quiet on the Western Front	3	0.024987979409470822
Original Score	All Quiet on the Western Front	Babylon	3	0.989167010224906
Original Song	RRR	RRR	1	0.9999840351172419
Best Picture	Everything Everywhere All at Once	The Banshees of Inisherin	2	0.5425237068279831
Production Design	All Quiet on the Western Front	All Quiet on the Western Front	1	0.9998460161482006
Short Film (Animated)	The Boy, the Mole, the Fox and the Horse	The Boy, the Mole, the Fox and the Horse	1	0.03998163454505988
Short Film (Live Action)	An Irish Goodbye	An Isish Goodbye	1	0.03998163454505988
Sound	Top Gun: Maverick	All Quiet on the Western Front	3	0.07100749756217425
Visual Effects	Avatar: The Way of Water	Avatar: The Way of Water	1	0.9999999999990346
Adpt. Screenplay	Women Talking	All Quiet on the Western Front	2	0.015589129238133949
Org. Screenplay	Everything Everywhere All at Once	The Banshees of Inisherin	2	0.9099785339654741

Table 1: Results vs. Predictions of the Model

7 Problems Encountered

When this project started, previous research had been done on the datasets and python libraries that could be useful in order to avoid having to create everything manually, which would have taken a huge amount of time. One of the best resources that was found was a library called CinemaGoer.

It consisted on a library that used the IMDB database to get all sorts of informations, not only for films, but also for actors, actresses and directors. After working with it for two months and developing the code to get information of nominations in other awards (it can still be found in the python code in case it is available again in the future), the information was erased from the library. A few months later, a new dataset of the Golden Globes was added to Kaggle but nothing else. The rest was created from scratch, so only the categories that seemed to have influence in the prediction were added.

Finally, the last problem appeared when trying to evaluate the predicting model. Usually, a set of the data is used to train the model and then the rest is used for testing, using methods as the Accuracy, Precision, F1 or R2 score to evaluate how good our model is working. The problem here was that for each year and category, one and only one winner needed to be selected which could only be done by comparing the probabilities of being classified as winner for every nominee.

8 Conclusion

The two main goals of this project were proving that race and gender really play a huge part in the decision of the Academy Awards winners, which has been clearly proven in the part of the exploration of the data, and the second one was doing a prediction of the 2023 winners, which wasn't completed as successfully as it was intended, having only 16 correct predictions out of the total of 23.

However, if we take into account the limited amount of data that could be used and that the mispredictions were quite logical and predictable, it can be argued that the results were somewhat decent. The idea is that this project doesn't end here, because the python code used is uploaded in `https://github.com/Gerard3443/TFG` and can be reused for future ceremonies by adding the information of the new nominees and changing just a few parameters.

9 References

[1] Raphael Fontes, David Lu (2023). The Oscar Award, 1927 - 2023 [Data set]. Kaggle. https://www.kaggle.com/datasets/unanimad/the-oscar-award

[2] Raphael Fontes (2023). Golden Globe Awards, 1944 - 2020 [Data set]. Kaggle. https://www.kaggle.com/datasets/unanimad/golden-globe-awards

[3] Raphael Fontes (2023). British Academy of Film and Television Arts Awards [Data set]. Kaggle. https://www.kaggle.com/datasets/unanimad/bafta-awards

[4] Screen Actors Guild Awards (2023). The 29th Annual Screen Actors Guild Awards [website]. https://www.sagawards.org/awards/nominees-and-recipients/29th-\annual-screen-actors-guild-awards?field_taxonomy_vocabulary_2_target_id=4

[5] Visual Effects Society (2023). Visual Effects Society Announces Winners of the 21st Annual VES Awards [website]. https://www.vesglobal.org/press-releases/visual-effects-society-announces-winners-of-the-21st-annual-ves-awards

[6] Cinemagoer. Cinemagoer [website]. https://cinemagoer.github.io

[7] Rebecca Keegan, Ben Zauzman (2022). Inside the Academmy's Inclusion Drive. https://www.hollywoodreporter.com/movies/movie-features/the-\academy-inclusion-data-1235116216

[8] Ben Zauzman (2023). Predicting This Year's Oscar Winners Using Just Math. https://www.hollywoodreporter.com/movies/movie-news/oscars-\predictions-2023-math-ben-zauzmer-1235342656

[9] Kenneth Grout, Owen Eagan (2021). Oscar Is a Man: Sexism and the Academy Awards. https://doi.org/10.51698/tripodos.2020.48p85-102