Final Report: Women In Film

Natalia Coronado, Misty Peng, Kamila Soltynska

Introduction

The topic our group decided to conduct our research on was the representation of women in film. Throughout time, women have been greatly underrepresented in the industry. Since the academy awards began in 1929, only five women have been nominated for the best director category, two of which have won. (Brown, 2022) In a study conducted in 2019 by the USC Annenberg Inclusion initiative, 1,300 popular films from 2007-2019, were examined for their inclusivity of often marginalised groups. The study found that across 1,518 content creators, only 10.7% of directors were women, 19.4% were writers, and 24.3% were producers. (Inclusion Initiative) This substantive disparity between the number of men and the number of women behind the camera, has led filmmaking to be a greatly male dominated industry, in which female led narratives are cast to the side, and not given the opportunity to succeed. This leads to inaccurate depictions of women in film, in which they are oftentimes shown fulfilling harmful stereotypical roles that negatively affect the way society views women..

In 1985, the Bechdel Test was created as a measure of gender depiction in film, and has since been used by many to determine the level of women's representation in movies. In order to pass the test, the film must get through three requirements; there must be at least two women in the movie, they must talk to each other, and they must talk about something other than a man. In an article published in 2021, it was stated that 40% of movies still do not pass the test, 10% don't even reach the first requirement. (Viswanathan, 2021) Why is the depiction of women in film so important? Essentially, it can be seen as a fight against inequality. As social cognitive psychotherapist April Seifert said, "the stereotypes we see in movies, books, and other media unconsciously shape our worldview. — The more strong women we see in films and other places, the more our brains will start to automatically associate 'strength' as a feminine trait, and as a result, the more women will be treated as strong, equal members of society." (Inclusion Initiative) In order to turn this into a machine learning problem, we set out to discover what factors contribute the most to whether a movie passes the Bechdel test or not. By using a decision tree approach, we will be able to look closely at the components which most have an effect on the performance on the test, and in essence, what components lead to better representation of women in film, such as, what movie genres, or which country. We expect to see high information gain in the categories that most contribute to the passing or failing of the test, possibly restriction (one of the stereotypical roles given to women in films is that of a temptress or housewife, two roles often seen in more dramatic films geared towards older audiences), or imdb rating, if movies with higher scores tend to pass the test more, or less, often.

Method

The dataset was acquired via Data.World (Data.World, 2022) and consisted of information regarding 1453 movies released from 2010 to 2014. This dataset expounded on 26 features for every movie from which eight most pertinent features were selected. This process involved setting and following criteria for inclusion. For example, the 'movie title' column (containing irrelevant textual data), the 'IMDb popularity score' (using an unclear measure), and the 'nominations' feature (which was similar to the 'award wins' category), were removed (Appendix A).

Subsequently, data within the remaining column features was modified from multiple nominal categories or qualitative data into suitable categories. The remaining columns include 'genre', 'restriction rating', 'runtime', 'language', 'IMDb rating', 'award wins', 'country', and the 'Bechdel test result.' First, the 'genre' column features were reduced from 22 possibilities to two for the initial decision tree, and four for the finalised decision tree: 'adventure,' 'nonfiction,' 'light,' and 'dark'. All other column data was recategorized into binary separations:

- *Restriction* rating became 'restricted' for 18+ audiences and 'not restricted' for anything under 18 based on Motion Picture Association ratings originally provided in the dataset
- *Runtime* became 'normal' if between 90-150 minutes and 'abnormal' for any runtimes outside of that window
- *Language* became 'English' if English was in the list of languages and 'not English' only if English did not appear at all on the list of languages
- Similarly, *Country* became 'USA' if the USA appeared at least once on the list of countries and 'not USA' only if USA was not included in the list at all
- *IMDb rating* became 'low' and 'high,' with low representing all movies with a rating below 6.4 that is the site average (citation)
- *Wins* became 'awards-winning' if the movie won two or more rewards and 'not awards-winning' if the movie won only one or none
- *Bechdel test* was already binary and remained 'Yes' or 'No'

The decision tree was implemented in Excel using the same methodology as Assignment 4. We counted the number of Bechdel test passes and fails for the data and calculated the information gain from each of the six categories. We utilized the filter feature to update the data according to the branch path leading to each node. Then, using the attribute with the highest information gain from the new set of data, we continued to the next branch unless a stopping point of ≥70% was reached, at which point the node would end as 'Passes' or 'Does Not Pass.' Otherwise, if five branch levels produced no significant results or the quantity of the data became too small, a branch was labeled 'Inconclusive.' We created one tree using this process for US movies and another for non-US to concur with the aim of comparing decision trees of movies made in different countries. The decision trees are visualized via SmartDraw (SmartDraw, 2022) in order to project clear paths through the decision trees. We commenced this process with a binary genre categorization ('dramatized' vs 'lifelike'). In an attempt to improve methodology following evaluation and produce more insightful results, a new pair of decision trees was constructed with a quaternary categorization.

Results and Discussion

We set our decision threshold at ≥ 70% and were able to achieve end nodes for 78% of the leaf nodes in the non-US tree (Fig B.1) and 65% of the leaf nodes in the US tree (Fig C.1). A majority of our branches ended in a decision with this threshold, but raising it to 75% would drastically decrease the number of decisions we could reach within six decision levels.

This shows that the performance of our decision tree can return meaningful results but is not excellent. The first set of trees we calculated separated genre into two categories; we changed this split to four ways to retain more of the original genre integrity of 22 categories.

Figure D shows the breakdown of how each feature category performed in each tree while Figure E focuses on how genre categories affected what decision the trees reached. For example, the 'light' category often resulted in a continuing node or 'no data/inconclusive' while the 'dark' category resolved into either 'pass' or 'fail' within two levels. This could be a result of 'light' movies being a mix of movies that pass and fail the Bechdel test while 'dark' movies have specific characteristics that we measured to determine how they perform on the test. On the other hand, the dark category in the US tree took the longest to resolve, eventually ending in multiple no data/inconclusive nodes, a result potentially from more data falling into the US category as opposed to non-US.

Our original trees with two genre categories resulted in node paths that did not make much intuitive sense. After adjusting our calculations to take into account four separate genre categories, we found interesting results. One of the most surprising results occurred in the non-US tree where the four-category feature produces the highest information gain. After splitting the root node into the four categories, the 'adventure' category immediately resulted in 'passes'. Non-US 'adventure' films include movies like 'Arrietty (2010)' and 'Xi you xiang mo pian (Journey to the West) (2013)', making up a category that are considered to all pass the Bechdel test. On the other hand, US 'adventure' films ended in 'pass' nodes 25% of the time and the remaining 75% were either 'continuing' or 'no data/inconclusive' nodes. Overall, the US tree had fewer conclusive nodes (Fig E). This was an unexpected result because there was more data from the US rather than outside of the US. The trees we built produced interesting pathways of representation for women in films. Using this dataset and the decision tree algorithm, we were able to create a connection from unlikely features such as IMDb rating and language to whether a movie passes the Bechdel test.

Future Direction

Some of the challenges we encountered had to do with the data itself, when choosing the nodes for our decision trees. After choosing the features which could have a greatest impact on the passing or failing of the test, we had to come up with ways to split the categories in half, such as low and high imdb scores and normal vs abnormal runtime. Outside research had to be conducted in order to find a meaningful split. (Johnston, 2009) Another challenge was the genre category, and the many genres it contained, 22 in total. Initially, we had chosen two categories, dramatised and non-dramatised, however, the genres are much more complex and diverse from each other so in our second decision trees we had to find a way to separate the category into four meaningful distinctions. Due to this change, the template for the first decision tree calculations had to be updated to fit this new separation of the data. The amount of data we had was also fairly small, so when creating the decision trees, some nodes did not contain enough information, leading to inconclusive results. In order to improve this, we could run the tests again with data containing a greater number of movies, possibly from a larger time frame than 2010-2014.

If we had an unlimited number of resources, perhaps we could use a different approach to investigating our research question, this approach being a generalized regression model and a random forest model, as done by researchers at the University of Rochester. (Yang et al., 2020) Linear regression would allow us to generate more concrete statements about the effect each category has on the passing or failing of the Bechdel test. The random forest model can be used to achieve data that is not highly dependent on the previous nodes, as is the case with the original decision tree, and its performance can be compared to the linear regression's performance, as it was in the aforementioned study. Some future study ideas would be, how are movies in which women are represented fairly perceived by the public, which could be done using features such as critics rating, audience rating, and marketing budgets. In addition we could research how movies directed and produced by women differ from those led by men. How large is the gap between the two and which aspects of female representation are affected the most, such as the number of lines the women in the movie have, or simply the number of female characters present.

Appendix

Appendix A - Rationale for the selection of dataset features

Feature name	**Included?**	**Reasoning and how the new data will be grouped**
Movie Title	N	Irrelevant, textual data.
Passes Bechdel Test?	Determiner	Used to determine if tree branches should terminate.
Number of Criteria Passed	N	Related to the Bechdel test, lacks relevance.
Clarity of Pass	N	Related to the Bechdel test, lacks relevance.
Has at least two [named] women in it	N	Criteria for the Bechdel test, lacks relevance.
The women talk to each other	N	Criteria for the Bechdel test, lacks relevance.
The women talk about something besides a man	N	Criteria for the Bechdel test, lacks relevance.
Genre	Y	Grouping for the old decision tree: Separated into two categories: dramatized and lifelike, based on the realism of the genres. If a movie falls into both, it is considered dramatized. Grouping for the new decision tree: Separated into four categories: light, dark, nonfiction, adventure
Director	N	Irrelevant, textual data.
Main Actors	N	Irrelevant, textual data.
Production Company	N	Too many variables to be meaningfully reduced.
Plot Keywords (top 10)	N	Too many variables to be meaningfully reduced.
Year	N	Difficulty in reduction into meaningful categories due to low temporal range (2010-2014)
Rating	Y	Grouping into restricted vs non-restricted categories
Runtime (min)	Y	Grouping into normal vs abnormal (90-150 min) categories. Movies within a 90 and 150 (90 and 150 included) minute range are labeled as normal. Movies falling outside of that range are considered abnormal. Range criteria chosen because most blockbuster movies fall within it.
Plot	N	Too many variables to be meaningfully reduced.
Language	Y	Grouped into English and non-English
Country	Y	Grouped into US and not-US categories.
Image	N	Data is just a link
IMDb Popularity Score	N	Unclear what scale is used to measure popularity.

Award Popularity	N	Unclear what scale is used to measure popularity of awards.
IMDb rating	Y	Grouped into high and low ratings, where scored above 6.4 are high and below are low. The split is determined by the average IMDb rating.
IMDb votes	N	Too similar/correlated with the Wins category
Nominations	N	Too similar/correlated with the Wins category
Wins	Y	Grouped into 0-1 and 2+ categories. Seperates movies into category of one or no awards and multi awards-winning movies
IMDB Link	N	Data is just a link

Appendix B - Non-US Decision Tree

Fig B.1 Whole Tree

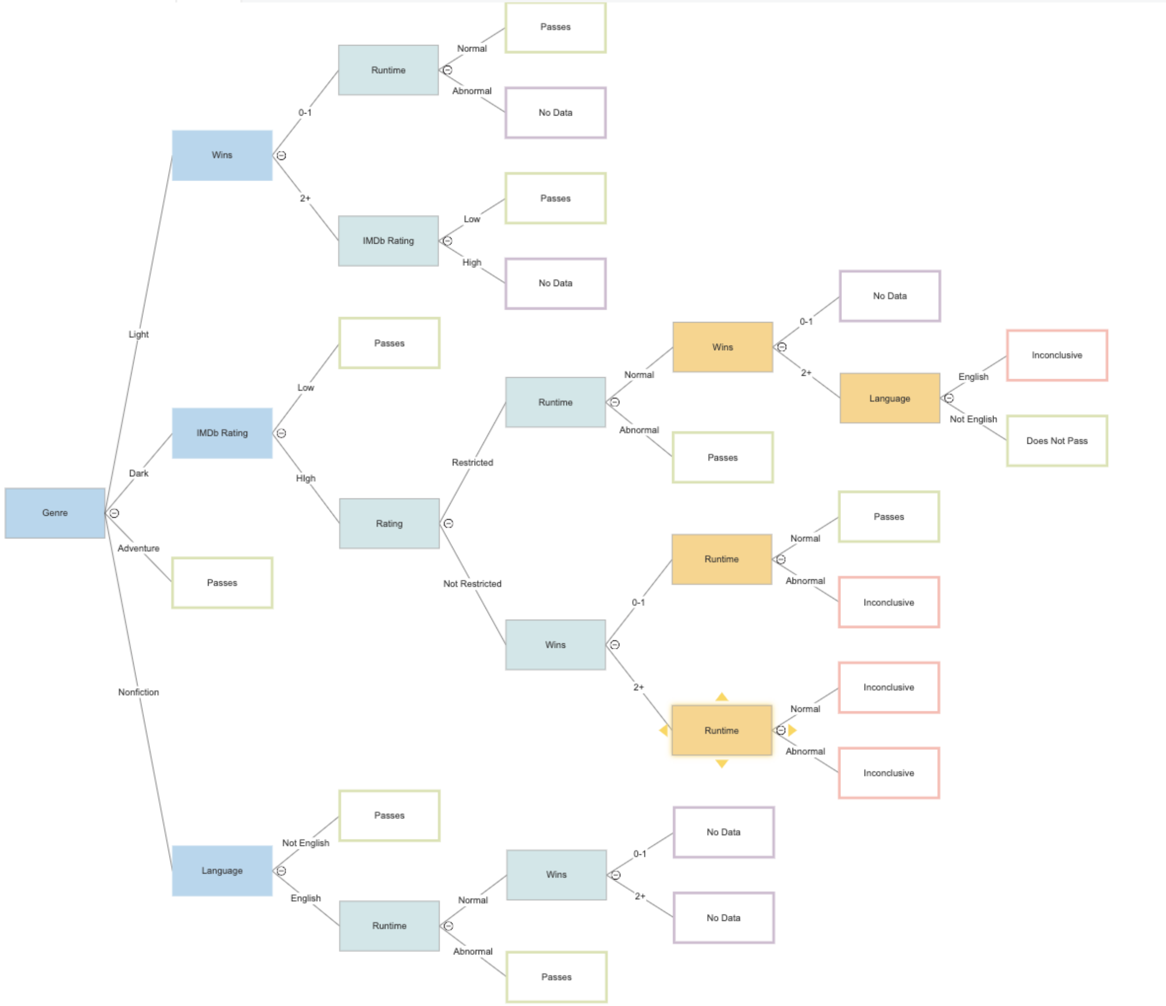


Fig B.2 Adventure branch

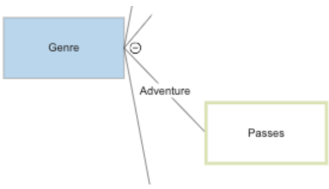


Fig B.3 Nonfiction branch

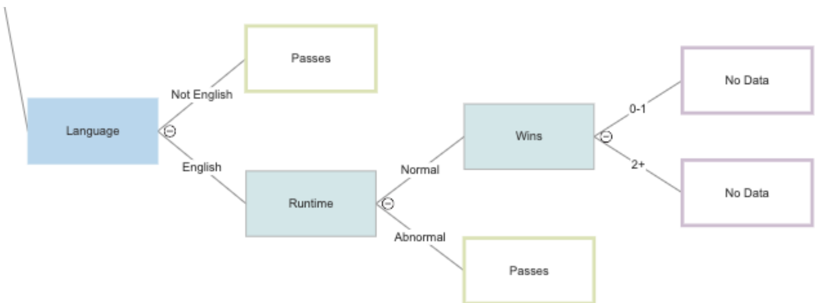


Fig B.4 Dark branch

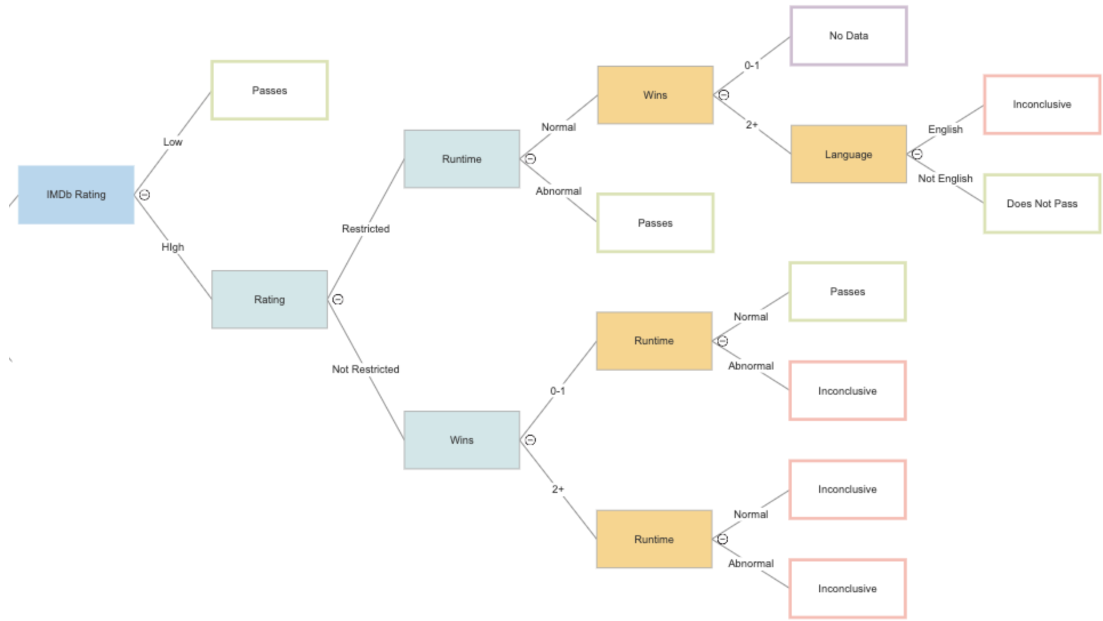
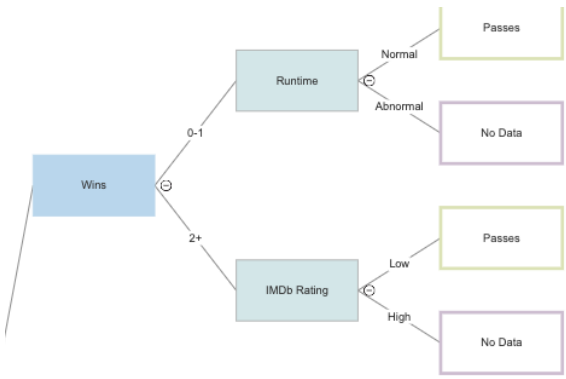


Fig B.5 Light branch



Appendix C - US Decision Tree

Fig C.1 Whole tree

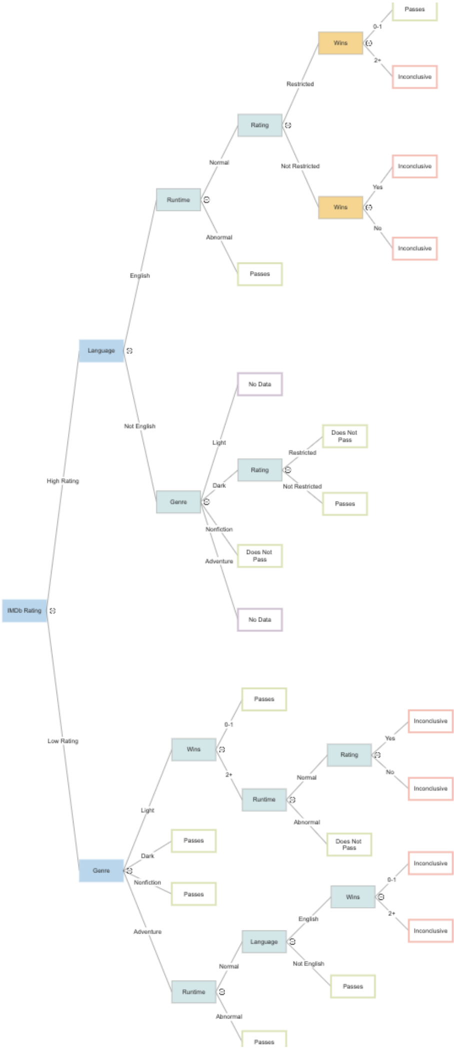


Fig C.2 High IMDb rating branch

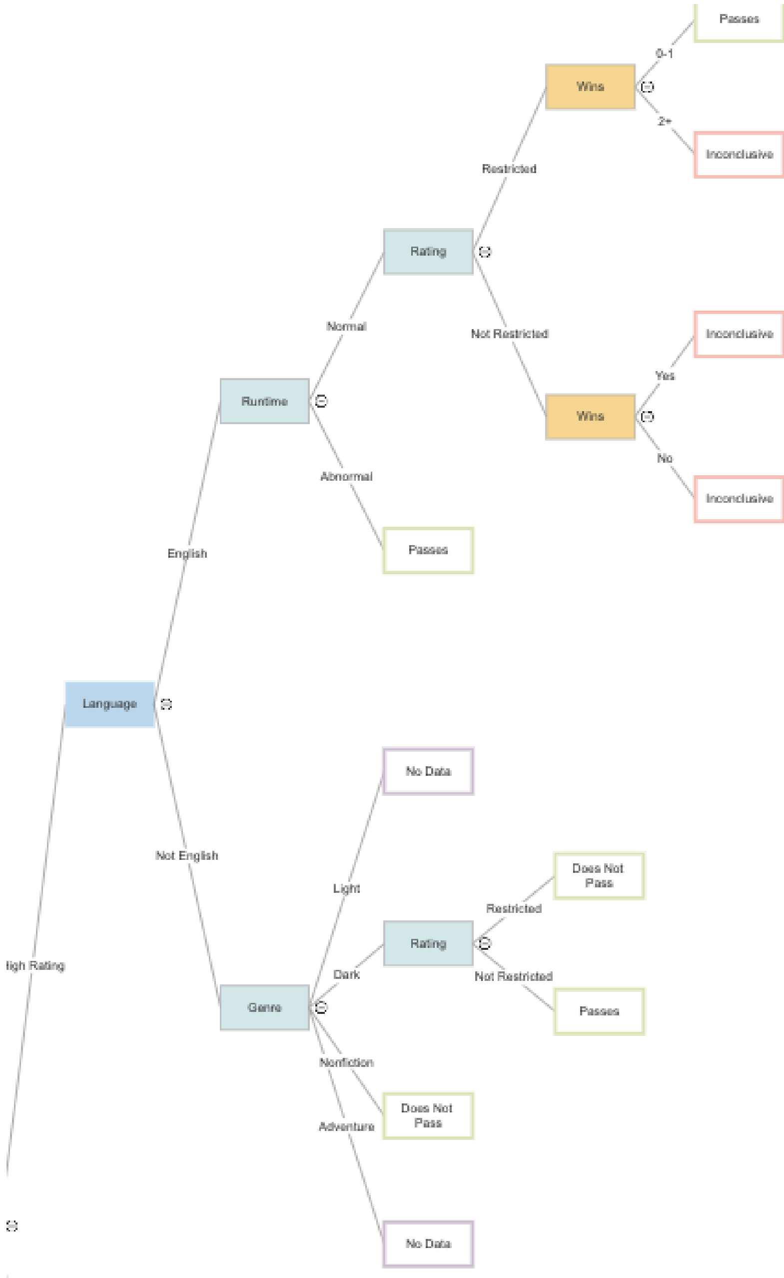
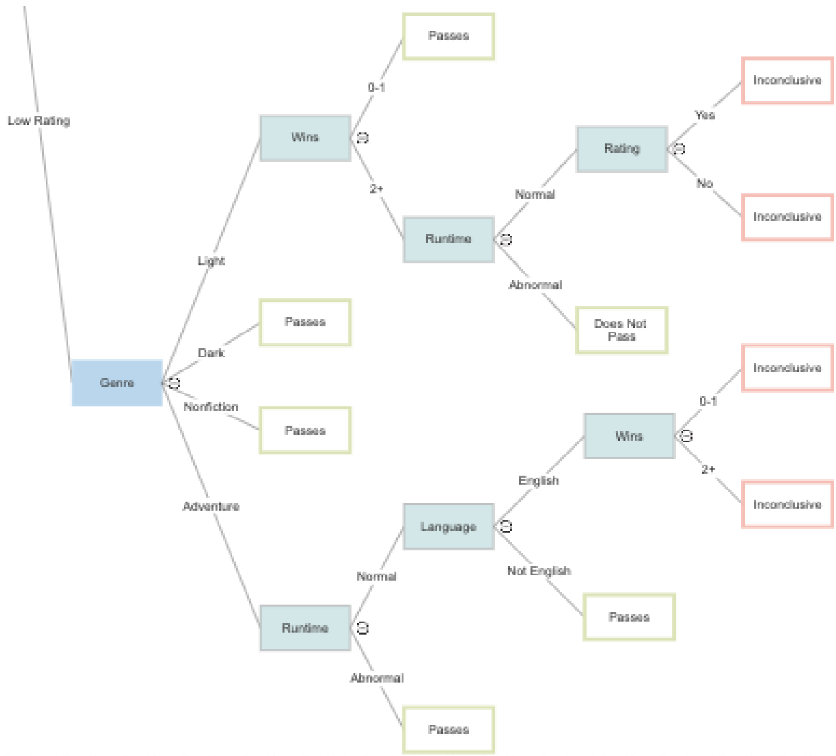


Fig C.3 Low IMDb rating branch



Appendix D - A color-coded chart visualization to indicate which nodes follow which features most often.

Feature	Category	Non US						US				
		pass	fail	continue	no data/ inconclusive	total		pass	fail	continue	no data/ inconclusive	total
	Abnormal	2	0	0	3	5		0	0	3	0	3
Runtime	Normal	2	0	2	1	5		2	1	0	0	3
	Not English	1	1	0	0	2		1	0	1	0	2
Language	English	0	0	1	1	2		0	0	2	0	2
IMDb Rating	high	0	0	1	1	2		0	0	1	0	1
	low	2	0	0	0	2		0	0	1	0	1
	awards-winning	0	0	2	2	4		0	0	1	3	4
Wins	not awards-winning	0	0	3	1	4		2	0	0	2	4
	light	0	0	1	0	1		0	0	1	1	2
	dark	0	0	1	0	1		1	0	1	0	2
	nonfiction	0	0	1	0	1		1	1	0	0	2
Genre	adventure	1	0	0	0	1		0	0	1	1	2
	restricted	0	0	1	0	1		0	1	1	1	3
Rating	not restricted	0	0	1	0	1		1	0	1	1	3

color code	percent of nodes
	0-24
	25-49
	50-99
	100

Appendix E - A color-coded chart visualization to follow what happens to the nodes following each genre category.

	Genre (US)					Genre (non US)			
level 1	light	dark	nonfiction	adventure		light	dark	nonfiction	adventure
pass	0	1	1	0	pass	0	0	0	1
fail	0	0	1	0	fail	0	0	0	0
continue	1	1	0	1	continue	1	1	1	0
no data/ inconclusive	1	0	0	1	no data/ inconclusive	0	0	0	0
total	2	2	2	2	**total**	1	1	1	1

level 2	light	dark	nonfiction	adventure		light	dark	nonfiction	adventure
pass	1	1	n/a	1	pass	0	1	1	n/a
fail	0	1	n/a	0	fail	0	0	0	n/a
continue	1	0	n/a	1	continue	2	1	1	n/a
no data/ inconclusive	0	0	n/a	0	no data/ inconclusive	0	0	0	n/a
total	2	2	n/a	2	**total**	2	2	2	n/a

level 3	light	dark	nonfiction	adventure		light	dark	nonfiction	adventure
pass	0	n/a	n/a	1	pass	2	0	1	n/a
fail	1	n/a	n/a	0	fail	0	0	0	n/a
continue	1	n/a	n/a	1	continue	0	2	1	n/a
no data/ inconclusive	0	n/a	n/a	0	no data/ inconclusive	2	0	0	n/a
total	2	n/a	n/a	2	**total**	4	2	2	n/a

level 4	light	dark	nonfiction	adventure		light	dark	nonfiction	adventure
pass	0	n/a	n/a	0	pass	n/a	1	0	n/a
fail	0	n/a	n/a	0	fail	n/a	0	0	n/a
continue	0	n/a	n/a	0	continue	n/a	3	0	n/a
no data/ inconclusive	2	n/a	n/a	2	no data/ inconclusive	n/a	0	2	n/a
total	2	n/a	n/a	2	**total**	n/a	4	2	n/a

level 5	light	dark	nonfiction	adventure		light	dark	nonfiction	adventure
pass	n/a	n/a	n/a	n/a	pass	n/a	1	n/a	n/a
fail	n/a	n/a	n/a	n/a	fail	n/a	0	n/a	n/a
continue	n/a	n/a	n/a	n/a	continue	n/a	1	n/a	n/a
no data/ inconclusive	n/a	n/a	n/a	n/a	no data/ inconclusive	n/a	4	n/a	n/a
total	n/a	n/a	n/a	n/a	**total**	n/a	6	n/a	n/a

level 6	light	dark	nonfiction	adventure		light	dark	nonfiction	adventure
pass	n/a	n/a	n/a	n/a	pass	n/a	0	n/a	n/a
fail	n/a	n/a	n/a	n/a	fail	n/a	1	n/a	n/a
continue	n/a	n/a	n/a	n/a	continue	n/a	0	n/a	n/a
no data/ inconclusive	n/a	n/a	n/a	n/a	no data/ inconclusive	n/a	1	n/a	n/a
total	n/a	n/a	n/a	n/a	**total**	n/a	2	n/a	n/a

Works Cited

Brown, T. (2022, March 27). *How many women have won the directing Oscar?* Los Angeles Times. Retrieved May 11, 2022, from https://www.latimes.com/entertainment-arts/movies/story/2022-03-27/oscars-2022-women-best-director-winners

Economy, C. (2015, July). *What are the Benefits of Watching Strong Women in Film?* Culture and Creativity. Retrieved May 12, 2022, from https://www.culturepartnership.eu/en/article/women-film

Inclusion Initiative, A. (n.d.). *Inequality in 1,300 popular films - USC Annenberg school for* ... Retrieved May 12, 2022, from https://assets.uscannenberg.org/docs/aii-inequality_1300_popular_films_09-08-2020.pdf

Johnston, N. (2009, October 9). *IMDb movie ratings over the years*. Nathaniel Johnston. Retrieved May 12, 2022, from http://www.njohnston.ca/2009/10/imdb-movie-ratings-over-the-years/

"SmartDraw." *SmartDraw - Create Flowcharts, Floor Plans, and Other Diagrams on Any Device*, https://www.smartdraw.com/?id=104640&gclid=CjwKCAjw9-KTBhBcEiwAr19igz1a9lvYA0UcS5Qr9gEGFHjYaCJR91vw_vvv_INIAIDYQ6pCjKwWiBoCWMQQAvD_BwE. Accessed 9 May 2022.

Viswanathan, G. (2021, July 14). *What is the Bechdel test and what is its relevance to today's film industry?, Nightingale*. Nightingale. Retrieved May 12, 2022, from https://nightingaledvs.com/what-is-the-bechdel-test-and-what-is-its-relevance-to-todays-film-industry/

"Women-in-Film - dataset by alice-c." *Data.World*, https://data.world/alice-c/women-in-film. Accessed 9 May 2022.

Yang, L., Xu, Z., & Luo, J. (2020). Measuring female representation and impact in films over time. *ACM/IMS Transactions on Data Science*, *1*(4), 1–14. https://doi.org/10.1145/3411213