

Análise Temporal Competitiva e Narrativa em RuPaul's Drag Race por Meio de Machine Learning

Kaique de Oliveira e Silva¹, João Guilherme Maia de Menezes¹

¹ Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

{kaiqueoliveir0}@dcc.ufmg.br

Abstract. *Competitive reality television combines objective performance and narrative construction, making outcome prediction a particularly challenging problem. This work investigates the temporal evolution of competitive and narrative patterns associated with the winners of RuPaul's Drag Race using machine learning techniques. Building upon the dataset developed in MSI I, a contestant-episode representation was created to track contestants' trajectories throughout the seasons. Competitive, narrative, visibility, and characterization variables were incorporated and evaluated using a Leave-One-Season-Out (LOSO) validation strategy. Experimental results showed that the best performance was achieved by a Balanced Logistic Regression model combined with preprocessing, feature selection, and temporal recency weighting techniques. The final model achieved 77.8% Top 1 accuracy, 94.4% Top 2 accuracy, and 100.0% Top 3 accuracy across the eighteen regular seasons analyzed. Beyond predictive performance, the temporal trajectory analysis revealed recurring patterns in winner construction, characterized by the gradual consolidation of winning probabilities throughout the competition. The results suggest that the interaction between competitive performance and narrative construction generates sufficiently consistent signals to identify potential winners before the end of a season.*

Keywords: *machine learning, predictive analytics, reality television, RuPaul's Drag Race, temporal analysis, Edgic.*

Resumo. *Reality shows competitivos combinam desempenho objetivo e construção narrativa, tornando a previsão de seus resultados um problema particularmente complexo. Este trabalho investiga a evolução temporal dos padrões competitivos e narrativos associados às vencedoras de RuPaul's Drag Race por meio de técnicas de aprendizado de máquina. A partir da base de dados desenvolvida em MSI I, foi construída uma representação temporal no formato participante-episódio, permitindo acompanhar a trajetória das competidoras ao longo das temporadas. Foram incorporadas variáveis competitivas, narrativas, de visibilidade e de caracterização das participantes, avaliadas por meio de uma estratégia de validação Leave-One-Season-Out (LOSO). Os experimentos demonstraram que a melhor configuração foi obtida por um modelo de Regressão Logística Balanceada associado a técnicas de pré-processamento, seleção de atributos e ponderação temporal por recência. O modelo alcançou 77,8% de acerto Top 1, 94,4% de acerto Top 2 e 100,0% de acerto Top 3 nas dezoito temporadas analisadas. Além do desempenho preditivo, a análise das trajetórias*

temporais revelou padrões recorrentes na construção das vencedoras, caracterizados pela consolidação gradual de suas probabilidades de vitória ao longo da competição. Os resultados indicam que a combinação entre desempenho competitivo e construção narrativa produz sinais suficientemente consistentes para permitir a identificação de futuras vencedoras antes do encerramento das temporadas.

Palavras-chave: *aprendizado de máquina, análise preditiva, reality shows, RuPaul's Drag Race, análise temporal, Edgic.*

1. Introdução

O gênero televisivo da *Reality TV* abrange uma ampla variedade de formatos, incluindo programas de competição, relacionamentos e estilo de vida. Em comum, esses programas buscam registrar interações humanas espontâneas, permitindo que o público acompanhe indivíduos reais em situações reais ou apresentadas como tal. Essa característica diferencia os realities de formatos ficcionais tradicionais e contribui para a criação de uma forte conexão entre audiência e participantes [Butler 2020].

Paralelamente, a análise de dados e as técnicas de aprendizado de máquina vêm sendo aplicadas em um número crescente de áreas. Empresas utilizam modelos preditivos para compreender o comportamento de consumidores, enquanto organizações esportivas empregam métodos estatísticos para avaliação de desempenho, definição de estratégias e suporte à tomada de decisão. Em ambos os casos, o objetivo é identificar padrões capazes de explicar eventos passados e antecipar resultados futuros [Davenport 2006, Mondello and Kamke 2014, Morgulev et al. 2018].

No contexto dos reality shows competitivos, a predição de resultados apresenta desafios particulares. Diferentemente de competições estritamente esportivas, o desempenho dos participantes não é determinado exclusivamente por indicadores objetivos. Aspectos relacionados à narrativa construída pela edição, à visibilidade concedida aos competidores e à percepção gerada junto ao público também exercem influência significativa sobre a trajetória dos participantes ao longo da temporada.

Entre os programas que exemplificam essa dinâmica destaca-se *RuPaul's Drag Race*, competição televisiva em que artistas drag disputam uma série de desafios semanais até a definição de uma vencedora. O programa se consolidou como um fenômeno cultural de alcance internacional, articulando performance, identidade, representação e competição televisiva [Chan 2023]. Embora o desempenho competitivo seja um componente fundamental da avaliação, a construção narrativa das participantes desempenha papel igualmente relevante. Dessa forma, a trajetória de uma competidora é influenciada tanto por seus resultados nos desafios quanto pela maneira como sua história é apresentada ao longo da temporada.

Durante a pesquisa desenvolvida em MSI I, foi construída uma base de dados contendo informações competitivas, narrativas e de caracterização das participantes das temporadas regulares de *RuPaul's Drag Race*. A partir dessa base, foram avaliados modelos de aprendizado de máquina voltados para a identificação de padrões associados às vencedoras da competição. Os resultados obtidos demonstraram que variáveis narrativas e competitivas apresentam capacidade preditiva significativa quando analisadas de forma conjunta.

Entretanto, a abordagem desenvolvida em MSI I considerava apenas informações consolidadas ao final das temporadas. Embora essa estratégia permita identificar características recorrentes das vencedoras, ela não responde a uma questão igualmente relevante: em que momento da competição torna-se possível identificar uma provável campeã? Em outras palavras, ainda não se compreende como o potencial preditivo dessas informações evolui à medida que novos episódios são exibidos.

Diante dessa lacuna, este trabalho propõe uma abordagem temporal para análise de *RuPaul's Drag Race*, reorganizando a base de dados em uma estrutura participante-episódio e aplicando técnicas de aprendizado de máquina para acompanhar a evolução das competidoras ao longo da competição. A partir dessa representação, busca-se compreender como padrões competitivos e narrativos se desenvolvem ao longo das temporadas e em que medida essas informações podem ser utilizadas para antecipar a identificação das futuras vencedoras.

1.1. Objetivos Gerais

Considerando a problemática apresentada, esta pesquisa tem como objetivo investigar a evolução temporal dos padrões competitivos e narrativos presentes em *RuPaul's Drag Race* por meio de técnicas de aprendizado de máquina. Busca-se compreender de que maneira a combinação entre desempenho competitivo, construção narrativa e presença televisiva influencia a probabilidade de uma participante tornar-se vencedora da competição ao longo da temporada.

Diferentemente da abordagem desenvolvida em MSI I, que analisava os dados de forma consolidada ao final das temporadas, este estudo concentra-se na dimensão temporal do problema, avaliando como a capacidade preditiva dos modelos evolui à medida que novos episódios são incorporados à análise. Dessa forma, procura-se identificar não apenas quais características estão associadas às vencedoras, mas também em que momento essas características passam a fornecer evidências suficientes para distingui-las das demais competidoras.

Nesse contexto, este trabalho busca responder às seguintes questões de pesquisa:

- *QP1*: É possível identificar, antes do encerramento da temporada, quais participantes apresentam maior probabilidade de se tornarem vencedoras da competição?
- *QP2*: Como a capacidade preditiva dos modelos evolui ao longo dos episódios de uma temporada?
- *QP3*: Existem padrões temporais recorrentes na construção competitiva e narrativa das vencedoras de *RuPaul's Drag Race*?
- *QP4*: A incorporação de informações temporais melhora a capacidade de identificação das vencedoras quando comparada a abordagens baseadas exclusivamente em dados consolidados?

1.2. Objetivos Específicos

Para responder às questões de pesquisa propostas, foram definidos os seguintes objetivos específicos:

1. Adaptar a base de dados desenvolvida em MSI I para uma estrutura temporal, organizada no formato participante-episódio, permitindo acompanhar a evolução das compe-

tidoras ao longo das temporadas.

2. Construir variáveis capazes de representar simultaneamente informações competitivas, narrativas e de presença televisiva em diferentes momentos da competição.
3. Desenvolver e avaliar modelos de aprendizado de máquina voltados para a estimativa da probabilidade de vitória das participantes ao longo da temporada.
4. Implementar uma estratégia de validação baseada em *Leave-One-Season-Out* (LOSO), permitindo avaliar a capacidade de generalização dos modelos para temporadas não observadas durante o treinamento.
5. Comparar diferentes representações dos dados, incluindo informações acumuladas, semanais e sequenciais, buscando identificar quais abordagens produzem melhor desempenho preditivo.
6. Investigar o impacto da incorporação de mecanismos de ponderação temporal, atribuindo maior relevância a informações provenientes de episódios mais próximos do desfecho da competição.
7. Analisar o comportamento das probabilidades preditas ao longo das temporadas, identificando padrões recorrentes de consolidação das participantes vencedoras.
8. Avaliar o desempenho dos modelos por meio de métricas de ranqueamento, verificando a frequência com que a vencedora real é posicionada entre as participantes mais prováveis de cada temporada.

2. Referencial Teórico

Esta seção apresenta os principais conceitos que fundamentam a pesquisa, abordando competições narrativas, aprendizado de máquina, análise preditiva, o sistema Edgic e os trabalhos relacionados à aplicação de técnicas computacionais em reality shows.

2.1. Reality TV e Competições Narrativas

Os reality shows de competição diferenciam-se de competições tradicionais por combinarem desempenho objetivo com construção narrativa. Embora os participantes sejam avaliados por meio de desafios e eliminações periódicas, a edição do programa também influencia a forma como suas trajetórias são apresentadas ao público, produzindo narrativas de crescimento, conflito, superação e desenvolvimento pessoal [Butler 2020, Blair 2010].

Nesse contexto, *RuPaul's Drag Race* representa um caso particularmente relevante. O programa combina avaliações técnicas, criatividade artística e desempenho performático com uma narrativa construída ao longo dos episódios, fazendo com que o sucesso das participantes dependa simultaneamente de fatores competitivos e editoriais. Essa característica torna o programa um ambiente adequado para estudos envolvendo modelagem preditiva baseada em múltiplas dimensões de informação [Chan 2023].

2.2. Aprendizado de Máquina e Classificação

O aprendizado de máquina (*Machine Learning*) consiste no desenvolvimento de modelos capazes de identificar padrões em conjuntos de dados e utilizá-los para realizar inferências sobre novas observações. Em problemas supervisionados de classificação, o objetivo é aprender uma função capaz de associar exemplos a categorias previamente definidas a partir de exemplos rotulados.

No contexto desta pesquisa, o problema foi formulado como uma tarefa de classificação binária, na qual o modelo estima a probabilidade de uma participante tornar-se

vencedora da temporada. Como existe apenas uma vencedora por edição, o conjunto de dados apresenta forte desbalanceamento entre as classes, tornando necessário o emprego de estratégias capazes de lidar adequadamente com essa característica.

Além da obtenção de previsões, modelos de aprendizado de máquina também possibilitam analisar a contribuição relativa das variáveis utilizadas, permitindo investigar quais fatores apresentam maior influência sobre o fenômeno estudado [Gulhane and Sajana 2021].

2.3. Análise Preditiva em Contextos Competitivos

A utilização de dados para apoiar processos de tomada de decisão consolidou-se em diferentes áreas, dando origem ao campo conhecido como *Analytics*. Segundo Davenport [Davenport 2006], organizações orientadas por dados utilizam modelos quantitativos para compreender padrões e obter vantagens competitivas.

No contexto esportivo, a área de *Sports Analytics* emprega métodos estatísticos para análise de desempenho, planejamento estratégico e avaliação de atletas e equipes [Mondello and Kamke 2014, Morgulev et al. 2018]. Embora reality shows apresentem diferenças importantes em relação às competições esportivas, ambos compartilham características relevantes, como a existência de eventos sucessivos, evolução temporal dos participantes e disponibilidade de informações históricas.

Dessa forma, técnicas de análise preditiva podem ser adaptadas para competições narrativas, incorporando não apenas indicadores de desempenho, mas também informações relacionadas à construção editorial das trajetórias dos participantes.

2.4. Edgic e Construção Narrativa

Uma das abordagens mais utilizadas por comunidades especializadas para analisar a narrativa de participantes em reality shows é o sistema conhecido como *Edgic*. O método foi desenvolvido originalmente para análises de *Survivor*, reality show competitivo norte-americano considerado um dos principais precursores do gênero moderno e referência para diversos programas de competição televisiva [Tisdale 2021]. A partir desse contexto, o Edgic passou a ser utilizado para identificar padrões editoriais por meio da classificação sistemática da representação dos participantes em cada episódio [Holmes 2015].

O sistema é tradicionalmente composto por três dimensões principais:

- **Rating:** representa o grau de complexidade da narrativa atribuída ao participante;
- **Tone:** identifica se a edição apresenta predominantemente características positivas, negativas ou neutras;
- **Visibility:** estima o nível de exposição do participante durante o episódio.

Embora envolva certo grau de subjetividade, o Edgic tornou-se uma ferramenta amplamente utilizada por comunidades de fãs para acompanhar a evolução narrativa de competidores em reality shows. Como consequência, foram construídas bases históricas contendo classificações consistentes para diferentes temporadas, utilizadas tanto por entusiastas quanto por pesquisas acadêmicas.

Neste trabalho, parte das informações narrativas foi obtida a partir das classificações de Edgic publicadas por um único usuário no fórum *Club That Suck* [Chi 2025].

A adoção dessa fonte única buscou garantir consistência na representação narrativa das participantes entre as diferentes temporadas, evitando variações que poderiam surgir da combinação de classificações produzidas por autores distintos. Dessa forma, todas as observações narrativas utilizadas no estudo seguem um mesmo conjunto implícito de critérios interpretativos.

2.5. Trabalhos Relacionados

A aplicação de técnicas de aprendizado de máquina a programas de entretenimento televisivo ainda representa uma área relativamente recente quando comparada a domínios tradicionais, como análise financeira ou esportiva. Ainda assim, estudos recentes demonstram o potencial dessas abordagens em diferentes contextos.

Em [Lee et al. 2022], foram empregados algoritmos de aprendizado de máquina para prever o desempenho de participantes do reality show *The Bachelor*, identificando características demográficas e históricas associadas ao sucesso na competição.

No estudo apresentado em [Quinn 2023], foram combinadas informações competitivas e narrativas para prever a progressão de participantes em *Survivor*, obtendo desempenho superior ao de modelos baseados exclusivamente em indicadores objetivos.

Uma abordagem distinta foi proposta por [Sakiyama et al. 2020], que utilizaram técnicas de processamento de linguagem natural para analisar comentários publicados em redes sociais, demonstrando que informações provenientes do público apresentam potencial para estimar acontecimentos em reality shows.

Além dos estudos voltados especificamente para programas televisivos competitivos, [Cammarano et al. 2024] investigaram a previsão da popularidade de séries por meio de técnicas de aprendizado de máquina, evidenciando a aplicabilidade desses métodos em diferentes produtos audiovisuais.

Apesar desses avanços, observa-se que os trabalhos existentes concentram-se predominantemente na predição de resultados finais ou na utilização de informações consolidadas após a ocorrência dos eventos analisados. Não foram identificados estudos que investigassem a evolução temporal da probabilidade de vitória em competições narrativas utilizando uma representação estruturada no formato participante-episódio.

Dessa forma, o presente trabalho diferencia-se da literatura existente ao propor uma abordagem temporal capaz de acompanhar, episódio a episódio, a evolução competitiva e narrativa das participantes, permitindo investigar não apenas quem possui maior probabilidade de vencer, mas também como essa probabilidade se consolida ao longo da competição.

3. Metodologia

Esta pesquisa tem como objetivo investigar a evolução temporal dos padrões competitivos e narrativos associados às vencedoras de *RuPaul's Drag Race*. Para isso, foi desenvolvida uma abordagem baseada em aprendizado de máquina capaz de acompanhar, episódio a episódio, a trajetória das participantes ao longo da competição.

A metodologia adotada compreende quatro etapas principais: construção da base temporal de dados, engenharia de atributos, desenvolvimento dos modelos preditivos e avaliação dos resultados por meio de validação cruzada entre temporadas.



Figura 1. Fluxo metodológico da abordagem proposta em MSI II, desde a transformação da base de dados até a avaliação dos modelos preditivos.

A Figura 1 apresenta uma visão geral das etapas desenvolvidas nesta pesquisa. Nas subseções seguintes, cada uma dessas etapas é descrita em maior nível de detalhamento.

3.1. Base de Dados

A base utilizada nesta pesquisa foi construída a partir dos dados coletados durante o desenvolvimento de MSI I. O conjunto contempla as dezoito temporadas regulares de *RuPaul's Drag Race* exibidas até o momento da realização deste estudo.

Os dados incluem informações competitivas, narrativas e de caracterização das participantes. Entre os atributos considerados encontram-se resultados obtidos nos desafios semanais, classificações narrativas derivadas do sistema Edgic, métricas de visibilidade, informações demográficas e características relacionadas ao perfil artístico das competidoras.

Enquanto a base desenvolvida em MSI I possuía uma estrutura consolidada ao final das temporadas, o presente trabalho exigiu uma reorganização dos dados para permitir análises temporais. Dessa forma, cada observação passou a representar uma combinação participante-episódio, possibilitando acompanhar a evolução individual das competidoras ao longo da competição.

Ao final desse processo, a base passou a conter múltiplos registros para cada participante, correspondentes aos diferentes momentos de sua trajetória na temporada.

3.2. Construção da Base Temporal

A principal adaptação metodológica realizada em MSI II consistiu na transformação da base consolidada em uma estrutura temporal.

Para cada participante foram geradas observações correspondentes aos episódios em que permaneceu ativa na competição. A cada novo episódio, os atributos acumulados foram atualizados para refletir o histórico disponível naquele momento específico da temporada.

Essa abordagem permite simular um cenário real de predição, no qual apenas informações conhecidas até determinado episódio podem ser utilizadas para estimar a probabilidade de vitória de cada participante.

A Tabela 1 ilustra a principal diferença entre a estrutura utilizada em MSI I e a representação temporal adotada neste trabalho.

Tabela 1. Exemplo simplificado da transformação da estrutura dos dados entre a base consolidada utilizada em MSI I e a base temporal participante-episódio desenvolvida em MSI II.

MSI I – Base Consolidada					
Nome	Total Win	Total CP	Total MOR	Ranking Final	Winner
Bebe Zahara Benet	2	4	2	Winner	1
Nina Flowers	1	3	4	Runner-Up	0

MSI II – Base Temporal					
Ep.	Nome	Competitivo	Narrativo	Seq. 3 ep.	Total Win
1	Bebe Zahara Benet	Safe	MOR	Safe	0
2	Bebe Zahara Benet	Safe	UTR	Safe > Safe	0
3	Bebe Zahara Benet	Win	CP	Safe > Safe > Win	1
4	Bebe Zahara Benet	High	MOR	Safe > Win > High	1

Observa-se que, enquanto a base utilizada em MSI I continha apenas uma observação consolidada para cada participante ao término da temporada, a estrutura proposta em MSI II representa a evolução da competição por meio de observações no formato participante-episódio. Cada registro passa a conter apenas as informações disponíveis até o episódio correspondente, incluindo atributos competitivos, narrativos, acumulados e sequenciais. Essa representação permite acompanhar a evolução das participantes ao longo da temporada e reproduz um cenário mais próximo de uma aplicação real, em que as previsões são realizadas progressivamente à medida que novos episódios são exibidos.

3.3. Engenharia de Atributos

Após a construção da base temporal, foi realizada uma etapa de engenharia de atributos com o objetivo de representar diferentes aspectos da trajetória das participantes ao longo da competição.

Os atributos utilizados podem ser agrupados em quatro categorias principais:

- **Atributos competitivos:** representam o desempenho da participante na competição, incluindo vitórias em desafios, colocações altas, posições de risco, eliminações e demais indicadores de desempenho acumulados até cada episódio.
- **Atributos narrativos:** correspondem às classificações derivadas do sistema Edgic, contemplando dimensões relacionadas à construção narrativa, ao tom editorial e ao nível de visibilidade das participantes.
- **Atributos temporais:** incluem representações acumuladas e sequenciais das informações competitivas e narrativas. Além dos totais observados até determinado episódio, foram construídas variáveis capazes de representar o comportamento recente das participantes por meio de sequências correspondentes aos episódios mais recentes da competição.
- **Atributos de caracterização:** compreendem informações relativamente estáveis das participantes, como características demográficas e atributos relacionados ao perfil artístico, preservados da base desenvolvida em MSI I.

Essa organização permitiu que os modelos explorassem simultaneamente informações históricas, tendências recentes e características individuais das competidoras, produzindo uma representação mais completa de suas trajetórias ao longo da temporada.

3.4. Tratamento dos Dados

Antes da etapa de treinamento, os dados passaram por um processo de pré-processamento integrado ao pipeline de aprendizado de máquina.

As variáveis categóricas foram transformadas por meio da técnica de *One-Hot Encoding*, enquanto os atributos numéricos foram padronizados utilizando normalização baseada em média e desvio padrão. Valores ausentes foram tratados por imputação automática, garantindo consistência ao conjunto de dados sem a necessidade de remoção de observações.

3.5. Formulação do Problema

O problema foi modelado como uma tarefa de classificação binária supervisionada.

Para cada observação participante-episódio, o modelo deveria estimar a probabilidade de a participante tornar-se vencedora da respectiva temporada.

A variável alvo foi definida da seguinte forma:

- $Winner = 1$, caso a participante tenha vencido a temporada;
- $Winner = 0$, para todas as demais participantes.

Embora o treinamento tenha sido realizado como um problema de classificação binária, o objetivo final do estudo consiste em ranquear as participantes dentro de cada temporada. Dessa forma, as probabilidades produzidas pelos modelos foram utilizadas para ordenar as competidoras e identificar aquelas com maior chance estimada de vitória.

3.6. Estratégia de Validação

A avaliação dos modelos foi realizada utilizando a estratégia *Leave-One-Season-Out* (LOSO). Estratégias de validação que separam grupos inteiros entre treino e teste são recomendadas quando há dependência entre observações de um mesmo grupo, pois evitam estimativas excessivamente otimistas de generalização [Varoquaux et al. 2017].

Em cada iteração do processo, uma temporada completa era removida do conjunto de treinamento e utilizada exclusivamente para teste. As demais temporadas eram utilizadas para ajustar o modelo.

Essa abordagem foi escolhida para evitar que informações provenientes de uma mesma temporada aparecessem simultaneamente nos conjuntos de treino e teste. Como participantes de uma mesma edição compartilham contexto competitivo e narrativo, divisões aleatórias poderiam produzir estimativas artificialmente otimistas do desempenho do modelo.

A validação LOSO permite avaliar a capacidade de generalização para temporadas completamente inéditas, simulando de forma mais realista a aplicação do sistema em futuras edições do programa.

3.7. Modelagem Preditiva

Diversas configurações foram avaliadas ao longo do desenvolvimento da pesquisa, incluindo representações baseadas exclusivamente em dados acumulados, conjuntos contendo informações semanais e abordagens que combinavam diferentes tipos de atributos.

Os experimentos demonstraram que a melhor performance foi obtida por uma Regressão Logística Balanceada integrada a um pipeline de pré-processamento e seleção automática de atributos.

O pipeline final foi composto pelas seguintes etapas:

1. Imputação de valores ausentes;
2. Padronização de atributos numéricos;
3. Codificação das variáveis categóricas;
4. Seleção automática de atributos por meio de *SelectKBest*;
5. Treinamento de uma Regressão Logística Balanceada.

A utilização de balanceamento de classes foi necessária devido à natureza do problema, uma vez que existe apenas uma vencedora por temporada, resultando em uma distribuição altamente desbalanceada da variável alvo.

3.8. Ponderação Temporal por Recência

Durante os experimentos observou-se que episódios mais próximos do final da competição tendem a conter sinais mais relevantes para a identificação da vencedora do que episódios iniciais.

Com base nessa hipótese, foi introduzido um mecanismo de ponderação temporal por recência. Cada observação recebeu um peso proporcional à sua posição dentro da temporada, atribuindo maior importância aos episódios mais recentes durante o treinamento.

Essa estratégia permitiu enfatizar informações produzidas em momentos mais próximos da coroação sem introduzir qualquer conhecimento sobre eventos futuros, preservando a validade metodológica do experimento.

3.9. Métricas de Avaliação

Como o objetivo principal do estudo consiste em identificar corretamente as participantes mais prováveis de vencer cada temporada, a avaliação foi realizada utilizando métricas de ranqueamento.

Para cada temporada, as participantes foram ordenadas de acordo com a probabilidade estimada pelo modelo no último episódio considerado. A partir desse ranking foram calculadas as seguintes métricas:

- Top 1 Final: proporção de temporadas em que a vencedora real foi posicionada em primeiro lugar;
- Top 2 Final: proporção de temporadas em que a vencedora real apareceu entre as duas primeiras posições;
- Top 3 Final: proporção de temporadas em que a vencedora real apareceu entre as três primeiras posições;
- Rank Médio da Vencedora: posição média ocupada pela vencedora real nos rankings produzidos pelo modelo.

Essas métricas permitem avaliar não apenas a capacidade de identificar corretamente a campeã, mas também o quão próxima ela se encontra do topo das previsões produzidas pelo sistema.

4. Resultados

Esta seção apresenta os resultados obtidos pelo modelo final selecionado. Inicialmente são discutidas as métricas globais de desempenho. Em seguida, são analisados o comportamento do modelo por temporada e a evolução temporal das probabilidades estimadas para as participantes.

4.1. Desempenho Geral do Modelo

A avaliação foi realizada utilizando a estratégia *Leave-One-Season-Out* (LOSO), permitindo estimar a capacidade de generalização do modelo para temporadas não utilizadas durante o treinamento.

A Figura 2 apresenta o desempenho obtido pelo modelo final nas métricas Top 1, Top 2 e Top 3.

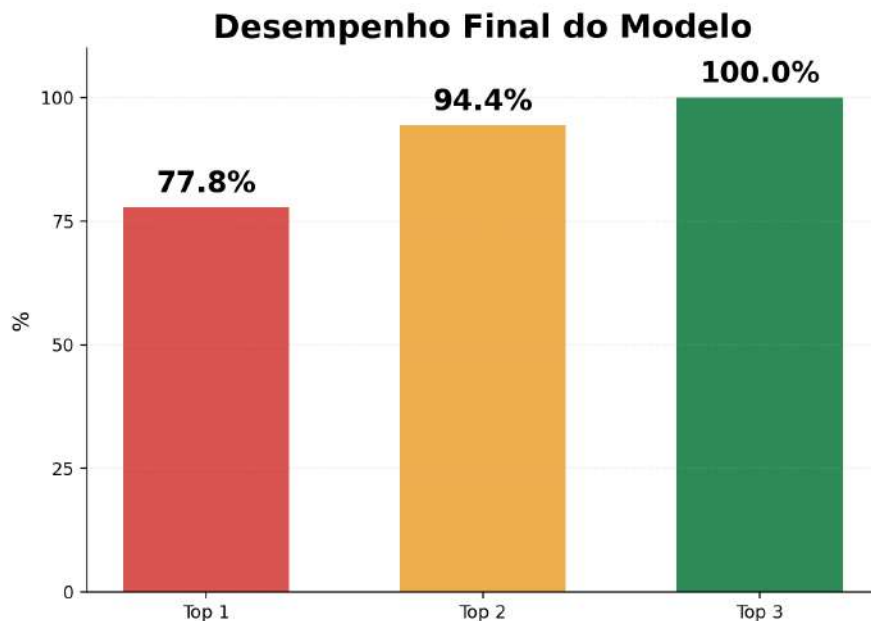


Figura 2. Desempenho do modelo final nas métricas Top 1, Top 2 e Top 3.

Observa-se que o modelo identificou corretamente a vencedora em 77,8% das temporadas analisadas. Em 94,4% dos casos, a campeã real foi posicionada entre as duas participantes mais bem classificadas e, em todas as dezoito temporadas avaliadas, figurou entre as três maiores probabilidades estimadas.

Esses resultados evidenciam elevada capacidade de ranqueamento, indicando que o modelo consegue identificar de forma consistente o conjunto de participantes com maior potencial de vitória.

4.2. Desempenho por Temporada

A Figura 3 apresenta a posição ocupada pela vencedora real no ranking produzido pelo modelo para cada temporada.

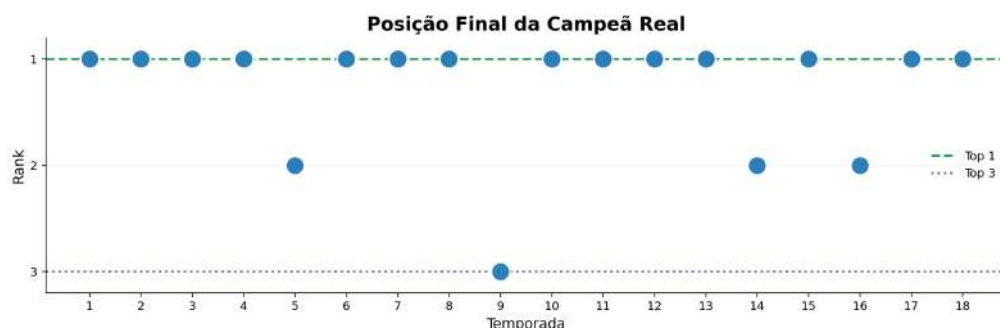


Figura 3. Posição final da vencedora real no ranking produzido pelo modelo.

Em quatorze das dezoito temporadas analisadas, a vencedora foi corretamente posicionada na primeira colocação. Nas quatro temporadas restantes, correspondentes às

temporadas 5, 9, 14 e 16, embora não tenha ocorrido acerto Top 1, a campeã permaneceu entre as participantes mais bem avaliadas pelo modelo, contribuindo para os elevados índices observados nas métricas Top 2 e Top 3.

Esses resultados mostram que, mesmo nos casos de erro, o modelo foi capaz de identificar corretamente o grupo de participantes com maior potencial competitivo e narrativo.

4.3. Evolução Temporal das Probabilidades

Além da classificação final, foi analisada a evolução das probabilidades produzidas pelo modelo ao longo dos episódios de cada temporada.

A Figura 4 apresenta um exemplo de consolidação gradual da vencedora ao longo da temporada.

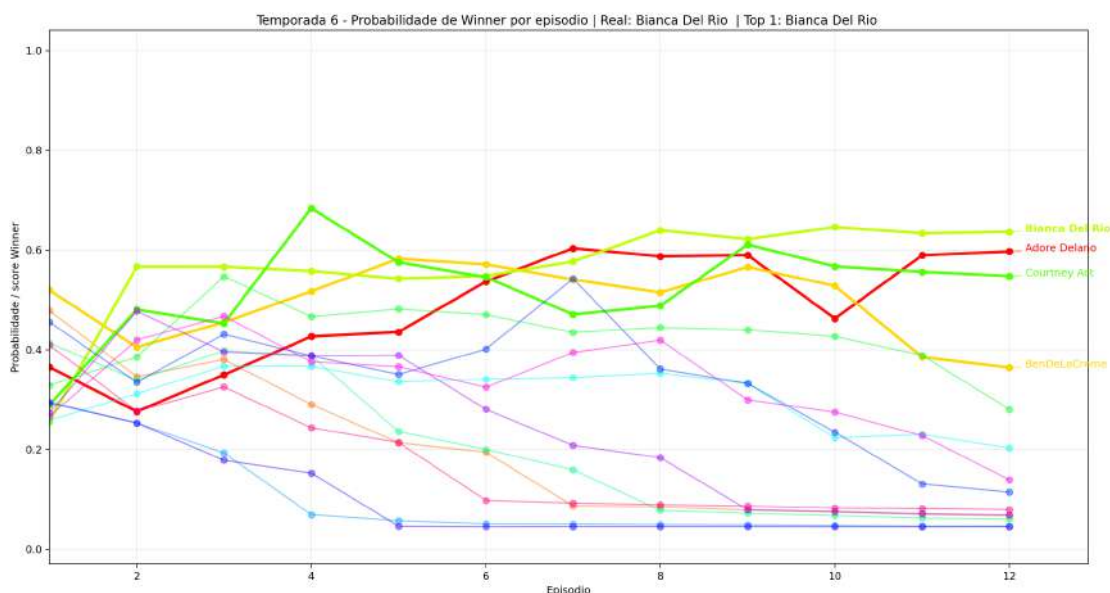


Figura 4. Exemplo de consolidação gradual da vencedora ao longo da temporada.

A análise da trajetória mostra que, nos episódios iniciais, as probabilidades permanecem mais distribuídas entre diferentes participantes. À medida que novos episódios são incorporados, a participante vencedora passa gradualmente a ocupar posições superiores no ranking, consolidando sua liderança nos episódios finais da competição.

Esse comportamento sugere que a identificação da vencedora não ocorre necessariamente de forma imediata, mas resulta da acumulação progressiva de sinais competitivos e narrativos ao longo da temporada.

A Figura 5 apresenta a trajetória da Temporada 9, caso em que a vencedora real foi posicionada em terceiro lugar pelo modelo.

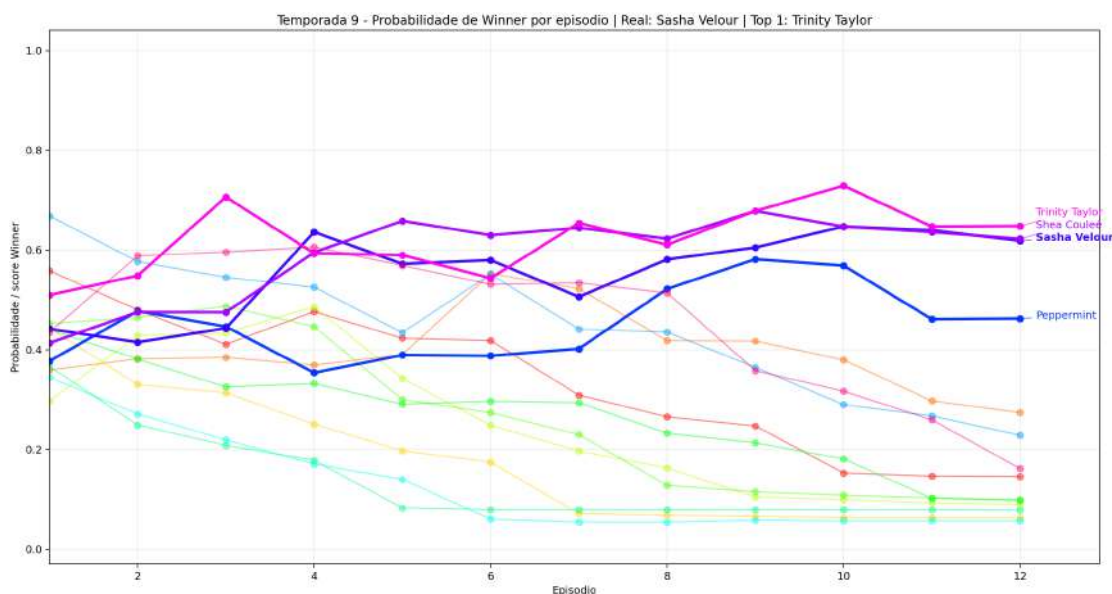


Figura 5. Evolução temporal das probabilidades na Temporada 9.

Nesse caso, observa-se que a diferença entre as participantes classificadas em segundo e terceiro lugar é bastante reduzida nos episódios finais. Assim, embora o modelo não tenha posicionado a vencedora em primeiro lugar, ele identificou corretamente o grupo de candidatas com maior potencial de vitória.

4.4. Padrões Observados

A análise conjunta dos resultados indica que as probabilidades estimadas pelo modelo tendem a seguir um comportamento progressivo. Nos episódios iniciais, há maior dispersão entre as participantes. Na fase intermediária, ocorre redução do número de candidatas com probabilidades competitivas. Nos episódios finais, observa-se maior concentração das probabilidades nas participantes mais relevantes da temporada.

Esse padrão reforça a hipótese de que a construção da vencedora em *RuPaul's Drag Race* ocorre de forma gradual, sendo influenciada pela combinação entre desempenho competitivo, construção narrativa e visibilidade ao longo da temporada.

5. Discussão

Os resultados obtidos demonstram que a incorporação da dimensão temporal amplia a capacidade de análise de competições narrativas. Mais do que identificar vencedoras ao final da temporada, a abordagem proposta permite acompanhar como as evidências associadas à vitória se acumulam episódio a episódio.

5.1. Contribuição da Dimensão Temporal

A principal contribuição deste trabalho consiste na transformação de uma base consolidada em uma representação temporal estruturada no formato participante-episódio.

Enquanto MSI I avaliava apenas o estado final de cada temporada, a abordagem proposta em MSI II possibilitou acompanhar a evolução das participantes ao longo da

competição. Essa mudança permitiu reproduzir um cenário mais próximo de uma aplicação real, no qual novas previsões podem ser feitas à medida que novos episódios são exibidos.

Como observado nas Figuras 4 e 5, as probabilidades estimadas pelo modelo não permanecem estáticas. Elas variam conforme novas informações competitivas e narrativas são incorporadas, indicando que a identificação das vencedoras depende de um processo gradual de acúmulo de evidências.

5.2. Competição e Construção Narrativa

Os experimentos indicam que o desempenho competitivo, embora fundamental, não é suficiente para explicar integralmente a definição das vencedoras.

A combinação entre resultados em desafios, atributos narrativos derivados do sistema Edgic e indicadores de presença televisiva produziu desempenho consistente nas diferentes temporadas avaliadas. Esse comportamento reforça a natureza híbrida de competições narrativas, nas quais desempenho e construção editorial coexistem na formação da trajetória das participantes.

Dessa forma, os resultados sugerem que a coroação em *RuPaul's Drag Race* não pode ser compreendida apenas como consequência direta do histórico competitivo. Ela emerge da interação entre performance, narrativa e visibilidade ao longo da temporada.

5.3. Análise dos Casos de Erro

As temporadas 5, 9, 14 e 16 representam os únicos casos em que o modelo não obteve acerto Top 1.

Apesar disso, em todas essas temporadas a vencedora permaneceu entre as participantes mais bem classificadas, indicando que o modelo foi capaz de reconhecer sua relevância competitiva e narrativa.

O caso da Temporada 9, apresentado na Figura 5, merece destaque por apresentar probabilidades finais bastante próximas entre as candidatas mais bem posicionadas. Esse comportamento sugere que o erro decorre principalmente da dificuldade de ordenação entre participantes com trajetórias semelhantes, e não da incapacidade de identificar corretamente o grupo de favoritas.

Esses resultados também sugerem que determinados fatores subjetivos, como decisões de produção, escolhas editoriais específicas ou acontecimentos externos à competição, podem exercer influência sobre o desfecho final sem estarem completamente representados pelas variáveis utilizadas.

5.4. Limitações

Entre as principais limitações da pesquisa destacam-se o número reduzido de temporadas disponíveis para treinamento e a natureza subjetiva de parte das variáveis narrativas derivadas do sistema Edgic.

Além disso, fatores externos à competição, como repercussão junto ao público, popularidade em redes sociais e decisões específicas da produção, não foram incorporados ao conjunto de atributos utilizado.

Ainda assim, o desempenho obtido demonstra que a representação proposta captura uma parcela significativa dos padrões associados à construção das vencedoras.

5.5. Resposta às Questões de Pesquisa

Os resultados obtidos permitem responder positivamente às questões propostas neste estudo.

Em relação à primeira questão, observou-se que é possível identificar participantes com elevada probabilidade de vitória antes do encerramento das temporadas, uma vez que a vencedora real esteve entre as três participantes mais prováveis em todas as edições analisadas.

Quanto à segunda questão, verificou-se que a capacidade preditiva evolui progressivamente ao longo dos episódios, acompanhando o acúmulo de informações competitivas e narrativas.

Em relação à terceira questão, foram observados padrões temporais recorrentes, especialmente a dispersão inicial das probabilidades, a redução gradual do grupo de favoritas e a consolidação final das candidatas mais fortes.

Por fim, quanto à quarta questão, os resultados indicam que a incorporação de informações temporais melhora a capacidade de identificação das vencedoras ao permitir que o modelo represente não apenas o estado final da competição, mas também o processo de construção das trajetórias ao longo da temporada.

6. Conclusão

O presente trabalho teve como objetivo investigar a evolução temporal dos padrões competitivos e narrativos associados às vencedoras de *RuPaul's Drag Race* por meio de técnicas de aprendizado de máquina. Partindo da base de dados desenvolvida em MSI I, foi proposta uma nova abordagem baseada em observações estruturadas no formato participante-episódio, permitindo analisar não apenas quais características estão associadas às vencedoras, mas também como essas características se acumulam ao longo da competição.

Para atingir esse objetivo, foram construídos modelos capazes de incorporar simultaneamente informações competitivas, narrativas e de presença televisiva. Além disso, foi adotada uma estratégia de validação *Leave-One-Season-Out* (LOSO), garantindo que a avaliação refletisse a capacidade de generalização para temporadas não observadas durante o treinamento.

Os resultados obtidos demonstraram que a incorporação da dimensão temporal contribui significativamente para a identificação de vencedoras em competições narrativas. O modelo final alcançou 77,8% de acerto Top 1, 94,4% de acerto Top 2 e 100% de acerto Top 3, posicionando a vencedora real entre as três participantes mais prováveis em todas as dezoito temporadas analisadas. Esses resultados evidenciam que a trajetória das futuras campeãs apresenta padrões suficientemente consistentes para serem identificados por métodos estatísticos antes do encerramento da competição.

A análise das probabilidades produzidas ao longo das temporadas revelou que a construção da vencedora ocorre de forma gradual. Nos episódios iniciais, as previsões

tendem a permanecer distribuídas entre diversas participantes, refletindo a incerteza característica dos estágios iniciais da competição. À medida que novos episódios são incorporados, observa-se a formação progressiva de um grupo reduzido de candidatas competitivas, culminando na consolidação das participantes com maior probabilidade de vitória. Esse comportamento sugere que a coroação em competições narrativas não depende exclusivamente de eventos isolados, mas da acumulação contínua de sinais competitivos e narrativos ao longo da temporada.

Outro resultado relevante foi a constatação de que o desempenho competitivo, embora importante, não é suficiente para explicar sozinho a definição das vencedoras. Os experimentos mostraram que a inclusão de variáveis narrativas e de presença televisiva produz ganhos significativos de desempenho, indicando que a construção editorial exerce papel complementar ao desempenho objetivo na trajetória das participantes.

Do ponto de vista metodológico, a pesquisa demonstra a viabilidade da aplicação de técnicas de aprendizado de máquina em contextos de entretenimento televisivo e contribui para ampliar a literatura relacionada à análise quantitativa de competições narrativas. Além disso, a abordagem temporal proposta oferece uma alternativa para estudos que buscam compreender não apenas os resultados finais de um fenômeno, mas também seu processo de construção ao longo do tempo.

Apesar dos resultados positivos, algumas limitações permanecem. O número reduzido de temporadas disponíveis restringe o volume total de dados para treinamento, enquanto parte das variáveis narrativas utilizadas depende de classificações subjetivas. Além disso, fatores externos à competição, como repercussão em redes sociais, percepção do público e decisões específicas de produção, não foram incorporados ao modelo e podem influenciar o resultado final.

Como trabalhos futuros, sugere-se a incorporação de fontes adicionais de informação, especialmente dados relacionados à recepção do público em plataformas digitais e redes sociais. Também podem ser exploradas abordagens voltadas à interpretação automática de narrativas televisivas por meio de processamento de linguagem natural. Por fim, a metodologia desenvolvida pode ser adaptada para outros reality shows competitivos, permitindo investigar se os padrões observados em *RuPaul's Drag Race* também se manifestam em diferentes formatos de competição narrativa.

Em síntese, os resultados obtidos indicam que competições narrativas apresentam estruturas suficientemente consistentes para possibilitar a identificação estatística de futuras vencedoras antes do encerramento da temporada. Mais do que prever resultados, esta pesquisa contribui para compreender como trajetórias competitivas e narrativas são construídas ao longo do tempo, oferecendo uma perspectiva quantitativa sobre um fenômeno tradicionalmente analisado de forma predominantemente qualitativa.

Referências

- [Blair 2010] Blair, J. L. (2010). Surviving reality tv: The ultimate challenge for reality show contestants. *Loy LA Ent L Rev*, 31(1).
- [Butler 2020] Butler, J. L. (2020). The rise and reign of reality tv. Master of science (m.s.) thesis, Drexel University.

- [Cammarano et al. 2024] Cammarano, M., Guarino, A., and Malandrino, D. e. a. (2024). Tv shows popularity prediction of genre-independent tv series through machine learning-based approaches. *Multimed Tools Appl*, 83:75757–75780.
- [Chan 2023] Chan, A. S. W. (2023). Rupaul’s drag race: A cultural phenomenon that challenges gender norms and sparks conversations across borders. *Journal of Homosexuality*, 71(8):1863–1866.
- [Chi 2025] Chi (2025). Chi’s rpdr edgic dump thread. Postagem no Fórum Online: clubsthatsuck.jcink.net. Acesso em: 3 dez. 2025.
- [Davenport 2006] Davenport, T. (2006). Competing on analytics. *Harvard Business Review*, 84:98–107, 134.
- [Gulhane and Sajana 2021] Gulhane, M. and Sajana, T. (2021). Human behavior prediction and analysis using machine learning – a review. *Turkish Journal of Computer and Mathematics Education*, 12(5):870–876.
- [Holmes 2015] Holmes, M. (2015). Survivor edgic – an introduction. Inside Survivor. Acesso em: 7 set. 2025.
- [Lee et al. 2022] Lee, A. J., Chesmore, G. E., Rocha, K. A., Farah, A., Sayeed, M., and Myles, J. (2022). Predicting winners of the reality tv dating show the bachelor using machine learning algorithms. Disponível em: <https://arxiv.org/abs/2203.16648>.
- [Mondello and Kamke 2014] Mondello, M. and Kamke, C. (2014). The introduction and application of sports analytics in professional sport organizations. *Journal of Applied Sport Management*, 6(2).
- [Morgulev et al. 2018] Morgulev, E., Azar, O. H., and Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5:213–222.
- [Quinn 2023] Quinn, K. (2023). Can machine learning models predict reality tv winners? the case of survivor. D-Lab. Acesso em: 7 set. 2025.
- [Sakiyama et al. 2020] Sakiyama, K., de Souza Rodrigues, L., and Matsubara, E. (2020). Can twitter data estimate reality show outcomes? In Cerri, R. and Prati, R., editors, *Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science*, volume 12319. Springer, Cham.
- [Tisdale 2021] Tisdale, S. (2021). *Survivor and the Endless Gaze: Survivor, Reality TV, and the Endless Gaze*. Gallery Books, New York.
- [Varoquaux et al. 2017] Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145(Part B):166–179.